UNIVERSIDAD NACIONAL DEL NORDESTE

FACULTAD DE CIENCIAS AGRARIAS

PROPUESTA DE MODELOS QUIMIOMÉTRICOS PARA ESTABLECER SISTEMAS DE TRAZABILIDAD DE NARANJAS PRODUCIDAS EN LA REGIÓN NORDESTE ARGENTINA

Ing. Agr. José Emilio GAIAD

Director. Roberto Gerardo PELLERANO

DOCTORADO EN RECURSOS NATURALES

2020

PUBLICACIONES

La realización de la presente Tesis Doctoral ha dado origen a la redacción y publicación de diversos trabajos de investigación en revistas de prestigio internacional:

- ✓ Gaiad, JE; Hidalgo, MJ; Villafañe, RN; Marchevsky, EJ; Pellerano, RG. 2016. Tracing the geographical origin of Argentinean lemon juices based on trace element profiles using advanced chemometric techniques. Microchemical Journal. 129: 243-248. https://doi.org/10.1016/j.microc.2016.07.002.

- ✓ Díaz, I; Mazza, SM.; Combarro, EF; Giménez, LI; Gaiad, JE. 2017. Machine learning applied to the prediction of citrus production. Spanish Journal of Agricultural Research. [S.l.] 15 (2), p. e0205, july 2017. ISSN 2171-9292. Available at: <https://revistas.inia.es/index.php/sjar/article/view/9090>. doi: http://dx.doi.org/10.5424/sjar/2017152-9090.

- ✓ Pérez-Rodríguez, M; Gaiad, JE; Hidalgo, MJ; Avanza, MV; Pellerano, RG. 2018. Classification of cowpea beans using multielemental fingerprinting combined with supervised learning. Food Control. https://doi: 10.1016/j.foodcont.2018.08.001.

ÍNDICE DE CONTENIDOS

ÍTEM	PÁGINA
CAPÍTULO I	
PROBLEMA DE ESTUDIO	
1.1. Introducción	1
1.2. Planteamiento del problema	9
1.3. Trazabilidad de los alimentos	11
1.4. Composición química de los vegetales	15
1.5. Huella dactilar	17
1.6. Objetivos	18
1.6.1. Objetivos Generales	19
1.6.2. Objetivos Particulares	19
1.7. Hipótesis	20
1.8. Referencias	20
CAPÍTULO II	
ESTADÍSTICA MULTIVARIANTE Y APRENDIZAJE AUTOMÁTICO	
2.1. Introducción	27
2.2. Análisis Multivariante de Datos	28
2.2.1. Medidas Descriptivas Multivariantes	31
2.2.2. Análisis gráfico	32
2.2.3. Análisis de la Variancia Multivariante	33
2.2.4. Análisis de Componentes Principales	34
2.2.5. Análisis Discriminante	36
2.3. Aprendizaje Automático	38

ÍTEM	PÁGINA
2.3.1. Aprendizaje Supervisado	41
2.3.1.1. Árboles de Decisión	42
2.3.1.2. Bosques Aleatorios	45
2.3.1.3. Algoritmos de Clasificación por Vecindad	46
2.3.1.4. Redes Neuronales Artificiales	47
2.3.1.5. Máquina de Vectores de Soporte	52
2.4. Entrenamiento de Modelos	55
2.5. Evaluación de Modelos	57
2.5.1. Matriz de Confusión	57
2.5.2. Índice Kappa	60
2.5.3. Métodos gráficos de evaluación de modelos de clasificación	61
2.6. Aplicaciones	62
2.7. Referencias	65
CAPÍTULO III	
MATERIAL Y MÉTODOS	
3.1. Área de estudio	71
3.1.1. Características de la región NOA	72
3.1.2. Características de la región NEA	73
3.2. Variedades estudiadas	76
3.2.1. Variedades de limonero	76
3.2.2. Variedades de mandarina	77
3.2.3. Variedades de naranjo	78
3.3. Muestras	79
3.3.1. Obtención de las muestras de frutos	79

ÍTEM	PÁGINA
3.3.2. Obtención de jugos	85
3.4. Análisis multielemental	86
3.4.1. Digestión de muestras	86
3.4.1.1. Pretratamiento de muestras para espectrometría de masas	86
3.4.1.2. Pretratamiento de muestras para espectroscopía de absorción atómica de llama y de emisión atómica	88
3.4.2. Determinación multielemental	89
3.4.2.1. Espectroscopía atómica de masas con plasma acoplado inductivamente	90
3.4.2.2. Espectroscopía de absorción atómica por llama	93
3.4.2.3. Espectroscopía óptica de emisión atómica por plasma de microondas	96
3.5. Datos	99
3.6. Análisis de datos	99
3.7. Referencias	104
CAPÍTULO IV	
DETERMINACIÓN DEL ORIGEN GEOGRÁFICO DE JUGOS DE LIMÓN BASADO EN PERFILES DE OLIGOELEMENTOS	
4.1. Introducción	109
4.2. Materiales y Métodos	110
4.2.1. Muestras y procedimiento analítico	110
4.2.2. Datos	111
4.2.3. Análisis de datos	111
4.3. Resultados y discusión	112
4.3.1. Caracterización multielemental	112
ÍTEM	**PÁGINA**
4.3.2. Análisis de componentes principales	113

4.3.3. Análisis supervisado de muestras 115

4.3.4. Comparación de modelos 117

4.4. Resumen de resultados 118

4.5. Referencias 119

CAPÍTULO V

DETERMINACIÓN DEL ORIGEN GEOGRÁFICO DE JUGOS DE MANDARINAS PRODUCIDAS EN EL NORDESTE DE ARGENTINA

5.1. Introducción 122

5.2. Materiales y Métodos 124

5.2.1. Obtención y procesamiento de las muestras 124

5.2.2. Datos 125

5.2.3. Análisis de datos 125

5.3. Resultados y discusión 126

5.3.1. Caracterización multielemental 126

5.3.2. Análisis exploratorio 127

5.3.3. Selección de modelos 128

5.3.3.1. Análisis Discriminante Lineal 129

5.3.3.2. Árboles de Decisión 131

5.3.3.3. K-Vecino más Cercano 134

5.3.3.4. Redes Neuronales Artificiales 135

5.3.3.5. Máquinas de Vectores Soporte 139

5.3.4. Comparación de modelos 140

5.3.5. Marcadores químicos de identidad 141

ÍTEM **PÁGINA**

5.4. Resumen de resultados 143

5.5. Referencias 145

CAPÍTULO VI

COMPOSICIÓN MINERAL Y MODELOS PARA DETERMINAR IDENTIDAD DE JUGOS DE NARANJAS PRODUCIDAS EN LA REGION NORDESTE ARGENTINA

6.1. Introducción 150

6.2. Materiales y Métodos 152

6.2.1. Obtención y procesamiento de las muestras 152

6.2.2. Datos 153

6.2.3. Análisis de datos 154

6.3. Resultados y discusión 154

6.3.1. Caracterización multielemental 155

6.3.2. Elementos de interés ambiental, toxicológico y nutricional 158

6.3.3. Análisis exploratorio 161

6.3.4. Propuesta de modelos clasificatorios 165

6.3.4.1. Selección de modelos con información de FAAS 166

6.3.4.1.1. Análisis Discriminante Lineal 166

6.3.4.1.2. Árboles de Decisión 168

6.3.4.1.3. K-Vecino más Cercano 171

6.3.4.1.4. Redes Neuronales Artificiales 171

6.3.4.1.5. Máquinas de Vectores Soporte 175

6.3.4.1.6. Comparación de modelos 176

6.3.4.2. Selección de modelos con información de MP-AES 177

ÍTEM **PÁGINA**

6.3.4.2.1. Análisis Discriminante Lineal 177

6.3.4.2.2. Árboles de Decisión 180

6.3.4.2.3. K-Vecino más Cercano 181

6.3.4.2.4. Redes Neuronales Artificiales 182

6.3.4.2.5. Máquinas de Vectores Soporte 186

6.3.4.2.6. Comparación de modelos 188

6.3.5. Marcadores químicos de identidad 189

6.4. Resumen de resultados 191

6.5. Referencias 194

CAPÍTULO VII

CONCLUSIONES GENERALES 201

RESUMEN

Los mercados más competitivos de frutos cítricos requieren conocer el origen geográfico e identidad de estos, para lo que se necesita comprobar la identidad física de las muestras. La composición mineral de los vegetales obedece a patrones generales definidos para especies y variedades, pero cierta variabilidad se debe a condiciones de los sitios en que crecen, por lo que los contenidos minerales permitirían diseñar modelos matemáticos para definir su trazabilidad.

Esta tesis se ha realizado con los objetivos evaluar la presencia de marcadores químicos de trazabilidad en jugos de frutas cítricas, mediante técnicas de huella dactilar y de aprendizaje automático, con especial énfasis en la determinación de la composición química multielemental y demostrar la capacidad de métodos de análisis de datos multivariantes y de aprendizaje automático para establecer modelos de predicción para autenticar o confirmar la identidad de jugos de frutos de limón producidos en las regiones NEA y NOA, y de mandarina y naranja producidas en la región NEA.

Se trabajó con información derivada de muestras de frutos de limonero (*Citrus limon* L., Osbeck) 'Eureka', 'Lisboa' y 'Génova' de cuatro zonas de NOA y NEA, de mandarino (tangor) 'Murcott' (*C. sinensis* L. x C. reshni) y 'Okitsu' *(C. unshiu* Marc.) y naranjo dulce (*C. sinensis* L.) 'Valencia late' y 'Salustiana', de cuatro zonas del NEA, caracterizadas por su composición multielemental.

La determinación de concentraciones de elementos en muestras digeridas de jugos de limón se llevó a cabo por espectrometría de masas por plasma acoplado (ICP-MS), las de jugos de naranja mediante espectroscopía de absorción atómica de llama (FAAS) y las de jugos de naranja y mandarina por espectroscopía de emisión atómica de plasma de microondas (MP-AES). Se aplicaron análisis de: Varianza (ANOVA) y Varianza Multivariado (MANOVA) y Prueba de Hottelling, Componentes Principales (PCA), Discriminante Lineal (LDA) y de Mínimos Parciales (PLS DA), K Vecinos más

Cercanos (KNN), Redes Neuronales Artificiales (ANN), Máquinas Vectoriales de Soporte (SVM) y Bosques Aleatorios (RF). Para comparar métodos y seleccionar modelos se emplearon criterios de sensibilidad, especificidad, porcentaje de acierto e índice κ. Los análisis se realizaron con InfoStat 2020 y R 3.2.1.

Se caracterizaron los jugos de diferentes regiones por sus contenidos de elementos minerales. En los de limón se observaron concentraciones de Fe, Zn y Rb mayores a 10 μg/g, de Al, Ba, Cu, Mn y Ni entre 1 y 10 μg/g y de La, Cr, Se, Li, Mo, Co, Sn, Sc, V y Bi menores a 1 μg/g. Los de mandarina y naranja presentaron perfiles similares, el elemento más abundante fue K (cercano a 1000 μg/g), Al, Mg y Ca > 10 μg/g, Mn, Cu, Zn y Sr entre 1 y 5 μg/g y Cd, Cr y Fe < 1 μg/g.

En los jugos de naranja se describieron elementos nocivos. Según establece el Código Alimentario Argentino, los promedios de Cu en 'Valencia late' superaron los máximos y el Pb superó los máximos en algunas muestras, no en promedio. Los niveles de los todos los elementos se encontraron por debajo de los máximos establecidos por la Association of American Food Control Oficials (AFCO). El contenido promedio de Cd de los jugos de naranja superó el máximo permitido por la OMS y la FAO.

Se detectaron potenciales marcadores químicos de trazabilidad. En los jugos de limón la diferenciación entre regiones se estableció por los contenidos de Fe, La, V, Cu y Zn. En los de mandarina por los contenidos de Al, Ca, Cr, Cu, Fe, K, Mg, Mn, Sr y Zn. En los de naranja, cuando la información provino de FAAS, mediante Mn, Zn, Na y K y cuando los datos se obtuvieron mediante MP-AES estos elementos fueron Al, Ba, Ca, Cd, Cr, Co, Cu, K, Mg, Mn, Mo, Ni, Sr y Zn.

Se demostró la eficacia de métodos de análisis multivariados y de aprendizaje automático para proponer modelos para la autenticación o confirmación de identidad de los frutos. El orden de tasas de éxito de los métodos de clasificación para jugos de limón por provincia fue: SVM 100% > RF = LDA = PLS-DA = KNN 95%. En jugos de mandarina: ANN 96% > SVM 94% > DT 91% > LDA 86% > KNN 85%. En el caso de los jugos de naranja, las técnicas de clasificación probadas pueden ordenarse en: LDA 96 %

> ANN 92% > SVM = DT 90% > KNN 67% (información de FAAS) y SVM 99% > ANN = DT = LDA 99% > KNN 74% (información de MP-AES).

ABSTRACT

The most competitive citrus fruit markets need to know the geographical origin and identity of citrus fruits, for which the physical identity of the samples needs to be verified. The mineral composition of the plants is due to general patterns defined for species and varieties, but some variability is due to conditions of the sites where they grow, so mineral content would allow mathematical models to be designed to define their traceability.

This thesis was carried out with the objectives of assess the presence of chemical traceability markers in citrus fruit juices, using fingerprinting and machine learning techniques, with particular emphasis on determining multi-element chemical composition and demonstrating the ability of multivariate data analysis and machine learning methods to establish prediction models to authenticate or confirm the identity of lemon fruit juices produced in the NEA and NOA regions , and mandarin and orange produced in the NEA region.

Information derived from samples of lemon fruits (*Citrus limon* L., Osbeck) 'Eureka', 'Lisbon' and 'Genove' from four areas of NOA and NEA, tangerine (tangor) 'Murcott' (*C. sinensis* L. x C. Reshni) and 'Okitsu' (*C. unshiu* Marc.) and sweet orange tree (*C. sinensis* L.) 'Valencia late' and 'Salustiana', four areas of the NEA, characterized by its multi-element composition.

Determination of element concentrations in digested samples of lemon juices was carried out by mass spectrometry by coupled plasma (ICP-MS), orange juices using atomic flame absorption spectroscopy (FAAS) and those of orange juices and mandarin by atomic spectrometry by microwave plasma (MP-AES). Analysis: variance (ANOVA) and Multivariate Variance (MANOVA) and Hottelling Test, Main Components (PCA), Linear Discriminant (LDA) and Partial Minimums (PLS DA), K Closest Neighbors (KNN),

Artificial Neural Networks (ANN), Vector Support Machines (SVM), and Random Forests (RF) were applied. Sensitivity, specificity, accuracy, and index. Analyses were performed with InfoStat 2020 and R 3.2.1.

Juices from different regions were characterized by their mineral content. In lemon concentrations of Fe, Zn and Rb greater than 10 μg/g, Al, Ba, Cu, Mn and Ni between 1 and 10 μg/g and La, Cr, Se, Li, Mo, Co, Sn, Sc, V and Bi less than 1 μg/g. Tangerine and orange had similar profiles, the most abundant element was K (about 1000 μg/g), Al, Mg and Ca > 10 μg/g, Mn, Cu, Zn and Sr between 1 and 5 μg/g and Cd, Cr and Fe < 1 μg/g. Elements considered harmful were described for orange juices. As established in the Argentine Food Code, contents of Cu in 'Valencia late' exceeded the highs and the Pb exceeded the highs in some samples, not on average. The levels of all elements were included below the maximums set by the Official Association of American Food Control (AFCO). The average Cd content of orange juices exceeded the maximum allowed by WHO and FAO.

Potential presence of traceability chemical markers was detected in NEA and NOA lemons and NEA tangerines and oranges. In lemon juices the differentiation between regions was established by contents of Fe, La, V, Cu and Zn. In tangerine the differentiation was based on contents of Al, Ca, Cr, Cu, Fe, K, Mg, Mn, Sr y Zn. In orange, when the information came from FAAS, through Mn, Zn, Na and K and when the data was obtained from MP-AES, by Al, Ba, Ca, Cd, Cr, Co, Cu, K, Mg, Mn, Mo, Ni, Sr and Zn.

The effectiveness of multivariate analysis and machine learning methods to propose models for authentication or confirmation fruit identity was demonstrated. The rank of success rate of classification methods for lemon juices by province was: SVM 100% > RF = LDA = PLS-DA = KNN 95%. In tangerine juices the rank of success rate was: ANN 96% > SVM 94% > DT 91% > LDA 86% > KNN 85%. In the case of orange juices, proven sorting techniques can be ordered: LDA 96% > ANN 92% > SVM 90% = DT 90% > KNN 67% (FAAS information) and SVM 99% > ANN 99% = DT 99% = LDA 99% > KNN 74% (MP-AES information).

CAPÍTULO I

PROBLEMA DE ESTUDIO

1.1. Introducción

Las frutas forman parte de los alimentos con mayor cantidad de nutrientes, sustancias protectoras y antioxidantes naturales, altamente beneficiosas para la salud humana, por lo que su inclusión en la dieta es común a todas las culturas y está generalizada en todos los países. Los cítricos, término que comprende a los frutos de diferentes especies del género *Citrus* (entre los que podemos mencionar principalmente limones, naranjas, pomelos y mandarinas), constituyen los tipos de frutas más consumidos en el mundo entero, por encontrarse entre los alimentos más accesibles (en términos de precios, disponibilidad en los diferentes mercados y presencia durante todo el año) y útiles en la dieta humana (fundamentalmente por sus altos contenidos de vitamina C, ácido fólico, fibra dietética y minerales, así como numerosos fitoquímicos, incluidos los flavonoides, aminoácidos, triterpenos, ácidos fenólicos y carotenoides) (de la Guardia *et al.*, 2005; Yi *et al.*, 2008; Roussos, 2011).

Los frutales del género *Citrus* presentan ciertas características generales comunes, son árboles o arbustos de hojas perennes, su altura puede oscilar entre los 5 y los 16 m (actualmente se combinan prácticas culturales y asociaciones de portainjertos y variedades que se traducen en plantas más pequeñas, que facilitan la realización de las tareas agrícolas y resultan más productivas). Sus tallos son erectos, muchos de ellos con ramas provistas de espinas y hojas con pecíolo alado. Las flores son muy fragantes, suelen estar reunidas en inflorescencias, generalmente en forma de corimbos, aunque más raramente aparecen aisladas. Presentan cinco pétalos y numerosos estambres. Destacan por sus grandes frutos carnosos que son hesperidios con piel gruesa, con un tamaño habitual entre los 3 y los 10 cm de diámetro. Florecen en primavera y los frutos se recogen desde fines del verano hasta el invierno (Agustí, 2010; Palacios, 2013).

En términos generales, se estima que la cantidad de agua necesaria para un huerto de cítricos oscila entre 9.000 y 12.000 m^3, por hectárea por año, lo que equivale a una precipitación anual de 900 a 1.200 mm, sin embargo, las precipitaciones mayores no son problemáticas siempre y cuando haya un buen drenaje del suelo. Se considera que la humedad relativa influye sobre la calidad de la fruta, los cítricos cultivados en regiones donde la humedad relativa es alta tienden a tener piel más delgada y suave, contienen mayor cantidad de jugo y son de mejor calidad, pero en casos extremos el exceso de humedad puede favorecer el desarrollo de enfermedades fungosas y de algunas plagas. El rango adecuado de humedad relativa puede considerarse entre 50% y 80%. En zonas ventosas es necesario establecer barreras o cortinas rompevientos para evitar la deshidratación, roturas de ramas, caída de flores, hojas y frutos, como así también contribuir al control de enfermedades y plagas. Las barreras deberán ser de árboles de crecimiento vertical, de rápido desarrollo, follaje denso y que no alberguen plagas y enfermedades comunes a los cítricos (Agustí, 2010; Palacios, 2013).

Los cítricos son originarios del sudeste asiático, de una vasta región comprendida entre 0° y 30° de latitud norte, que incluye al sur y sudeste de China, India, Myanmar, Tailandia, Filipinas, Borneo y Sumatra, entre otros, donde se concentra la mayor cantidad de especies cítricas y afines. En la actualidad, y como puede observarse en la Figura 1.1, su cultivo se ha extendido a todo el planeta, centralizado en dos franjas bien definidas, entre los 16° y 41° de latitud norte y entre los 11° y 35° de latitud sur, que abarcan, en los distintos continentes, desde California hasta Argentina, desde la Cuenca del Mediterráneo hasta Sudáfrica, y desde Japón hasta Australia (Palacios, 2013).

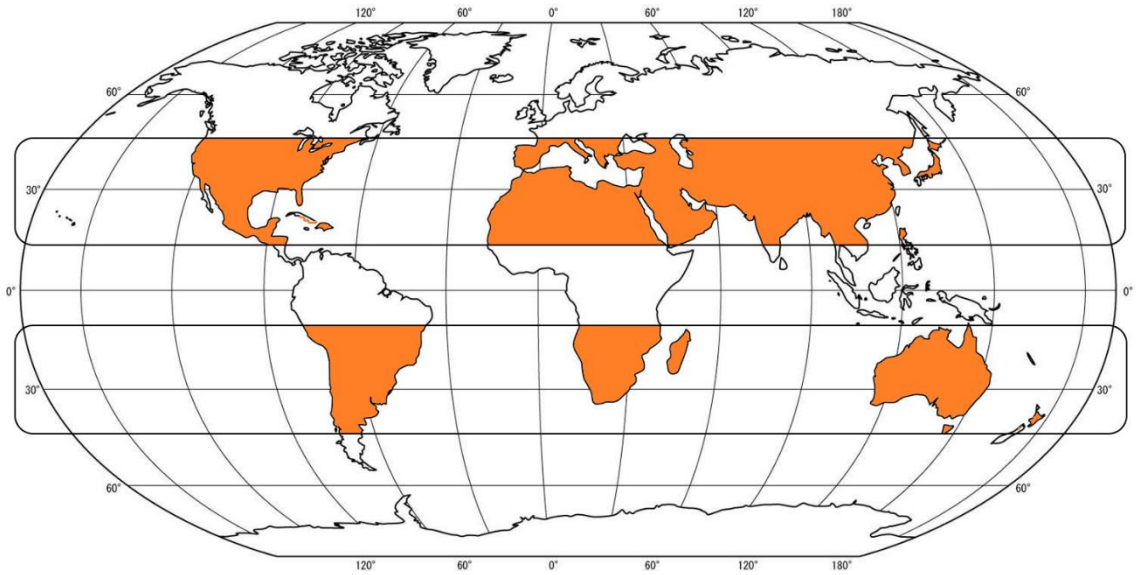

Figura 1.1. Distribución mundial de la producción de citrus, áreas productoras
* Fuente: elaboración propia

Estas especies se caracterizan por ser árboles de mediano porte, que mantienen en general una forma esférica o elíptica y presentan hojas verdes brillantes todo el año, tienen flores con perfume característico, tronco robusto y ramas principales engrosadas, bien distribuidas.

Los limoneros (*Citrus limon* L., Burm.) son árboles de hoja perenne de tronco delgado. Alcanzan una altura que oscila entre los 3 y los 6 m y poseen numerosas ramas delgadas, muchas de las cuales cuentan con espinas. Las hojas miden hasta 10 centímetros de largo, muestran una forma elíptica u ovalada y están ligeramente dentadas en los bordes, poseen un color verde oscuro en el haz y verde pálido en el envés. Las flores, son pequeñas y crecen solitarias o en grupos de 2 o más, con un aroma particular. Su color es blanco con un capullo violáceo, 5 sépalos cóncavos y 4 o 5 pétalos de unos 2 cm de longitud. Los frutos son ovalados y generalmente tienen una protuberancia en un extremo. Su piel es relativamente gruesa y su color es amarillo claro o verdoso, es áspera al tacto, debido a las pequeñas vesículas en la superficie. Por dentro, la pulpa es amarillo pálido a verde y se compone de sacos unidos entre sí rellenos de zumo. Son jugosos y ácidos, aunque por lo general presentan varias semillas, algunas variedades carecen de ellas (Agustí, 2010; Palacios, 2013).

El término mandarinos agrupa a varias especies (*C. unshiu* Marcovitch (Satsumas)*, C. nobilis* Loureiro (King), *C. deliciosa* Tenore (Común), *C. Clementina* Hort. ex Tanaka (Clementinas), *C. reticulata* Blanco (Dancy, Ponkan), *C. reshni* Hort. ex Tanaka (Cleopatra), *C. sunki* Hort. ex Tanaka (Sunki y Suenkat)). Los árboles poseen un tamaño menor al de los naranjos y una forma más redondeada en su copa, su envergadura va desde los 2 a 4 m de altura. Esto los convierte, además de proveedores de agradables frutos, en árboles ornamentales en los jardines. Las hojas son algo más alargadas que las del naranjo y presentan flores pequeñas blancas que pueden crecer en pequeños ramilletes de dos o tres unidades y a veces se presentan solas. Sus frutos son de color naranja intenso y su forma varía de acuerdo con cada variedad, pero en general son más pequeñas que una naranja. La cáscara se desprende fácilmente del resto del fruto, donde encontramos numerosos gajos fáciles de separar entre ellos y provistos de gran cantidad de jugo, siendo su sabor dulce y agradable (Palacios, 2013).

El naranjo dulce (*C. sinensis* L.) es el cítrico más cultivado en el mundo, se caracteriza por sus plantas de tamaño mediano a grande, hojas brillantes verde oscuro, prácticamente sin espinas o con muy pocas. Los frutos del naranjo dulce son de tamaño mediano, con o sin semillas, jugo agridulce con 0,8 a 1% de acidez. Constan de un exocarpio, llamado flavedo, que presenta vesículas que contienen aceites esenciales; el mesocarpio, denominado albedo, de color blanco; y el endocarpio, que presenta los tricomas que contienen el jugo. Son frutas hipocalóricas e hiposódicas compuestas principalmente por agua. Su contenido de grasa, proteínas y fibra es muy bajo, siendo los hidratos de carbono el segundo componente de mayor presencia. Asimismo, se caracterizan por su alto contenido de vitamina C y aportes de vitaminas B_1, B_2 y provitamina A. En general, por su época de maduración, a las variedades de naranjas se las clasifica en tempranas, intermedias y tardías, de ello dependerá el momento de cosecha (Agustí, 2010; Palacios, 2013).

La producción mundial de frutas se encuentra en aumento. En el año 2017 fue de 1.012.189.228 t, lo que significa un incremento del 0,72% respecto a 2016 (producción de 1.004.885.444 t). Argentina participó con 11.034.721 t, que corresponden a un 1,1% del total. El 15% de la producción mundial de frutos corresponde a cítricos, que se

mantiene alrededor de las 146.599.168 t, entre las que 73.313.089 t corresponden a naranjas (50%), 17.218.173 t a limones y limas (11,74%), 33.414.126 t a mandarinas, tangerinas, clementinas y satsumas (22,79%), 9.063.143 t a pomelos y toronjas (6,18%) y 13.590.637 a otras frutas cítricas (9,27%). La producción argentina para el año 2017 fue de 3.272.771 t y representa el 2,22% de la producción mundial de cítricos, el 51,22% corresponde a limones y limas, el 31,22% a naranjas, el 14,02% a mandarinas, tangerinas, clementinas y satsumas, el 3,42% a pomelos y toronjas (Federcitrus, 2018; FAO, 2017).

La República Argentina tiene tradición citrícola y se destaca por su producción de limones, mandarinas, naranjas y, en menor medida, pomelos. La fruta producida permite abastecer la demanda del país durante todo el año y es muy oportuna para las exportaciones en contra estación a los países del hemisferio norte. En el año 2017 la Argentina se ubicó octava a nivel mundial en cuanto a la producción de frutos cítricos, con un total de 3.272.771 t. Alrededor del 27% de la producción se destinó al consumo interno como fruta fresca, cerca del 46% fue destinado a la industrialización (incluye la producción de jugos y esencias, entre otros), y aproximadamente un 11% fue exportado a diferentes mercados, existe además un 16% de pérdida atribuible a diversos factores. Los principales compradores de los cítricos argentinos son: España (18,77%), Rusia (18,45%) y Holanda (10,75%)[1]. Entre los destinos regionales, Paraguay se ubica en el primer lugar (8,25%) (SENASA, 2014).

En la Tabla 1.1 se presenta la producción mundial de citrus agrupadas por países, en las campañas 2016/2017 y 2017/2018, donde se observa claramente que los principales países productores de frutas cítricas, con producciones anuales superiores a los 10 millones de toneladas son China y Brasil (en orden decreciente).

[1]https://www.magyp.gob.ar/sitio/areas/ss_mercados_agropecuarios/areas/frutas/_archivos/000030_Informes/000029_Anuario%20de%20Frutas%20-%202019.pdf

Tabla 1.1. Producción mundial de citrus por país, campañas 2016/2017 y 2017/2018 (miles de t)

País	Campaña 2016/2017	Campaña 2017/2018
China	32.200	33.300
Brasil	20.400	17.340
México	7.584	7.620
España	7.248	6.357
EE. UU.	7.045	5.687
Egipto	4.110	4.295
Turquía	3.975	4.065
Argentina	3.273	3.284
Sudáfrica	2.734	2.400
Italia	2.603	2.379
Marruecos	2.330	1.993
Grecia	1.131	1.116
Perú	1.096	1.190
Japón	1.070	990
Australia	674	674
Corea del Sur	600	560
Israel	569	593
Vietnam	590	590
Chile	491	S/D
Costa Rica	322	325
Uruguay	264	270
Guatemala	175	175
Chipre	105	100
Otros	434	415

*Fuente: Federcitrus (2018)

Los principales destinos de la exportación de frutos cítricos argentinos se resumen en la Tabla 1.2 (campaña 2016/17) agrupados por especie. Se observa que los limones fueron la fruta mayormente exportada por Argentina seguidos por las naranjas y en menor medida mandarinas.

Tabla 1.2. Destino de las exportaciones argentinas de frutos cítricos (t), campaña 2016/2017, discriminado por especie

País	Limones	Naranjas	Mandarinas	Pomelos	Totales	Porcentajes
España	38.588	27.914	73	135	66.711	18,77
Rusia	42.295	3.134	20.024	111	65.564	18,45
Holanda	31.006	6.681	511	21	38.219	10,75
Italia	31.246	3.516	-	21	34.783	9,79
Paraguay	-	28.351	895	182	29.429	8,25
Canadá	10.298	1.760	5.185	116	17.360	4,88
Grecia	12.141	-	-	-	12.141	3,42
Ucrania	9.951	437	84	58	10.531	2,96
Gran Bretaña	7.221	106	643	-	7.970	2,24
Francia	7.160	234	-	-	7.395	2,08
Filipinas	71	578	6.509	-	7.158	2,01
Polonia	6.643	-	-	-	6.643	1,87
Alemania	5.084	96	-	-	5.180	1,46
Rumania	4.336	-	-	-	4.336	1,22
Dinamarca	3.017	-	-	-	3.017	0,85
Malasia	1.653	293	938	-	2.829	0,80
Indonesia	1.146	143	1.444	-	2.734	0,77
Eslovenia	2.578	48	-	-	2.626	0,74
Bélgica	2.181	188	-	-	2.369	0,67
Emiratos Árabes	517	521	1.325	-	2.363	0,66
Arabia Saudita	87	1.448	797	-	2.331	0,66
Singapur	1.229	457	573	-	2.259	0,64
Portugal	1.306	639	256	-	2.201	0,62
Kosovo	1.904	48	-	-	1.952	0,55
Hong Kong	1.214	68	261	-	1.543	0,43
Resto del mundo	12.380	2.366	941	63	15.749	4,43
Total	235.254	78.972	40.461	708	355.396	

*Fuente: Federcitrus 2018

En Argentina la producción de cítricos data de varios siglos, hacia mediados del siglo XVIII, ya existían importantes fincas productoras de naranjas. Datos de 2017 indican que la actividad citrícola concentra alrededor de 5.300 productores que, si bien se encuentran distribuidos en 10 provincias, el 75% se concentra en 4 de ellas. Incluye también un total de 330 empaques de frutas cítricas, de los cuales 75 son específicos para exportación y, para el procesamiento de la fruta, 22 plantas industriales. El subsector citrícola en general, ocupa alrededor de 100.000 trabajadores directos (Federcitrus, 2018).

Argentina posee varias zonas con condiciones ecológicas ideales para el desarrollo de la producción de cítricos, los cultivos están situados en lugares privilegiados de América del Sur, entre el trópico de Capricornio y el paralelo de 35°S

(Figura 1.2). El desarrollo de los cultivos de citrus en Argentina se extiende, principalmente a dos regiones: el Noroeste Argentino (NOA), donde se producen naranjas, pomelos y limones (especialmente en la provincia de Tucumán), y el Nordeste Argentino (NEA), donde predominan los cultivos de naranjas y mandarinas que, a través de diversas variedades orientadas a los gustos de los distintos mercados, se cosechan y exportan a lo largo de casi todo el año.

En el NOA, Tucumán concentra una de las principales regiones productoras de limones del mundo, centraliza el 72,9% de la superficie nacional dedicada al cultivo de limonero y el 30,3 % del total de la superficie nacional implantada con citrus. Salta presenta la mayor superficie productora de pomelos, con el 33,3% del total nacional de este cultivo y el 9,9% del total nacional de superficie citrícola. Jujuy presenta el 6,1% de la superficie citrícola nacional, siendo su principal producción las naranjas, que representan el 54,9% de la superficie citrícola provincial[2].

En la Región NEA, las provincias de Entre Ríos, Corrientes, Misiones y Formosa contribuyen con el 37,2% de la producción citrícola total del país. Si se analiza por especie, en el NEA se produce el 67,0% de las naranjas, el 87,5% de las mandarinas, el 5,1% de los limones y el 30,9% de los pomelos del país[3].

En el NEA, Entre Ríos es la principal provincia productora de naranjas y mandarinas, con el 42,0% y 53,0%, respectivamente, de la superficie nacional dedicada a esos cultivos. Concentra el 26,1% de la superficie dedicada al cultivo de cítricos en el país. Corrientes es la segunda productora de naranjas del país, con el 29,6% de la superficie total nacional y una superficie total de cítricos que representa el 19,0% del total de la superficie citrícola de Argentina. Misiones aporta el 2,3% de los cítricos producidos en el país, mandarino es la principal especie con el 30,5% de la superficie citrícola provincial. Formosa es la principal productora de pomelos de esta región, concentra el 21,7% de la superficie implantada con este cultivo en el país (Federcitrus, 2018).

[2]http://www.alimentosargentinos.gob.ar/HomeAlimentos/Cadenas%20de%20Valor%20de%20Alimentos%20y%20Bebidas/informes/LIMON_Resumen_Cadena_Septiembre_2019.pdf
[3]http://www.senasa.gob.ar/sites/default/files/ARBOL_SENASA/INFORMACION/INFORMES%20Y%20ESTADISTICAS/Informes%20y%20estadisticas%20Vegetal/FRUTALES/citricos_argentinos_de_excelencia.pdf

Figura 1.2. Zonas citrícolas de la República Argentina
*Fuente: Federcitrus 2018

1.2. Planteamiento del problema

En la actualidad, los integrantes de la cadena agroalimentaria se interesan por conocer y garantizar el origen geográfico e identidad de los productos agropecuarios y, en especial, de aquellos que intervienen en la producción de alimentos derivados. La certificación de origen puede ser un elemento esencial para asegurar la autenticidad de un determinado producto alimentario dado que protege, sobre todo, a los productos regionales y confirma características de calidad relacionadas con su origen (Drivelos & Georgiou, 2012; Luykx & van Ruth, 2008).

Debido a las exigencias de los mercados de exportación, los productores y las autoridades sanitarias de Argentina decidieron implementar un sistema de trazabilidad en el marco del Programa de Certificación de Cítricos de Exportación a la Unión

Europea y otros mercados con similares restricciones cuarentenarias. Luego de varios años de trabajo de planificación, tareas en las quintas y en las plantas de empaque, el sistema ha sido puesto en operación. Desde las platas de empaque se informa directamente al Sistema de Información sobre Trazabilidad Citrícola del NEA (SITC®-NEA). El Servicio Nacional de Sanidad y Calidad Agroalimentaria (SENASA) puede controlar en puerto, dado que el SITC®-NEA, opera en la red global[4]. Los importadores europeos y de otros países pueden conocer, mediante el acceso libre al sitio, quién transportó, quién exportó, quién despachó, quién produjo, de qué establecimiento y de qué lote de éste se originó la fruta cítrica. El SITC®-NEA es el primer sistema de información de Argentina, y quizás de otros países, que permite conocer todo el proceso de un producto, en este caso de los cítricos, desde el campo hasta el destino final.

Si bien se ha desarrollado el sistema registral SITC®, para lograr la trazabilidad de los cítricos producidos, el mismo implica documentar los procesos desde la cosecha hasta el destino final. Estos sistemas se basan en información documental y no contemplan mecanismos que permitan comprobar la identidad física de las muestras en cualquier etapa de la cadena productiva, por lo que pueden ser vulnerables a pérdida de información o maniobras de adulteración o contaminación. Resolver este problema implica la necesidad de contar con un mecanismo de identificación de las muestras físicas, que podría estar basado en la composición química de las mismas.

Se plantea, entonces, la necesidad de definir estrategias para establecer la trazabilidad y rastreabilidad de frutos cítricos, aunque no se disponga de información de toda la cadena de producción, procesamiento y comercialización, o complementariamente a ella. Debido a ello, este trabajo de tesis se orientó a la generación de conocimiento sobre variables químicas que aporten información para los sistemas de trazabilidad de cítricos argentinos vigentes.

1.3. Trazabilidad de los alimentos

[4] https://www.kyas.com.ar/producto/sitc

Desde mediados del siglo pasado se viene estableciendo una visión de cadena de abastecimiento, que se sintetiza finalmente en una visión sistémica, la expresión “gestión de la cadena de abastecimiento” (SCM en inglés) aparece en la década de los ochenta, evoluciona y se instala rápidamente (Melnyk *et al.*, 2009).

El interés de los consumidores por conocer el origen geográfico de los productos agropecuarios y en especial de los alimentos, ha crecido y adquirido una importancia trascendental en muchos países del mundo, por lo que la falta de registro en los procesos de la cadena de suministro ocasiona pérdida de oportunidades de comercialización en mercados más competitivos. El seguimiento y control en los procesos de almacenamiento, distribución y transformación de la cadena de suministro es primordial para garantizar la calidad de los alimentos, especialmente en procesos con alta variabilidad. En este sentido, las tecnologías de trazabilidad permiten el control y seguimiento en los procesos de la cadena de suministro de alimentos (Drivelos & Georgiou, 2012; Luykx & van Ruth, 2008).

Un producto confiable para un consumidor debe ser certificado y esa garantía debe estar presente en la etiqueta de venta de dicho producto. El sello representa para el consumidor, en términos de calidad, que el producto está "trazado", dado que está explícitamente descripto, es confiablemente controlado, está sistemáticamente verificado y es pasible de sanción para el caso de no cumplir con lo especificado. Es decir, se conoce la procedencia, los procesos y el destino del producto. La “trazabilidad” se puede definir como el conjunto de acciones, medidas y procedimientos técnicos que permite identificar y registrar un producto desde su origen hasta el final de la cadena de comercialización. La trazabilidad permite rastrear la cadena de producción, los procedimientos mediante los que se obtuvieron dichos productos y facilita el ingreso de estos a mercados específicos, más rentables, que actualmente exigen conocer de manera certera el origen del producto (Bertaccini *et al*., 2013; Bosona & Gebresenbet, 2013; Felmer *et al.*, 2006).

La experiencia ha demostrado que la imposibilidad de localizar el origen de los alimentos, además, puede poner en peligro el funcionamiento del mercado interior de alimentos. Es por tanto necesario establecer un sistema exhaustivo de trazabilidad en

las empresas alimentarias para poder proceder a retiradas específicas y precisas de productos, o bien informar a los consumidores o a los funcionarios encargados del control, y evitar así una mayor perturbación innecesaria en caso de problemas de seguridad alimentaria (Parlamento Europeo, 2002).

Los sistemas de trazabilidad tienen por objetivos, la Certificación de Procesos de Producción (EE. UU.) y la Seguridad Alimentaria (Unión Europea y la Organización de Naciones Unidas para la Agricultura y la Alimentación).

En los Estados Unidos, la ley de bioterrorismo (2002) establece que la persona que fabrica, procesa, empaca, transporta, distribuye, recibe, posee o importa alimentos a Estados Unidos, tiene la responsabilidad de establecer y mantener registros y el órgano de control encargado de realizar inspecciones en caso de sospechas razonables es la Food and Drug Administration (FDA). En el año 2011, a través de la ley de modernización de seguridad alimentaria de la FDA (Food Safety Modernization Act, FSMA), se establece un nuevo sistema de supervisión de la inocuidad de los alimentos, centrado en aplicar de forma integral los mejores recursos científicos disponibles para prevenir los problemas que pueden causar enfermedades en la gente. Considera el sistema alimentario en su totalidad y por tanto el concepto de responsabilidad de todos sus participantes de la cadena de suministro. La FDA puede verificar y bloquear los alimentos de establecimientos o países que se nieguen a permitir inspección. Busca garantizar la seguridad de oferta de alimentos importados y nacionales, centrándose en la prevención de la contaminación (Rincón Ballesteros, 2016).

Los conceptos de seguridad alimentaria de la Unión Europea y el de la Organización de Naciones Unidas para la Agricultura y la Alimentación (FAO) difieren entre sí. Para la FAO “existe seguridad alimentaria cuando todas las personas tienen en todo momento acceso físico, social y económico a suficientes alimentos inocuos y nutritivos para satisfacer sus necesidades alimenticias y sus preferencias en cuanto a los alimentos a fin de llevar una vida activa y sana. Los cuatro pilares de la seguridad alimentaria son la disponibilidad, el acceso, la utilización y la estabilidad. La dimensión nutricional es parte integrante del concepto de seguridad alimentaria” (FAO, 2009).

Para la Unión Europea la seguridad alimentaria sólo incluye al pilar de la salud y seguridad de los consumidores, afectando a la normativa relativa a la higiene de los productos alimenticios, a la salud y bienestar de los animales, a la sanidad vegetal, a la prevención de los riesgos de contaminación por sustancias externas y al etiquetado adecuado de dichos productos.

La FAO declaró que la gestión de la seguridad y calidad de los alimentos es una responsabilidad compartida de todos los actores de la cadena alimentaria, incluidos los gobiernos, la industria y los consumidores, pero al no existir un consenso común o estandarización de criterios, no se pueden realizar acciones conjuntas (FAO, 2003). En 2002 la Food Standards Agency (FSA) explica que los sistemas de trazabilidad habían adquirido una importancia considerable en lo que respecta a la alimentación, después de una serie de incidentes de seguridad alimentaria en la que los sistemas de trazabilidad habían demostrado ser débiles o ausentes; razón por la cual hoy en día los consumidores demandan evidencia verificable de la trazabilidad y rastreabilidad actuando como un criterio importante en la calidad y seguridad alimentaria (Aung & Chang, 2014). Se pone de manifiesto que la reconstrucción de la confianza pública en la cadena alimentaria está centrada en el diseño y la aplicación de la trazabilidad en toda la cadena productiva, desde la granja hasta el usuario final, para poder suministrar a los consumidores alimentos de alta calidad, inocuos y nutritivos (Opara, 2003).

La Unión Europea en su ley general de alimentos establece que debe existir un sistema de trazabilidad obligatoria para todos los alimentos y piensos que se venden en países de la Unión Europea, expresa que el detalle de la trazabilidad ha de extenderse también a cada ingrediente de la comida y debe figurar en los sistemas de etiquetado de productos alimenticios [5].

Dentro de las normas que tienen influencia mundial se pueden considerar las ISO específicamente 9001 (2005), 22000 (2005), estas establecen requisitos especificados para un sistema de gestión de seguridad alimentaria de una organización en la cadena alimentaria, con el fin de demostrar su capacidad para controlar los peligros de

[5] https://europa.eu/european-union/topics/food-safety_es

inocuidad. Estos estándares incluyen los métodos de análisis de los riesgos alimentarios de HACCP y el enfoque del sistema de gestión de la norma ISO 9001. Se cuenta con normativa local en cada país y a nivel mundial como el *Codex Alimentarius* que promueve la coordinación de todos los trabajos sobre normas alimentarias emprendidos por las organizaciones internacionales gubernamentales y no gubernamentales [6] [7] [8].

Existen, además, normas comerciales propuestas por diferentes actores dentro del sistema de comercio internacional, entre ellas GS1 (Estados Unidos), GAP (GlobalGAP), Guías de buenas prácticas para la rastreabilidad (British Retail Consortium), estas describen procesos de negocios en las que se detallan la cadena de trazabilidad y facilitan el intercambio de datos de trazabilidad, adoptando estándares de identificación de productos con fines comerciales (Rincón Ballesteros, 2016).

Las normas de trazabilidad surgen como consecuencia de las exigencias de los consumidores que, en los últimos 20 años, dieron mayor importancia a la seguridad alimentaria. Los consumidores de mayor poder adquisitivo dan prioridad a factores no necesariamente económicos, como ser: que el producto sea identificable desde el origen, que el producto sea diferenciable con respecto a productos alternativos o sustitutos, que sea seguro en términos de salud (evitar enfermedades, intoxicaciones, etc), que sea saludable para la dieta (nivel de grasa, vitaminas, proteínas, etc.) y que sea conveniente en términos de comodidad y simplicidad de cocción.

1.4. Composición química de los vegetales

Los vegetales requieren para su crecimiento y metabolismo el abastecimiento y absorción de compuestos químicos que se denominan nutrientes. El 90-95% del peso

[6] http://www.fao.org/3/i7407es/I7407ES.pdf
[7] https://www.who.int/foodsafety/areas_work/food-standard/es/
[8] http://www.anmat.gov.ar/webanmat/normativas_alimentos.asp

seco del material vegetal está constituido por C, O e H, que son los principales constituyentes de los compuestos orgánicos, y el 5-10% restante corresponde a otros elementos cuya presencia es esencial para completar su desarrollo normal y su ciclo biológico.

Debido a su papel fisiológico se les llama elementos esenciales y se definen como aquellos sin los cuales las plantas no pueden completar su ciclo de vida, son irremplazables por otros elementos, y están involucrados directamente en el metabolismo de la planta. De acuerdo con la concentración en que son requeridos por la planta, se clasifican en macro y micronutrientes (elementos a nivel de vestigios). Los elementos a nivel de vestigios son aquellos que están presentes a bajas concentraciones (mg/kg o menor) en la mayoría de los suelos, plantas, y demás organismos vivos (Adriano, 2001; Fageria *et al.*, 2002, Kabata Pendias, 2010).

A pesar de que los micronutrientes se requieren en pequeñas cantidades por las plantas, su influencia es tan importante como los macronutrientes. Esta clasificación tiene una validez relativa, ya que, en algunos casos, ciertos macronutrientes se acumulan en cantidades menores que otros micronutrientes. Por otro lado, además de los nutrientes se pueden encontrar en la planta otros elementos, sin función biológica conocida hasta ahora, y otros cuya presencia conduce a disfunciones, ya que tienden a acumularse y son altamente tóxicos (Bi, Cd, Hg, Pb y Sn).

La nutrición vegetal estudia el conjunto de procesos mediante los cuales los vegetales toman sustancias del exterior para sintetizar sus componentes celulares o usarlas como fuente de energía, es decir se relaciona con el abastecimiento y absorción de los compuestos químicos necesarios para el crecimiento y metabolismo de plantas. Si bien la composición mineral de los vegetales obedece a patrones generales definidos para las diferentes especies y variedades, existe cierto grado de variabilidad que se debe, en gran medida, a condiciones de los sitios donde las plantas crecen. El suelo es uno de los elementos clave en todos los ecosistemas terrestres, en los que cumple funciones esenciales de soporte físico para el crecimiento de las plantas y fundamentalmente les proporciona el agua, el aire y los nutrientes esenciales para su desarrollo, es el factor principal que controla el flujo del agua en el ciclo

hidrológico, así como de las especies químicas dentro de los ciclos biogeoquímicos (Weil & Brady, 2017).

Esta influencia de las condiciones locales en la composición de los tejidos vegetales ha permitido que la determinación de diversos compuestos orgánicos haya sido utilizada para caracterizar el origen específico de productos vegetales. Sin embargo, el rango normal de compuestos orgánicos en los alimentos varía con algunas prácticas agrícolas como la fertilización, las condiciones climáticas en el año de cultivo, la historia de los campos, así como la ubicación y características del suelo, por lo que a veces es difícil ser contundente acerca de la autenticidad de origen a partir de la determinación de los componentes orgánicos. Por otro lado, el contenido de compuestos orgánicos es susceptible de sufrir variaciones o descomposiciones en el tiempo, de acuerdo con las condiciones ambientales de almacenamiento de las muestras, resultando su determinación válida solamente en determinados intervalos de tiempo. Por su parte, el contenido de minerales es una excelente alternativa debido a la correlación entre elementos a nivel de vestigios, las condiciones de cultivo y ambientales en origen (Gonzálvez & de la Guardia, 2013), además de que sus niveles se mantienen constantes independientemente de la alteración de las muestras.

La composición química de los vegetales, incluyendo la concentración de distintos elementos inorgánicos, se ve afectada por diversos factores como: especie botánica, variedad, naturaleza química y tipo de suelo, acidez del suelo, disponibilidad de agua y nutrientes y condiciones climáticas, entre otros. Es decir que la composición mineral de los vegetales se verá fuertemente influenciada por las condiciones en las que fueron producidas (Watson *et al.*, 2012). Debido a ello, las diferencias en los contenidos minerales pueden ser aprovechadas desde el punto de vista quimiométrico para proponer modelos matemáticos que permitan determinar su trazabilidad (Ariyama & Yasui, 2006; Forina *et al.*, 2009).

Los elementos entran a un sistema agrícola a través de procesos naturales y antropogénicos. El suelo hereda elementos de la roca madre y por otro lado los procesos antropogénicos implican la entrada de elementos a través del uso de fertilizantes, desechos orgánicos, desperdicios industriales y municipales, irrigación y

depósitos húmedos o secos. Estos procesos contribuyen a generar valores diferentes de concentración en los elementos en el sistema. Por otra parte, según la especie química en que se encuentren presentes los elementos en las diferentes fases del suelo, así será la disponibilidad relativa para las plantas y, por tanto, su incorporación a la biomasa, lo que finalmente se va a ver reflejado en las variaciones en la composición de los vegetales que crecen en ellos[9].

El análisis de la composición mineral ha sido utilizado en diversos estudios para determinar la procedencia geográfica de diferentes alimentos, entre los que se puede destacar la determinación de la procedencia geográfica de vino (Dutra *et al.*, 2013; Geana *et al.*, 2013) mieles (Batista *et al.*, 2012; Di Bella *et al.*, 2015), arroz (Cheajesadagul *et al.*, 2013; Chung *et al.*, 2015; Maione *et al.*, 2016) y leche (Magdas *et al.*, 2016), entre otros.

1.5. Huella dactilar

La evaluación de los contenidos de elementos a nivel de vestigio ha sido propuesta para determinar el origen geográfico de las muestras. El término técnicas de "huella dactilar" describe a una variedad de métodos analíticos que pueden medir la composición de productos alimentarios de una manera no selectiva, esto es, mediante la colección de un espectro, cromatograma o datos de composición multielemental. El procesado matemático de la información contenida en esas "huellas dactilares" permitiría la caracterización del producto alimentario. Los métodos que pueden proporcionar un perfil mineral característico se pueden usar para la determinación del origen geográfico y, por lo tanto, proporcionar una herramienta valiosa para la autenticación y trazabilidad de alimentos (Esslinger *et al.*, 2014).

9 http://www.fao.org/3/w1309s/w1309s04.htm#:~:text=Los%20minerales%20provienen%20de%20la,de%20vegetales%20y%20animales%20muertos.

La evaluación de los contenidos de elementos a nivel de vestigios ha sido propuesta para determinar el origen geográfico de las muestras. Es así como técnicas de un solo elemento (espectrometría de absorción atómica y espectrometría de absorción atómica electrotérmica) y técnicas multi-elementales (espectrometría de masas con plasma acoplado inductivamente o espectrometría de emisión óptica con plasma acoplado inductivamente) se han empleado con éxito en la determinación del origen geográfico de material vegetal o en la autenticación de productos designados de origen (Gonzálvez & de la Guardia, 2013).

Las técnicas de huella dactilar proporcionan una gran cantidad de información para cada muestra analizada, se destacan por su alta precisión y bajos límites de detección, analizando la mayoría de los elementos e isótopos presentes en la tabla periódica de manera simultánea en un breve lapso. Esta capacidad de determinar de forma simultánea la concentración de numerosos elementos químicos, permite obtener una gran cantidad de información con una única inyección de muestra.

1.6. Objetivos

En la propuesta original de esta tesis, se planteaba trabajar con muestras de frutos de naranjo dulce (*Citrus sinensis* L.) de distintas procedencias de la región del NEA. Sin embargo, al realizar los trabajos de campo fue factible recoger muestras y obtener información de otras especies de cítricos de importancia en el país, por lo que se trabajó con información empírica derivada de un conjunto de muestras de frutos frescos de las variedades, de limonero (*C. limon* L., Osbeck) 'Eureka', 'Lisboa' y 'Génova' procedentes de localidades ubicadas en las regiones NOA y NEA, de mandarino (tangor) 'Murcott' (*C. sinensis* L. x *C. reshni*) y 'Okitsu' (*C. unshiu* Marc.) y naranjo dulce (*C. sinensis* L.) 'Valencia late' y 'Salustiana', de distintas procedencias de la región del NEA.

Estas muestras fueron caracterizadas desde el punto de vista físico químico y por su composición multielemental, y analizadas mediante métodos quimiométricos, con la finalidad de generar conocimientos sobre variables químicas que aporten información para los sistemas de trazabilidad de cítricos argentinos vigentes, para lo que se plantearon los objetivos e hipótesis que se presentan a continuación.

1.6.1. Objetivos Generales

Evaluar la presencia de marcadores químicos de trazabilidad en jugos de frutas cítricas, mediante técnicas de huella dactilar y de aprendizaje automático, con especial énfasis en la determinación de la composición química multielemental.

Demostrar la capacidad de métodos de análisis de datos multivariantes y de aprendizaje automático para establecer modelos de predicción para autenticar o confirmar la identidad de jugos de frutos de limón producidos en las regiones NEA y NOA, y de mandarina y naranja producidas en la región NEA.

1.6.2. Objetivos Particulares

- Contribuir al conocimiento de la composición química inorgánica de jugos de frutos de limón, mandarina y naranja.
- Aplicar métodos analíticos sensibles y selectivos para la determinación multielemental, de modo de explorar la presencia de marcadores químicos de identidad en jugos de frutos de limón producidos en las regiones NEA y NOA, y de mandarina y naranja producidas en la región NEA.

- Proponer modelos que permitan establecer la trazabilidad química de las muestras mediante el uso de las llamadas técnicas de huella dactilar y de herramientas de análisis multivariado de datos y de aprendizaje automático.

- Generar información de los productos alimenticios objeto de este estudio sobre su composición química inorgánica (presencia de elementos de interés ambiental, toxicológico y/o nutricional).

1.7. Hipótesis

Es posible establecer modelos estadísticos multivariados y de aprendizaje automático basados en la composición mineral (macroelementos, microelementos y elementos a nivel de vestigios) determinada por técnicas analíticas espectrométricas, que contribuyan a establecer sistemas de trazabilidad química de jugos de frutos cítricos.

1.8. Referencias

Aceto, M. 2016. 8 - The Use of ICP-MS in Food Traceability, Editor(s): Montserrat Espiñeira, Francisco J. Santaclara, In Woodhead Publishing Series in Food Science, Technology and Nutrition, Advances in Food Traceability Techniques and Technologies, Woodhead Publishing, 2016, Pages 137-164.

Adriano, DC. 2001. Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals. Springer Science & Business Media.

Agustí, M. 2010. Citricultura. Ediciones Mundi-Prensa, Madrid, España. pp. 507.

Ariyama, K; Yasui, A. 2006. The determination technique of the geographic origin of welsh onions by mineral composition and perspectives for the future. Japan Agricultural Research Quarterly. 40: 333-339.

Aung, MM; Chang, YS. 2014. Traceability in a food supply chain: Safety and quality perspectives. Food Control. 39: 172-184.

Balcaen, L; Bolea-Fernández, E; Resano, M; Vanhaecke, F. 2015. Inductively coupled plasma - Tandem mass spectrometry (ICP-MS/MS): A powerful and universal tool for the interference-free determination of (ultra)trace elements – A tutorial review. Analytica Chemica Acta. 894: 7-19.

Batista, L; da Silva, L; Rocha, B; Rodríguez, J; Berretta-Silva, A; Bonates, T; Gomes, V; Barbosa, R; Barbosa, F. 2012. Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. International Food Research. 49: 209-215.

Bertaccini, L; Cocchi, M; Li Vigni, M; Marchetti, A; Salvatore, E; Sighinolfi, S; Silvestri, M; Durante, C. 2013. The impact of chemometrics on food traceability. Chemometrics in Food Chemistry. Marini Ed. Pp 371-410.

Bosona, T; Gebresenbet, G. 2013. Food traceability as an integral part of logistics management in food and agricultural supply chain. Food Control. 33: 32-48.

Camuña Aguilar, JF. 1994. Desarrollo de plasmas inducidos por microondas como fuentes de excitación espectro-química y su aplicación al análisis de elementos traza. Tesis Doctorado en Química. Universidad de Oviedo.

Cheajesadagul, P; Arnaudguilhem, C; Shiowatana, J; Siripinyanond, A; Szpunar, J. 2013. Discrimination of geographical origin of rice based on multi-element fingerprinting by high resolution inductively coupled plasma mass spectrometry. Food Chemistry. 141: 3504-3509.

Chung, Y; Kim, J; Lee, J; Kim, S. 2015. Discrimination of geographical origin of rice (Oryza sativa L.) by multielement analysis using inductively coupled plasma atomic emission spectroscopy and multivariate analysis. Journal of Cereal Science. 65: 252-259.

De la Guardia, M; Trípoli, E; Giammanco, S; Finotti E. 2005. Flavanones in Citrus fruit: Structure–antioxidant activity relationships. Food Research International. 38 (10): 1161-1166.

Di Bella, G; Lo Turco, V; Potortí, A; Bua, G; Fede, M; Dugo, G. 2015. Geographical discrimination of Italian honey by multielement analysis with a chemometric focus. Journal of Food Composition Analysis. 44: 25-35.

Djedjibegovic, J; Larssen, T; Skrbo, A; Marjanovic, A; Sobriedad M. 2005. Contents of Cd, Cu, Hg and Pb in fish of river Neretva (Bosnia y Herzegovina) determined by inductively coupled plasma spectrometry (ICP-MS). Food Chemistry. 131 (2): 469-476.

Drivelos, SA; Georgiou, CA. 2012. Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. Trends in Analytical Chemistry. 40: 38-51.

Dutra, S; Adami, L; Marcon, A; Carnieli, G; Roani, C; Spinelli, F; Leonardelli, S; Vanderlinde, R. 2013. Characterization of wines according to the geographical origin by analysis of isotopes and minerals and the influence of harvest on the isotope values. Food Chemmistry. 141: 2148-2153.

Esslinger, S; Riedl, J; Fauhl-Hassek, C. 2014. Potential and limitations of non-targeted fingerprinting for authentication of food in official control. Food Research International. 60: 189-204.

Fageria, N; Baligar, V; Clark, R. 2002. Micronutrients in Crop Production. In: Advances in Agronomy. Academic Press. 185-268.

FAO. 2003. Estrategia de la FAO para un enfoque de la cadena alimentaria para la inocuidad y la calidad de los alimentos: Un documento marco para el desarrollo de la futura dirección estratégica. Disponible en línea: http://www.fao.org/docrep/meeting/006/y8350e.htm. Visita: 10/07/2019.

FAO. 2009. Declaración de la Cumbre Mundial sobre la Seguridad Alimentaria, Roma, noviembre de 2009. Disponible en línea: http://www.fao.org/fileadmin/templates/wsfs/Summit/Docs/final_Declaration/K60505_WSFS/OEWG_06.pdf. Visita: 04/09/2019.

FAO. 2017. Citrus fruits statistics 2017. Disponible en línea: http://www.fao.org/economic/est/est-commodities/citricos/es/. Visita: 04/09/2019.

Federcitrus. 2018. La actividad citrícola Argentina. Disponible en línea: https://www.federcitrus.org/. Visita: 04/09/2019.

Felmer, R; Chavez, R; Catrileo, A; Rojas, C. 2006. Tecnologías actuales y emergentes para la identificación animal y su aplicación en trazabilidad animal. Archivos de Medicina Veterinaria. INIA. 38, 3.

Forina, M; Casale, M; Olivieri, P. 2009. Application of chemometrics to food chemistry. In Brown, SD; Tauler, R; Walczak, B (Eds) Comprehensive Chemometrics. Amsterdam, Elsevier. Pp 75-128.

Geana, Y; Iordache, A; Ionete, R; Marinescu, A; Ranca, A; Culea, M. 2013. Geographical origin identification of Romanian wines by ICP-MS elemental analysis. Food Chemistry. 138: 1125-1134.

Gonzálvez, A; Armenta, S; de la Guardia, M. 2009. Trace-element composition and stable-isotope ratio for discrimination of foods with Protected Designation of Origin. TrAC Trends in Analytical Chemistry. 28 (11): 1295-1311.

Gonzálvez, A; de la Guardia, M. 2013. Basic Chemometric Tools. In: de la Guardia, M; Gonzálvez, A. Food protected designation of origin: Methodologies and applications. Comprehensive Analytical Chemistry. 60 (12): 299-315.

Kabata Pendias, A. 2010. Trace Elements in Soils and Plants, Fourth Edition: Taylor & Francis.

Londoño Posso, DM. 2013. Validación del método de determinación de Calcio y Magnesio por espectroscopia de absorción atómica de llama para el laboratorio de aguas y alimentos de la Universidad Tecnológica de Pereira. Repositorio Institucional. 126 pp.

Luykx, D; Van Ruth, S. 2008. An overview of analytical methods for determining the geographical origin of food products. Food Chemistry. 107: 897-911.

Magdas, D; Dehelean, A; Feher, Y; Cristea, G; Puscas, R; Dan, S; Cordea, D. 2016. Discrimination markers for the geographical and species origin of raw milk within Romania. International Dairy Journal. 61: 135-141.

Maione, C; Batista, B; Campiglia, A; Barbosa, F; Barbosa, R. 2016. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. Computers and Electronics in Agriculture. 121: 101-107.

Matusewicz, H; Bulska, E. 2018. Inorganic trace analytics: trace elements analysis and speciation. De Gruyter. Berlín. 459 pp.

Melnyk, S; Lummus, R; Vokurka, R; Quemaduras, L; Sandor, J. 2009. Mapping the future of supply chain management: a Delphi study. International Journal of Production Research. 47 (16): 4629-4653.

Molina, NA. 2006. Situación actual de la citricultura en Corrientes. Tierra Correntina: año 1 Nro. 1. Archivo Documental de INTA.

Molina, CA. 2013. Validación de las técnicas para determinación de Molibdeno y Cromo en agua residual, tratada, cruda y de Níquel en agua residual mediante espectrometría de absorción atómica por llama directa para el laboratorio de análisis de aguas y alimentos UTP. Universidad Tecnológica de Pereira. Repositorio Institucional. 149 pp.

Opara, LU. 2003. Traceability in agriculture and food supply chain: A review of basic concepts, technological implications, future prospects. International System for Agricultural Science and Technology AGRIS. 1 (1): 101-106.

Palacios, J. 2013. Citricultura. Talleres Gráficos Alfa Beta S.A. ISBN: 9789874383266. 518 pp.

Parlamento Europeo. 2002. Reglamento N° 178/2002 del Parlamento Europeo y del Consejo. Disponible en línea: https://www.mapa.gob.es/es/pesca/temas/calidad-seguridad-alimentaria/R178-2002_tcm30-285773.pdf. Visita: 24/07/2019.

Rincón Ballesteros, DL. 2016. Conceptualización de la trazabilidad en la cadena de abastecimiento. Universidad Distrital Francisco José de Caldas. Repositorio Institucional. 83 pp.

Roussos, PA. 2011. Phytochemicals and antioxidant capacity of orange (Citrus sinensis (l.) Osbeck cv. Salustiana) juice produced under organic and integrated farming system in Greece. Scientia Horticulturae. 129: 253-258.

SENASA. 2014. Escenarios y Tendencias. Informe Estadístico Cítricos Argentinos de Excelencia. SENASA. Ministerio de Agricultura, Ganadería y Pesca, Presidencia de la Nación.

Thomas, R. 2008. Guía práctica para ICP-MS: A Beginner's Tutorial, Second Edition: CRC Press.

Villafañe, R. 2018. Estudio de la composición mineral de forrajeras nativas de la provincia de Corrientes. Propuesta de modelos quimiométricos para evaluar propiedades químicas y eventual origen geográfico. Tesis para optar al título de Doctor en Recursos Naturales, Facultad de Ciencias Agrarias, Universidad Nacional del Nordeste. Abril de 2018.

Watson, C; Edwards, A; Dahlin, A; Eriksson, J; Lindstrom, B; Linse, L; Owens, K; Topp, C; Walker, R. 2012. Using soil and plant properties and farm management

practices to improve the micronutrient composition of food and feed Journal of Geochemical Exploration. 121: 15-24.

Weil, B; Brady, R. 2017. The nature and properties of soils. Pearson Education. 15th edition.

Yi, LT; Li, JM; Li, YC; Panorámica, Y; Xu, Q; Kong, LD. 2008. Antidepressant-like behavioral and neurochemical effects of the citrus-associated chemical apigenin. Life Sciences. 82: 741-751.

CAPÍTULO II

ESTADÍSTICA MULTIVARIANTE Y APRENDIZAJE AUTOMÁTICO

2.1. Introducción

Cuando se cuenta con gran cantidad de información sobre una muestra es muy difícil visualizar patrones de comportamiento y la mayoría de esta información puede no ser útil, por lo que se deben aplicar herramientas de análisis a los datos para extraer conocimiento que será provechoso para resolver el problema a investigar (Brereton, 2009; Otto, 2007; Varmuza & Filzmoser, 2016).

El uso de herramientas matemáticas, estadísticas y de aprendizaje automático en el proceso de medida químico ha generado un nuevo campo denominado Quimiometría, cuyo progreso en el siglo XXI se ha debido principalmente a la mejora de los programas estadísticos utilizados para ello (Kumar *et al.*, 2014). Estas técnicas se utilizan para la búsqueda de clasificaciones y de información oculta en las matrices complejas de datos generadas por las herramientas de análisis químico, simplificando enormemente la tarea de evaluación de las características de nuevas muestras. Dentro de la Quimiometría el reconocimiento de patrones ha ocupado un lugar importante (Brereton, 2009).

A su vez, podemos definir el Aprendizaje Automático como una rama de la Inteligencia Artificial cuyo principal objetivo es desarrollar modelos de predicción o clasificación a partir de evidencias previas, frecuentemente llamados ejemplos de entrenamiento. Identifica patrones subyacentes en un conjunto de datos que permite realizar agrupaciones de los mismos en distintas clases de modo que es posible asignar nuevas evidencias o ejemplos a uno de esos grupos de forma automática (Russel & Norving, 2004). El Aprendizaje Automático se encuentra en la intersección entre la

Programación, la Ingeniería y la Estadística. Puede ser aplicado a diversos problemas en variados y diversos campos, desde la Política a las Ciencias de la Tierra, todo campo que pueda obtener datos e interpretarlos puede beneficiarse de las técnicas de Aprendizaje Automático.

2.2. Análisis Multivariante de datos

Los problemas que la ciencia debe abordar involucran de manera simultánea a varias variables, por lo que para describir cualquier situación empírica se requiere tenerlas en cuenta y su análisis por separado no permite visualizar su comportamiento conjunto en la realidad. Se debe trabajar, entonces, con datos que provienen de la observación de múltiples variables sobre un conjunto de individuos o elementos de una población o muestra, y en ese caso las técnicas de Análisis Multivariante son especialmente adecuadas para descubrir patrones generales de comportamiento de los datos. El análisis de datos multivariante comprende el estudio estadístico de varias variables medidas simultáneamente en un conjunto de elementos, con la finalidad de resumir los datos mediante un pequeño conjunto de nuevas variables, encontrar grupos en los datos (si existen), clasificar nuevas observaciones en grupos definidos o relacionar dos o más conjuntos de variables (Cuadras, 2010; Hair *et al*., 2005; Johnsson, 2000; Peña, 2002).

El Análisis Multivariante es la parte de la Estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resulten de observar un número $p > 1$ de variables estadísticas sobre una muestra de n individuos. La información estadística en análisis multivariante es de carácter multidimensional, por lo tanto, la geometría, el cálculo matricial y las distribuciones multivariantes juegan un papel fundamental. La información multivariante es una matriz de datos, pero a menudo, en análisis multivariante la información de entrada puede consistir en matrices de distancias o similaridades, que miden el grado de discrepancia entre los individuos (Cuadras, 2010).

Si disponemos de *n* individuos (ω_1, ω_2, …, ω_n) y *p* variables (X_1, X_2, …, X_p), sea x_{ij} la observación de la variable X_j en el individuo ω_i. La matriz de datos multivariantes *X*, de orden *n* x *p*, con un término general (x_{ij}), consiste en un arreglo rectangular de escalares de *n* filas por *p* columnas. Tenemos entonces una matriz *X,* en la que las filas se identifican con los individuos y las columnas con las variables.

$$X = \begin{bmatrix} x_{11} \ldots x_{1j} \ldots x_{1p} \\ \ldots\ldots\ldots\ldots\ldots \\ x_{i1} \ldots x_{ij} \ldots x_{ip} \\ \ldots\ldots\ldots\ldots\ldots \\ x_{n1} \ldots x_{nj} \ldots x_{np} \end{bmatrix}$$

El problema principal del estudio simultáneo de *p* variables es que los datos se encuentran en un espacio de *p* dimensiones que es inabordable visualmente e inimaginable desde la comprensión humana. La descripción de una realidad compleja se puede simplificar al utilizar unas pocas variables que concentran la información del conjunto de las *p* variables, lo que permite su representación gráfica, la visualización del conjunto total y la comparación de diferentes conjuntos de datos o momentos, mejorando el conocimiento de la realidad estudiada. El Análisis Multivariante de datos proporciona métodos objetivos para definir esas nuevas variables que permiten describir adecuadamente la realidad compleja. Por otra parte, identificar algunas variables que presenten valores comunes en los elementos de grupos o subpoblaciones, permite establecer mecanismos de clasificación de los elementos en dichos grupos (Hair *et al*., 2005; Johnsson, 2000; Peña, 2002).

Las técnicas de Análisis Multivariante tienen aplicaciones en todos los campos científicos. Comenzaron desarrollándose para resolver problemas de clasificación en Biología, se extendieron para encontrar variables indicadoras y factores en Psicometría, Marketing y las Ciencias Sociales y han alcanzado una gran aplicación en Ingeniería y Ciencias de la Computación, como herramientas para resumir información y diseñar sistemas de clasificación automática y de reconocimiento de patrones. En la

actualidad existe una gran cantidad y variedad de métodos, con orígenes teóricos distintos, lo que puede producir una sensación de confusión al que se introduce por primera vez a estas técnicas (Hair *et al*., 2005; Peña, 2002).

Las técnicas de Análisis Multivariante pueden clasificarse de diferentes maneras, según se trate de métodos descriptivos (o exploratorios) o métodos explicativos o confirmatorios; según se aplique a una o varias poblaciones, y según intervengan uno o más grupos de variables; según objetivos perseguidos: estudio de variables o de individuos. El primer caso, puede plantearse a dos niveles, en el primero se desea extraer información que contienen los datos disponibles, en él se encuentran los métodos conocidos como métodos de exploración de datos. Estas técnicas de exploración de datos multivariantes se han popularizado en los últimos años en Ingeniería y Ciencias de la Computación con el nombre de "minería de datos", lo que indica su capacidad para extraer información a partir de la materia prima, los datos. Estas herramientas no permiten directamente obtener conclusiones generales acerca del proceso o sistema que genera los datos, por lo que un segundo nivel, cuando se pretende obtener conclusiones sobre la población que ha generado los datos, requiere la construcción de un modelo que explique su generación y permita prever datos futuros. En este segundo nivel se puede generar conocimiento sobre un problema que va más allá del análisis particular de los datos disponibles, comprende los métodos de inferencia multivariante (Hair *et al*., 2005; Peña, 2002).

El primer paso en la descripción de datos multivariantes es describir cada variable y comprender la estructura de dependencia que existe entre ellas. Siempre que sea posible conviene utilizar técnicas gráficas para resumir y representar la información contenida en los datos, y analizar la forma de medir las variables para obtener una representación lo más simple posible. En un primer análisis conjunto de las variables se aplican las denominadas técnicas de análisis exploratorio de datos, que incluyen herramientas gráficas y analíticas y se utilizan para una primera descripción del comportamiento multivariante de los individuos bajo estudio.

2.2.1. Medidas descriptivas multivariantes

De la misma manera que calculamos estadísticas univariantes de posición (media aritmética, mediana, moda, entre otras) y de dispersión (variancia, desviación estándar, entre otras), podemos calcular estadísticas descriptivas para el caso multivariante (Peña, 2002).

La medida de centralización más utilizada para describir datos multivariantes es el vector de medias, que es un vector de dimensión *p* cuyos componentes son las medias de cada una de las *p* variables.

$\overline{\mathbf{X}} = (\overline{x}_1,..., \overline{x}_j,...\overline{x}_p)$ es el vector (fila) de las medias de las variables:

$$x_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Las medidas de centralización escalares basadas en el orden de las observaciones no pueden generalizarse fácilmente al caso multivariante. Por ejemplo, podemos calcular el vector de medianas, pero este punto no tiene necesariamente una situación como centro de los datos. Esta dificultad proviene de la falta de un orden natural de los datos multivariantes.

Para variables escalares la variabilidad respecto a la media se mide habitualmente por la varianza, o su raíz cuadrada, la desviación típica, mientras que la relación lineal entre dos variables se mide por la covarianza. Esta información para un problema multivariante puede presentarse de forma compacta en la matriz de varianzas y covarianzas, que es una matriz cuadrada y simétrica que contiene en la diagonal las varianzas y fuera de la diagonal las covarianzas entre las variables.

S = ($s_{jj'}$) es la matriz *pxp* de varianzas y covarianzas muestrales:

$$s_{jj'} = \frac{1}{n}\sum_{1}^{n}(x_{ij} - \overline{x}_j)(x_{j'} - \overline{x}_{j'})$$

Naturalmente $\bar{X}$ y ***S*** son medidas multivariantes de tendencia central y dispersión.

Una forma de resumir la variabilidad de un conjunto de variables es mediante la traza de su matriz de varianzas y covarianzas y se define la variabilidad total de los datos por:

$$T = tr(S) = \Sigma \lambda_i$$

También es posible calcular la Varianza Generalizada como el determinante de *S* ($|S| = \lambda_1.\lambda_2 \ldots \lambda_p$), cuya raíz cuadrada (la Desviación Típica Global) es una medida del volumen p-dimensional que ocupan los datos.

Un objetivo fundamental de la descripción de los datos multivariantes es comprender la estructura de dependencias entre las variables. Estas dependencias pueden estudiarse: entre pares de variables (matriz de correlación); entre una variable y todas las demás (regresión múltiple); entre pares de variables, pero eliminando el efecto de las demás variables (correlaciones parciales); entre el conjunto de todas las variables (coeficiente de dependencia).

2.2.2. Análisis gráfico

Obtener buenas representaciones gráficas de datos multivariantes es un problema difícil. Recordemos que las correlaciones miden las relaciones lineales entre las variables, y pueden ser mal interpretadas cuando las relaciones son no lineales. Por esa razón se intenta transformar las variables para que las variables transformadas tengan relaciones aproximadamente lineales. Por último, los datos multivariantes contienen con frecuencia observaciones que son heterogéneas con el resto que, si no son detectadas, pueden alterar completamente el análisis descriptivo de las variables originales. Existen métodos para detectar los datos atípicos.

El primer paso de cualquier análisis multivariante es representar gráficamente las variables individualmente, mediante un histograma o un diagrama de caja. Estas representaciones son muy útiles para detectar asimetrías, heterogeneidad, datos atípicos etc. En segundo lugar, conviene construir los diagramas de dispersión de las variables por pares, con p variables existen p (p-1) /2 gráficos posibles que pueden disponerse en forma de matriz y son muy útiles para entender el tipo de relación existente entre pares de variables, e identificar puntos atípicos en la relación bivariante. Para más de tres variables se utilizan principalmente dos tipos de métodos gráficos. El primero, es mostrar los datos mediante figuras planas, asociando cada variable a una característica del gráfico (Figuras de Chernoff, Gráficos de Estrella). El segundo, es buscar conjuntos de proyecciones en una y dos dimensiones que revelen aspectos característicos de los datos. Una forma simple de resumir un vector de variables es construir una variable escalar como combinación lineal de sus valores (Peña, 2002).

2.2.3. Análisis de la Variancia Multivariante

El análisis multivariante de la varianza (MANOVA) es una generalización en $p > 1$ variables del análisis de la varianza (ANOVA). Es una técnica que puede ser usada para explorar simultáneamente la relación entre varias variables categóricas independientes (usualmente denominadas tratamientos) y dos o más variables métricas dependientes (variables respuesta). Permite probar la hipótesis de nulidad de efecto de los tratamientos. Además de identificar si los cambios en las variables independientes tienen efectos significativos en las variables dependientes, la técnica también intenta identificar las interacciones entre las variables independientes y su grado de asociación con las dependientes (Hair *et al*., 2005).

Las distribuciones estadísticas más comunes son la lambda (Λ) de Wilks, la traza de Pillai-M, la traza de Lawley-Hotelling y la raíz mayor de Roy. La discusión continúa sobre los méritos de cada una, aunque la raíz más grande que conduce sólo a una cota

de significancia no es de interés práctico. Una complicación más es que la distribución de estas estadísticas bajo la hipótesis nula no es sencilla y sólo puede ser aproximada, excepto en unos casos de pocas dimensiones. En el caso de dos grupos, todas las estadísticas son equivalentes y las pruebas se reducen a la distribución T^2 de Hotelling (Cuadras, 2010).

2.2.4. Análisis de Componentes Principales

Es un método de Ordenación que permite la representación geométrica de los individuos en dimensión reducida de modo que se expresen sus diferencias y analogías de la mejor forma posible.

Un problema central en el análisis de datos multivariantes es la reducción de las dimensiones: si es posible describir con precisión los valores de p variables por un pequeño subconjunto $r < p$ de ellas, se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información. Se desea encontrar un subespacio de dimensión menor que p tal que al proyectar sobre él los puntos conserven su estructura con la menor distorsión posible.

El Análisis de Componentes Principales (*Principal Component Analysis*, PCA), dadas n observaciones de p variables tiene como objetivo analizar si es posible representar adecuadamente esta información con un número de variables menor que p construidas como combinaciones lineales de las originales. Este análisis permite, por un lado, representar óptimamente en un espacio de dimensión reducida observaciones de un espacio general p-dimensional, y por otro lado transformar las variables originales, en general correlacionadas, en nuevas variables no correlacionadas, facilitando la interpretación de los datos. Representar puntos p dimensionales con la mínima pérdida de información en un espacio de menor dimensión es equivalente a sustituir las p variables originales por un conjunto de nuevas variables, que resuman óptimamente la información. Esto supone que las

nuevas variables deben tener globalmente máxima correlación con las originales o, en otros términos, deben permitir predecir las variables originales con la máxima precisión, para lo cual esas nuevas variables deben retener la máxima variabilidad (Hair *et al*., 2005; Peña, 2002).

El *biplot* es un gráfico en el que se representan las observaciones en las posiciones dadas por las dos primeras componentes principales definidas por el PCA. Sobre el mismo plano se representan los *n* individuos mediante puntos y simultáneamente, mediante vectores, las variables (las columnas de la matriz de datos *X*), en posiciones que hacen interpretables las relaciones entre ellas y las observaciones (Di Rienzo *et al*., 2020).

Las componentes principales son nuevas variables con las propiedades:

1. Conservan la variabilidad inicial (la suma de las varianzas de las componentes es igual a la de las X, ídem para la varianza generalizada).

$$tr(S) = \sum_{i=1}^{p} Var(x_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} Var(Z_i)$$

Siendo *x* las variables originales, λ los valores propios de la matriz de varianzas covarianzas y *Z* las nuevas variables o componentes principales.

2. La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado a su vector propio y la suma de valores propios de la matriz.

Var (z_h) = λ_h la porción explicada es $\lambda_h / \Sigma\lambda_i$

3. La correlación entre un componente principal y la variable *X* es proporcional al coeficiente de esa variable en la definición de la componente.

4. Las *r* componentes principales ($r<p$) proporcionan la predicción lineal óptima con *r* variables del conjunto de variables *X*.

Para seleccionar el número de componentes, existen diversos métodos:

- ✓ Realizar un gráfico de λ_i frente a *i*. Buscar el punto en que la gráfica se hace horizontal.
- ✓ Seleccionar los componentes para cubrir una proporción determinada de varianza.
- ✓ Desechar aquellas componentes asociadas a valores propios inferiores a un valor determinado.

Como resultado del PCA se puede visualizar la matriz de covarianza o correlación sobre la que se realiza el análisis, la correlación de cada componente principal con las variables originales, el coeficiente de correlación cofenética, gráficos Biplot y árbol de recorrido mínimo, los autovectores y autovalores. Los autovectores proveen los coeficientes de las coordenadas de las diferentes componentes principales para cada variable involucrada en el análisis. Se puede seleccionar la cantidad de componentes principales a emplear e interesa la proporción de variabilidad total que conserva cada una de las componentes seleccionadas, la que viene expresada por los correspondientes autovalores.

2.2.5. Análisis Discriminante

El Análisis Discriminante (*Discriminant Analysis*, DA) es un método de Clasificación que se orienta a la búsqueda de grupos similares, lo más homogéneos posibles, en los cuales se clasificarán los individuos a analizar.

Se utiliza cuando se dispone de un conjunto de elementos o individuos que pueden venir de dos o más poblaciones distintas. En cada elemento se ha observado un conjunto *p* de variables aleatorias, cuya distribución en un espacio *p*-dimensional ($\Re^p$) se conoce en las poblaciones consideradas y se desea clasificar un nuevo elemento, con valores de las variables conocidas, en una de las poblaciones. Esta técnica se encuentra dentro de las denominadas de clasificación supervisada, para indicar que se debe disponer primero de una muestra de elementos bien clasificados

que sirve de pauta o modelo para la clasificación de las siguientes observaciones (Cuadras, 2010; Peña, 2002).

Es frecuente encontrar en la bibliografía que el DA es presentado como una técnica de reducción de dimensión al igual que el PCA, aunque la mayor diferencia con ésta consiste en que el DA busca maximizar la variancia inter-grupos y minimizar la variancia intra-grupos, mientras que el PCA busca ejes ortogonales de máxima variancia en el conjunto de datos. El objetivo en el análisis discriminante es encontrar el subespacio que optimiza la separación de esas clases (Raschka, 2015).

Una regla discriminante es un criterio que permite asignar un individuo a una población, y que a menudo es planteado mediante una función discriminante y una regla de clasificación que asigna el individuo a una u otra población según el valor que toma la función discriminante. El problema más sencillo consiste en clasificar a los elementos en una de dos poblaciones, esta función (correspondiente a la ecuación de un hiperplano) divide $\Re^p$ en dos regiones. En el caso de que se trate de $k \geq 3$ poblaciones, el espacio $\Re^p$, se tendrá que dividir en k regiones y la función discriminante deberá definir la cantidad de hiperplanos necesarios para ello. Como es posible equivocarse al asignar un individuo a una población a la que no pertenece, el método permite calcular la probabilidad de clasificación errónea (Cuadras, 2010).

Existen varios enfoques posibles para abordar el problema de clasificación mediante DA. El primero, es el análisis discriminante clásico debido a Fisher, basado en la normalidad multivariante de las variables consideradas y que es óptimo bajo dicho supuesto. Si todas las variables son continuas, es frecuente que, aunque los datos originales no sean normales es posible transformar las variables para que lo sean y si se tienen variables discretas y continuas para clasificar, la hipótesis de normalidad multivariante es poco realista, pero existen otros enfoques al problema que pueden funcionar mejor en estos casos. Estos métodos funcionan bien con variables cuantitativas o cuando se conoce la densidad. Pero a menudo las variables son binarias, categóricas o mixtas. Aplicando el principio de que siempre es posible definir una distancia entre observaciones, es posible dar una versión del análisis discriminante utilizando solamente distancias (Cuadras, 2010; Peña, 2002).

Al hacer el DA, una información útil para decidir la bondad del modelo es provista por las tasas de error aparente (estimadores de la probabilidad de una mala clasificación), obtenidas al clasificar las observaciones en los grupos en cuestión a partir del uso de la función discriminante construida. Las tasas de error aparente tienden a subestimar el error, son útiles cuando se disponen de grandes tamaño de muestra en cada población (Raschka, 2015).

El Análisis Discriminante por Mínimos Cuadrados Parciales (PLS-DA) es un método de clasificación lineal basado en el algoritmo de regresión de mínimos cuadrados parciales (PLS) para construir modelos predictivos cuando los factores (variables independientes) son muchos y altamente colineales. El algoritmo PLS busca variables latentes con covarianza máxima con las variables dependientes que representan la pertenencia a la clase. Como tal, PLS tiene en cuenta la variable dependiente al definir variables latentes. Esta técnica se convierte en una herramienta establecida en el modelado quimiométrico, ya que a menudo es posible interpretar el factor extraído en términos del sistema físico subyacente. En general, a menudo se informa que el método PLS-DA funciona bien en la práctica (Ballabio y Consonni, 2013).

2.3. Aprendizaje Automático

El campo de la Estadística se enfrenta constantemente a los problemas que traen la ciencia y la industria. En los primeros tiempos, estos problemas a menudo consistían en experimentos agrícolas e industriales y tuvieron un alcance relativamente pequeño. Con la llegada de las computadoras y la era de la información, los problemas relacionados con el análisis de datos han cambiado tanto en tamaño como en complejidad. Desafíos en las áreas de almacenamiento de datos, organización y búsqueda han llevado a un nuevo campo, la "minería de datos". Grandes cantidades de datos se están generado en muchos campos y extraer patrones y tendencias,

entender "lo que el dato dice", lo que se llama "aprender de los datos" es una tarea fundamental (Hastie *et al*., 2017).

El Aprendizaje Automático es un campo multidisciplinar relacionado con la estadística, la teoría de la información, la teoría de los juegos y la optimización. Puede verse como una rama de la Informática y de la Inteligencia Artificial, pero en contraste con la Inteligencia Artificial tradicional, el Aprendizaje Automático no trata de construir una imitación automatizada del comportamiento de la inteligencia humana, sino más bien de utilizar los puntos fuertes y las capacidades especiales de las computadoras para complementar la inteligencia humana, a menudo realizando tareas que van mucho más allá de las capacidades humanas. Por ejemplo, la capacidad de escanear y procesar enormes bases de datos permite a los métodos de aprendizaje automático detectar patrones que la percepción humana no sería capaz de reconocer.

El Aprendizaje Automático obtiene conclusiones que se ajustan al entorno de donde se extraen los ejemplos con los que se construye el clasificador. Estos datos que nos sirven para realizar el aprendizaje generalmente se obtienen aleatoriamente. Esta descripción del Aprendizaje Automático pone de relieve su relación con la Estadística. De hecho, hay mucho en común entre las dos disciplinas, tanto en lo que respecta a los objetivos como a las técnicas utilizadas. Sin embargo, hay algunas diferencias significativas.

A diferencia de la Estadística, en el Aprendizaje Automático las consideraciones algorítmicas juegan un papel importante. El Aprendizaje Automático se refiere a la ejecución del aprendizaje por computadoras. Desarrollamos algoritmos para realizar las tareas de aprendizaje y nos preocupamos por su eficiencia computacional. Otra diferencia es que mientras que la Estadística está a menudo interesada en el comportamiento asintótico, la teoría del Aprendizaje Automático se centra en los límites de las muestras finitas.

El Aprendizaje Automático pues es un área recientemente desarrollada, que combina la Estadística con desarrollos paralelos en Ciencias de la Computación y la Información. El campo del Aprendizaje Automático involucra la cuestión de cómo

construir programas informáticos que mejoran automáticamente con la experiencia (Hastie *et al*., 2017; James *et al*., 2017). En los últimos años se han desarrollado muchas aplicaciones exitosas de Aprendizaje Automático, que incluyen programas de minería de datos y sistemas de filtrado de información, entre otros. El Aprendizaje Automático se basa en conceptos y resultados de muchos campos, incluyendo Estadística, Inteligencia Artificial, Filosofía, Teoría de la Información, Biología, Ciencia Cognitiva, Complejidad Computacional y Teoría del Control (Mitchel, 1997).

Comprende un vasto conjunto de herramientas para comprender los datos, que pueden clasificarse como supervisadas (aprendizaje predictivo/inductivo) o no supervisadas (aprendizaje descriptivo/deductivo). En términos generales, el aprendizaje supervisado implica la construcción de un modelo para predecir o estimar, una salida basada en una o más entradas. Esto es, en el aprendizaje supervisado, el objetivo es predecir el valor de una medida de resultado basada en una serie de medidas de insumos. Con el aprendizaje no supervisado, hay insumos, pero ninguna salida supervisada, sin embargo, se pueden comprender relaciones y estructura de tales datos; en el aprendizaje no supervisado el objetivo es describir las asociaciones y patrones entre un conjunto de medidas de insumos (Hastie *et al*., 2017; James *et al*., 2017).

Los métodos no supervisados se usan para describir y reconocer las diferencias que se encuentran presentes en la matriz de datos analizada, sin tener en cuenta ninguna información relacionada a la distribución y/o clasificación previa de las muestras. Este tipo de métodos resultan muy útiles y son aplicados en general para realizar una primera descripción de los datos a analizar (análisis exploratorio de datos) y de esta manera predecir y/o diseñar el método de modelado a aplicar en una etapa subsiguiente.

Los métodos supervisados por su parte resultan más adecuados para establecer un modelo que permita determinar la procedencia geográfica, dado que tienen en cuenta al objetivo del análisis al ser calculados. Todos los métodos pertenecientes a esta clase comparten el hecho de que se realizan en dos etapas:

- ✓ *Etapa de calibración o entrenamiento*, durante la cual se desarrollan algoritmos de resolución de acuerdo con las restricciones matemáticas de cada método, en base a la muestra de objetos de estudio previamente clasificados.

- ✓ *Etapa de validación*, donde se pone a prueba el modelo con muestras problemas pertenecientes a clases o grupos conocidos y se calcula la tasa de error de clasificación.

2.3.1. Aprendizaje Supervisado

Dado un conjunto de entrenamiento con n ejemplos (x_1, y_1) (x_2, y_2), …, (x_n, y_n) tales que cada uno se obtiene a partir de una función f desconocida ($y_j = f(x_j)$, donde x_j es un vector de valores $(x_{j1}, \ldots, x_{jk})$ para los atributos de entrada $(A_{j1}, \ldots, A_{jk})$, y_j es el valor del objetivo. Se trata de encontrar una función h que aproxime de la mejor forma posible a f. En el dominio de aprendizaje supervisado, para cada observación de la(s) medida(s) del predictor (x_i, i a 1, ..., n) hay una medida de respuesta asociada y_i. Deseamos ajustar un modelo que relacione la respuesta a los predictores, con el objetivo de predecir con precisión la respuesta para futuras observaciones (predicción) o comprender mejor la relación entre la respuesta y los predictores (inferencia). Muchos métodos clásicos de aprendizaje, como la Regresión Lineal y la Regresión Logística, así como enfoques más modernos como las Máquinas de Vectores Soporte, operan en el dominio de aprendizaje supervisado. Se pueden distinguir dos situaciones de aprendizaje supervisado, de Clasificación, cuando se trate de variables categóricas y de Regresión, cuando las variables sean cuantitativas (James *et al*., 2017).

Figura 2.1. Métodos de Aprendizaje Supervisados, según el tipo de variables a los que se aplican y principales características

2.3.1.1. Árboles de Decisión

Los Árboles de Decisión (*Decision Trees*, DT) son métodos basados en árboles para la regresión y la clasificación que implican la estratificación o segmentación del espacio predictor en un número de regiones simples. Con el fin de hacer una predicción para una observación dada, normalmente se usa el entrenamiento de las observaciones para la región a la que pertenece. Dado que el conjunto de reglas de división utilizados para segmentar el espacio predictor se puede resumir en un árbol, estos tipos de enfoques se conocen como métodos de árbol de decisión. Los DT clasifican las instancias ordenándolas por el árbol desde la raíz hasta algún nodo hoja, lo que proporciona la clasificación de la instancia. Cada nodo del árbol especifica una prueba de algún atributo de la instancia y cada rama descendente (James *et al.*, 2017; Mitchel, 1997).

Los DT se encuentran entre los métodos de aprendizaje más utilizados y prácticos para la inferencia inductiva, permiten aproximar funciones para variables de cualquier tipo y capaces de aprender expresiones disyuntivas. Son útiles, de sencilla

interpretación, pueden ser aplicados para problemas de Regresión o de Clasificación. Además, tienen la habilidad de seleccionar las variables que son relevantes para la clasificación. Sin embargo, por lo general logran menor porcentaje de acierto en la predicción que otros enfoques de aprendizaje supervisado (James *et al.*, 2017; Mitchel, 1997).

En el proceso de construcción de un árbol de decisión, en términos generales, intervienen los siguientes pasos:

- Se usa la estrategia "divide y vencerás". Utiliza los valores de las variables predictoras para dividir el conjunto de entrenamiento en conjuntos cada vez más pequeños de la misma clase (si es posible).

- Como nodo raíz se elige la variable predictora que mejor predice la variable objetivo.

- Los ejemplos se dividen en grupos según los distintos valores de la variable seleccionada. Esta decisión produce las primeras ramas del árbol.

El procedimiento anterior se repite hasta que se cumple la condición de parada que se establezca.

La división recursiva divide los datos en grupos, los cuales a su vez se dividen en grupos aún más pequeños, hasta que el proceso se detiene cuando el algoritmo determina que los datos dentro de los grupos son suficientemente homogéneos, o se encuentra algún otro criterio de detención. Los DT clasifican las instancias ordenándolas por el árbol desde el nodo raíz (que representa el conjunto de datos completo y todas las observaciones pertenecen a una sola región dado que no ha habido ninguna división). Luego, el algoritmo debe decidir un atributo para empezar la división, y elige el atributo de mayor poder de predicción dentro de una clase (o grupo) determinado, hasta algún nodo hoja, lo que proporciona la clasificación de la instancia. Cada nodo del árbol especifica una prueba de algún atributo de la instancia y cada rama descendente de ese nodo corresponde a uno de los valores posibles para este atributo. Una instancia se clasifica comenzando en el nodo raíz del árbol, probando el

atributo especificado por este nodo y, a continuación, desviando hacia abajo la rama de árbol correspondiente al valor del atributo en el ejemplo dado. A continuación, este proceso se repite para el subárbol en el nuevo nodo, eligiendo el mejor parámetro cada vez para crear otro nodo de decisión, hasta que se encuentra un criterio de finalización (James *et al*., 2017; Mitchel, 1997).

En el caso de los árboles de clasificación los nodos finales del árbol, llamados hojas, se etiquetan con uno de los valores de la variable a predecir. Cuando se construyen árboles de regresión, en las hojas puede haber un valor fijo o una regresión que estime la variable objetivo.

Al interpretar los resultados de un árbol de clasificación, a menudo interesa no sólo la predicción de clase correspondiente a una región de nodo terminal determinada, sino también en las proporciones de clase entre las observaciones de entrenamiento que caen en esa región. Una ventaja del algoritmo recursivo binario con que se construyen los DT es su interpretabilidad, las características de partición del espacio son completamente descriptas en un solo árbol. Con más de dos entradas, las particiones son difíciles de dibujar, pero la representación de árbol binario funciona en el mismo sentido (Hastie *et al*., 2017; James *et al*., 2017).

Entre los árboles de decisión más usados se encuentran C4.5 y CART. C4.5 es quizá el árbol de decisión más extensamente utilizado (Quinlan, 2014). Sus características fundamentales son:

- Utiliza la razón de ganancia como medida de impureza de los nodos.
- Realiza poda (postpoda).
- La condición de parada depende de un umbral.
- Trabaja con datos faltantes y con variables numéricas.
- Construye árboles de cualquier aridad.

C5.0 es la evolución de C4.5, aunque sus ventajas son principalmente a nivel de implementación. En general es más rápido, consume menos memoria y es más preciso. Los árboles que genera suelen ser más pequeños. La principal novedad que incorpora es que incorpora la técnica de boosting al entrenamiento. En este sentido, C5.0 genera árboles de decisión que combina para realizar predicciones. Además, presenta otras funcionalidades, como el permitir un tratamiento diferenciado de los errores de clasificación (no todos los errores pesan lo mismo).

CART (*Classification and Regression Trees*) (Breiman *et al.*, 1984) crea árboles binarios, utiliza el índice de Gini como criterio de selección de nodos. Realiza un mecanismo de poda según crece el árbol, basado en el ratio coste/complejidad.

2.3.1.2. Bosques Aleatorios

Un árbol de decisión provee un modelo simple, pero a menudo resulta demasiado simple o específico por lo que se ha establecido que varios modelos trabajando juntos proporcionan mejores resultados que un solo modelo realizando todo el trabajo. La técnica de Bosques Aleatorios (*Random Forest*, RF) funciona mediante el ensamble de algoritmos de bajo aprendizaje, árboles de decisión en este caso, para mejorar el porcentaje de acierto de la técnica global (Hastie *et al.*, 2017).

Los RF tienden a producir modelos muy precisos porque el conjunto reduce la inestabilidad que se puede observar cuando se construyen arboles de decisión. Son métodos robustos al ruido por lo que pequeños cambios en el conjunto de datos de entrenamiento, tendrán un impacto nulo o ínfimo. Los bosques aleatorios son una técnica muy eficaz en casos de clasificación no lineal como las máquinas de vectores soporte o las redes neuronales.

La aleatoriedad introducida por el método se presenta tanto en la selección de observaciones como de variables y permite que el método sea robusto al ruido, a los datos extremos y al sobreajuste, cuando se lo compara con un solo árbol de decisión.

La aleatoriedad también tiene beneficios en cuanto al costo computacional. Las ventajas de los RF se pueden resumir en que a menudo requieren muy poco preprocesamiento de los datos, se evita la necesidad de selección de variables ya que el algoritmo lo realiza durante el análisis, cada árbol es un modelo independiente y el modelo resultante tiende a no sobre ajustar el conjunto de datos de entrenamiento (Williams, 2011).

2.3.1.3. Algoritmos de Clasificación por Vecindad

Los métodos de aprendizaje basados en instancias, como el vecino más cercano y la regresión localmente ponderada, son enfoques conceptualmente sencillos para aproximar funciones objetivo de variables categóricas o numéricas. Aprender en estos algoritmos consiste simplemente en almacenar los datos de entrenamiento presentados. Cuando se encuentra una nueva instancia de consulta, se recupera un conjunto de instancias relacionadas similares de la memoria y se utiliza para clasificar la nueva instancia de consulta. El algoritmo del Vecino más Cercano (*K-Nearest Neighbor*, KNN) es el método basado en instancias más básico. Este algoritmo supone que todas las instancias corresponden a puntos en el espacio *n*-dimensional, permite que una instancia arbitraria *x* sea descrita por el vector de características (Mitchel, 1997).

A efectos de hacer una predicción para una observación $X = x$, se identifican las observaciones de entrenamiento k más cercanas a x. A continuación, X se asigna a la clase a la que pertenece la mayoría de estas observaciones. Por lo tanto, KNN es un enfoque completamente no paramétrico: no se asumen sobre la forma del límite de decisión. El método KNN requiere la selección de k, el número de vecinos (James *et al.*, 2017).

En el aprendizaje del vecino más cercano, la función objetivo puede ser para variables categóricas o cuantitativas. En el caso del aprendizaje de funciones objetivos para variables categóricas, el valor $f(x,)$ devuelto por el algoritmo KNN como su

estimación de $f(x)$, es sólo el valor más común de f entre los ejemplos de entrenamiento k más cercanos a x. El algoritmo KNN se adapta fácilmente a la aproximación de funciones objetivo para variables cuantitativas, para lo cual el algoritmo calcula el valor medio de los ejemplos de entrenamiento k más cercanos en lugar de calcular su valor más común (Mitchel, 1997).

2.3.1.4. Redes Neuronales Artificiales

Las redes neuronales artificiales (*Artificial Neural Net*, ANN) proporcionan un método general y práctico para aprender funciones para variables cuantitativas y categóricas a partir de ejemplos. El aprendizaje ANN es sólido para los errores en los datos de entrenamiento y se ha aplicado con éxito a problemas como la interpretación de escenas visuales, el reconocimiento de voz y el aprendizaje de estrategias de control de robots (Mitchel, 1997).

Las ANN son modelos computacionales que inicialmente trataban de reproducir la actividad neuronal humana. El cerebro puede considerarse un sistema altamente complejo, donde se calcula que hay aproximadamente 100 mil millones (10111011) neuronas en la corteza cerebral que forman un entramado de más de 500 billones de conexiones neuronales (la media de conexiones de una neurona oscila entre 5000 y 10000).

El cerebro se caracteriza por presentar gran velocidad de proceso, tratar gran cantidad de información que captura a través de los sentidos y almacena. Además, es capaz de reaccionar y aprender ante situaciones nuevas. Por otra parte, almacena información redundante, no es estable, y sobre todo, tiene un poder desconocido.

Tareas que son muy sencillas para el ser humano, como por ejemplo reconocer visualmente a un familiar, o identificar un olor, son tareas computacionalmente muy complejas.

En analogía aproximada, las redes neuronales artificiales se construyen a partir de un conjunto densamente interconectado de unidades simples, donde cada unidad toma una serie de entradas de valor real (posiblemente las salidas de otras unidades) y produce una sola salida de valor real (que puede convertirse en la entrada a muchas otras unidades) (Mitchel, 1997).

Las ANN son un sistema de procesamiento de información que tiene propiedades inspiradas en las redes neuronales biológicas: el procesamiento de información ocurre en muchos elementos simples llamados neuronas. Las señales son transferidas entre neuronas a través de enlaces de conexión. Cada conexión tiene un peso asociado representando la sinapsis. Cada neurona aplica una función de activación a su entrada de red para determinar su salida, representando la generación de potenciales de acción. En la siguiente figura se presenta la estructura básica de una red neuronal.

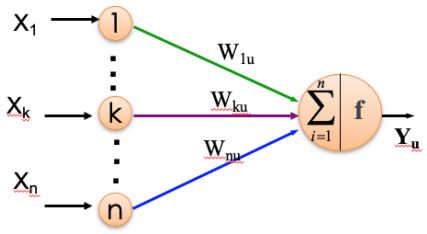

Figura 2.2. Estructura básica de una red neuronal

La ANN utilizan las neuronas como unidades mínimas de procesamiento conectadas entre sí, para producir modelos complejos. Existen muchos tipos de redes neuronales, en función de las siguientes características:

- Función de activación, que transforma las señales de entrada combinadas de una neurona en una sola señal de salida para ser transmitida a través de la red.
- La topología de la red que describe el número de neuronas en el modelo, el número de capas y la forma en que la que están conectadas.
- El algoritmo de entrenamiento con el que se aprenden los pesos de las conexiones, en función de los cuales se inhibe o excita una señal.

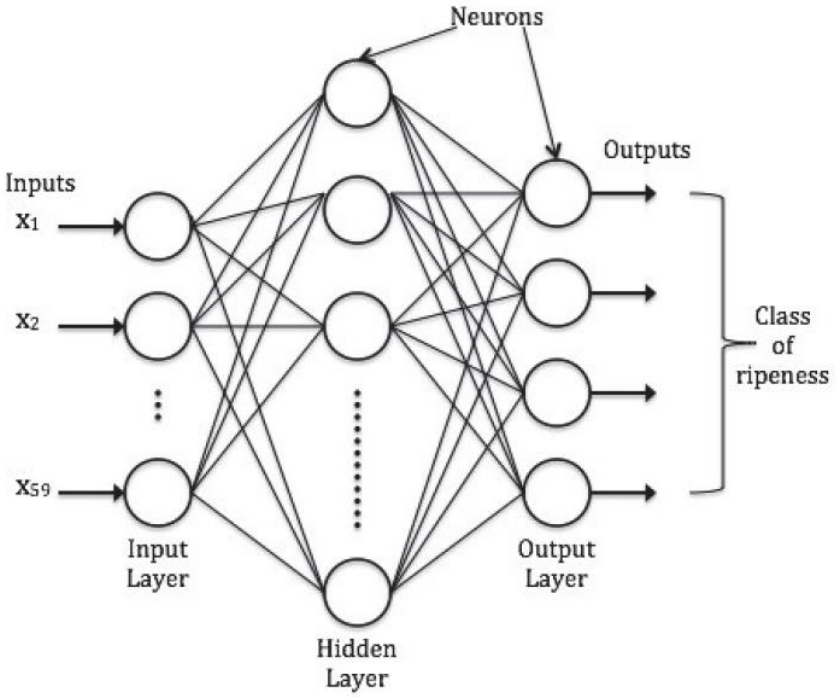

Figura 2.3. Estructura de una red neuronal multicapa (Norasyikin *et al*., 2012)

Cuando las nodos de entrada están conectados directamente con la capa de salida, hablamos de red de una capa. Las redes de una sola capa se pueden usar para la clasificación de patrones básicos, en particular para los patrones que son linealmente separables, pero se requieren redes más sofisticadas para la mayoría de las tareas de aprendizaje.

Una forma de crear redes más complejas consiste en añadir capas ocultas a la red, dando lugar a las redes multicapa.

Atendiendo a la dirección en la que se transmite la información en una red, podemos clasificarlas en:

- Redes *feedforward*: Los datos de entrada se mueven desde la capa de entrada hacia la capa de salida. Una red neuronal con múltiples capas ocultas se denomina red neuronal profunda y su entrenamiento aprendizaje profundo. Ejemplo de estas redes son los integrantes de la familia del Perceptrón.

- Red recurrente (o red de retroalimentación): los datos viajan en los dos sentidos, utilizando bucles. Esta propiedad, más similar al funcionamiento de una red neuronal biológica, permite resolver problemas extremadamente complejos. Si además se le dota a las neuronas de memoria a corto plazo el poder de estas redes aumenta considerablemente.

Además del número de capas y de flexibilidad de direcciones en las que viaja la información, las redes neuronales también pueden variar en complejidad por el número de neuronas en cada capa. Obviamente, el número de neuronas en la capa de entrada está predeterminado por el número de características en los datos de entrada. De manera similar, la cantidad de nodos de salida está predeterminada por el número de valores de la variable de salida. Sin embargo, decir la cantidad de neuronas ocultas que se deben añadir es un arte, ya que no existe una regla establecida para determinar el número de neuronas en la capas oculta. El número apropiado depende de la cantidad de nodos de entrada, la cantidad de datos de entrenamiento, la cantidad de datos ruidosos y la complejidad de la tarea de aprendizaje, entre muchos otros factores.

En general, las topologías de red más complejas con un mayor número de conexiones de red permiten el aprendizaje de problemas más complejos. Un mayor número de neuronas resultará en un modelo que refleje más de cerca los datos de entrenamiento, pero se puede producir sobreajuste, además de que pueden ser computacionalmente muy costosas. La mejor práctica es utilizar el menor número de nodos que den como resultado un rendimiento adecuado en un conjunto de datos de validación.

La topología de la red es precisamente lo que se debe fijar en un proceso de entrenamiento. A medida que la red neuronal procesa los datos de entrada, las conexiones entre las neuronas se fortalecen o se debilitan, de manera similar a cómo se desarrolla el cerebro de un humano a medida que experimenta con el medio. Los pesos de conexión de la red se ajustan para reflejar los patrones observados en el tiempo.

Ajustar los pesos de conexión de una red neuronal es habitualmente computacionalmente costoso. Por esa razón las redes neuronales, aunque desarrolladas desde hace muchos años, no se empezaron a utilizar hasta que se desarrolló la estrategia de retropropagación.

En síntesis, el algoritmo de retropropagación itera a través de muchos ciclos de dos procesos. Cada ciclo es conocido como una época. Debido a que la red no contiene conocimiento a priori, los pesos iniciales generalmente se establecen al azar. Luego, el algoritmo itera hasta que se alcanza un criterio de parada. Cada época en el algoritmo de propagación hacia atrás incluye:

- Una fase hacia adelante en la que las neuronas se activan en secuencia desde la capa de entrada a la capa de salida, aplicando los pesos y la función de activación de cada neurona en el camino. Al llegar a la capa final, se produce una señal de salida.

- Una fase hacia atrás en la que la señal de salida de la red se compara con el valor objetivo deseado en los datos de entrenamiento. La diferencia entre la señal de salida de la red y el valor objetivo da como resultado un error que se propaga hacia atrás en la red para modificar los pesos de conexión entre las neuronas y reducir los errores futuros. En esta fase, usa la derivada de la función de activación de cada neurona para identificar el gradiente en la dirección de cada uno de los pesos entrantes, de ahí la importancia de tener una función de activación diferenciable.

Es un proceso iterativo a partir de un conjunto de datos de entrenamiento, comparando la predicción de la red para cada ejemplo con el valor objetivo real conocido puede ser la etiqueta de clase conocida del ejemplo de entrenamiento si el problema es de clasificación, o un valor continuo para predicción numérica. Para cada ejemplo de entrenamiento, los pesos se modifican para minimizar el error cuadrático medio entre la predicción de la red y el valor objetivo real. Estas modificaciones se realizan en la dirección "hacia atrás" (es decir, desde la capa de salida) a través de cada capa oculta hasta la primera capa oculta (de ahí el nombre de propagación hacia

atrás). Aunque no está garantizado, en general las ponderaciones eventualmente convergerán y el proceso de aprendizaje se detendrá. Los pasos se definen a continuación.

- Inicializar los pesos: los pesos en la red se inicializan a números aleatorios pequeños. Cada unidad tiene un sesgo asociado, que se inicializan de manera similar. Cada ejemplo de entrenamiento, *X*, se procesa mediante los siguientes pasos.

- Propagación hacia delante: el ejemplo de entrenamiento se alimenta a la capa de entrada de la red. Las entradas pasan a través de las unidades de entrada, sin cambios. A continuación, se calculan la entrada y salida neta de cada unidad en las capas ocultas y de salida. Cada una de estas unidades tiene una serie de entradas que son, de hecho, las salidas de las unidades conectadas a ella en la capa anterior.

- Error de retropropagación: el error se propaga hacia atrás al actualizar los pesos y sesgos para reflejar el error de predicción de la red. Todo este proceso se debe realizar para cada ejemplo que se presenta a la red. También pueden acumularse las modificaciones y realizar la actualización de pesos después de presentar varios ejemplos a la red (épocas).

- Condición de parada.

2.3.1.5. Máquina de Vectores de Soporte

Técnicas como los Análisis de Regresión y Discriminante, que permiten generar hiperplanos de separación entre grupos de observaciones son óptimos en el caso que dos clases son linealmente separables, en el caso de que las clases se superponen y no son linealmente separables, estas técnicas se pueden generalizar a lo que se conoce como la Máquina de Vectores de Soporte (*Support Vector Machines*, SVM). Son un enfoque para la clasificación que se desarrolló en la comunidad de Ciencias de la

Computación en la década de 1990 y que ha crecido en popularidad desde entonces, han demostrado funcionar bien en una variedad de ajustes.

Una SVM crea hiperplanos que realizan una partición de los datos en grupos homogéneos con respecto a la clase de los datos. El aprendizaje SVM combina aspectos de los métodos basados en instancias (KNN) y de los modelos de Regresión Lineal, obteniendo un método que es muy potente, permitiendo modelar relaciones altamente complejas.

Las SVM se entienden más fácilmente cuando se utilizan para clasificación binaria, que es cómo se han concebido inicialmente, aunque pueden ser utilizados para resolver cualquier tipo de problema, tanto de clasificación binaria o multicategoría como de regresión.

Si el conjunto de datos a tratar es linealmente separable, como los datos que se muestran en la siguiente figura:

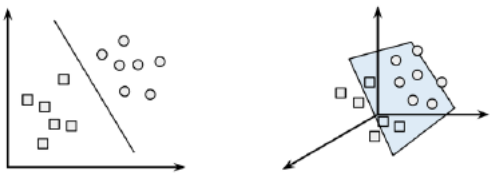

Figura 2.4. Ejemplo de frontera de separación para un problema bidimensional (izquierda) o tridimensional (derecha).

Las SVM utilizan este límite, que llamamos hiperplano, para dividir los datos en grupos de valores de clase similares. La figura anterior muestra cómo los ejemplos están perfectamente separados según su clase (círculos o cuadrados) por una línea recta (hiperplano en un espacio de dimensión 2) o una superficie (un hiperplano en un espacio de dimensión 3).

En el caso bidimensional, el algoritmo SVM intenta identificar ese hiperplano que separa las dos clases. La figura siguiente muestra varias posibilidades. SVM decide, de entre todas las posibilidades factibles, aquella que se corresponde con el Hiperplano de Margen Máximo, que es el que crea la mayor separación entre las dos clases.

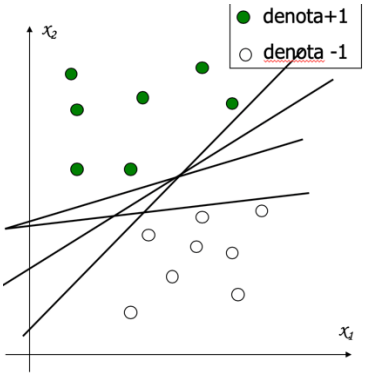

Figura 2.5. Diferentes separadores lineales para un conjunto de datos

Aunque cualquiera de las líneas que separan los círculos blancos de los verdes clasificaría correctamente todos los datos, es probable que la línea que conduce a la mayor separación generalice mejor los datos futuros. El margen máximo mejorará la posibilidad de que, a pesar de exista ruido, los ejemplos se clasifiquen mejor.

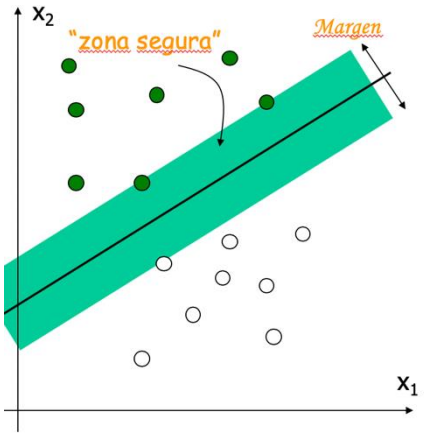

Figura 2.6. Margen máximo de separación para un conjunto de datos

Los vectores de soporte son los puntos de cada clase que son los más cercanos al margen máximo. Cada clase debe tener al menos un vector de soporte, pero es posible tener más de uno. De hecho, solo se utilizan los vectores soporte para determinar el margen máximo. Y esta es la característica fundamental de las SVM ya que estos vectores soporte generan el modelo de clasificación.

El esquema de funcionamiento anterior permite resolver cualquier tipo de problema introduciendo funciones *kernel* que transforman el espacio de entrada,

posiblemente no linealmente separable, en un espacio de características donde el conjunto sí es linealmente separable. La solución que proporciona SVM es un subconjunto del conjunto de entrenamiento que se denomina conjunto de vectores soporte.

Los vectores proporcionan una forma muy compacta de almacenar un modelo de clasificación, incluso si el número de características es extremadamente grande. (Hastie et al., 2017; James et al., 2017).

2.4. Entrenamiento de Modelos

En el proceso de entrenamiento se suele dividir los datos disponibles en tres conjuntos diferenciados:

- Entrenamiento/training: los datos a usar para ajustar el modelo.
- Prueba: los datos a usar para evaluar el modelo entrenado.
- Validación: datos para buscar la configuración óptima del modelo.

Es posible no disponer de conjunto de validación si no hay muchos datos disponibles. No obstante, la distribución de las muestras puede afectar al modelo obtenido. Supongamos un conjunto de 10 muestras, 4 de ellas de una clase B. Supongamos que se eligen justo estas 4 para el conjunto de prueba. Entonces, el modelo seguramente presente muy malas prestaciones.

Para solucionar este problema se repite el proceso de dividir en tres subconjuntos y entrenar el modelo. Con ello se dispone entonces de datos de prestaciones provenientes de varias pruebas de entrenamiento y de validación (y posiblemente el conjunto de prueba final) que permiten obtener una visión estadística del comportamiento del modelo. Normalmente, el conjunto de datos de validación (de existir) se reserva inicialmente y no forma parte de esta selección para la repetición del proceso de aprendizaje con diferentes subconjuntos.

Existen varias formas de realizar este análisis estadístico del resultado de aprender un modelo para un problema concreto. Las técnicas más conocidas son:

- Validación cruzada: muestreo sin reposición. En este modo, las muestras se dividen en conjuntos disjuntos de entrenamiento y prueba. Podemos distinguir entre:

  - *Leave-one-out*: sea un conjunto de N muestras. El modelo se entrena N veces, cada vez dejando una muestra diferente para el conjunto de prueba.

  - *Leave-one-group-out*: similar al anterior, pero en vez de considerar una sola muestra para prueba, se reserva un grupo de K muestras.

  - *k-fold cross validation*: se divide el conjunto de datos original en K grupos de muestras disjuntos. Se realizarán *K* entrenamientos/tests, donde en el proceso *K-esimo* se usa el subconjunto *K* para test y el resto de los datos para entrenamiento. Es típico usar *K* con valor 10.

  - 5x2 cv: el conjunto de datos se divide en dos partes, una para entrenamiento y la otra para test. Seguidamente, se intercambian los subconjuntos y se entrena y evalúa la prueba otra vez. El conjunto completo de dos entrenamientos se repite 5 veces de forma independiente, obteniendo 10 ejecuciones del proceso de entrenamiento y test.

- *Bootstrap*: muestreo con reposición. En este modo, cada conjunto de entrenamiento está formado por muestras del conjunto de datos, pero pudiéndose introducir la misma muestra varias veces (Mitchel, 1997).

2.5. Evaluación de Modelos

2.5.1. Matriz de Confusión

La matriz de confusión para un caso binario contiene cuatro valores característicos: verdaderos positivos (VP), falso positivos (FP), falso negativos (FN), y verdaderos negativos (VN). Los casos verdaderos positivos (VP) y los verdaderos negativos (VN) son los casos que han sido clasificados correctamente mientras que los casos de falso negativo (FN) y falso positivo (FP) son casos clasificados erróneamente. Estos cuatro parámetros se presentan en la Figura 2.7 (Takaya & Rehmsmeier, 2015).

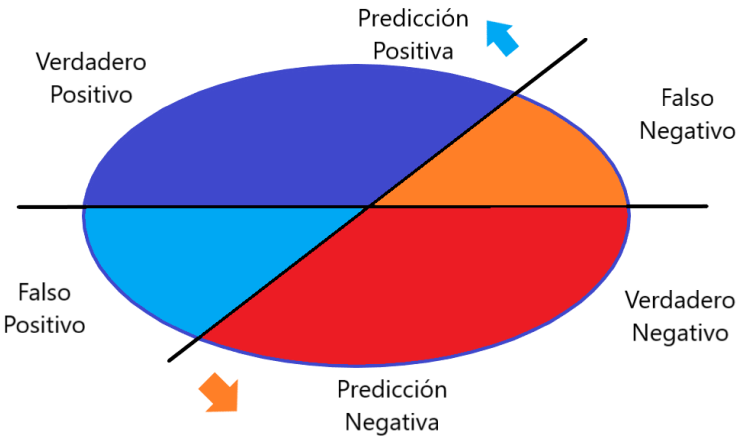

Figura 2.7. Valores de la matriz de confusión para un caso binario

La exactitud es la suma de los casos positivos y los casos negativos correctamente clasificados sobre el total de las muestras utilizadas en el proceso de clasificación. En la Figura 2.8 se ilustra el concepto de exactitud (Takaya & Rehmsmeier, 2015).

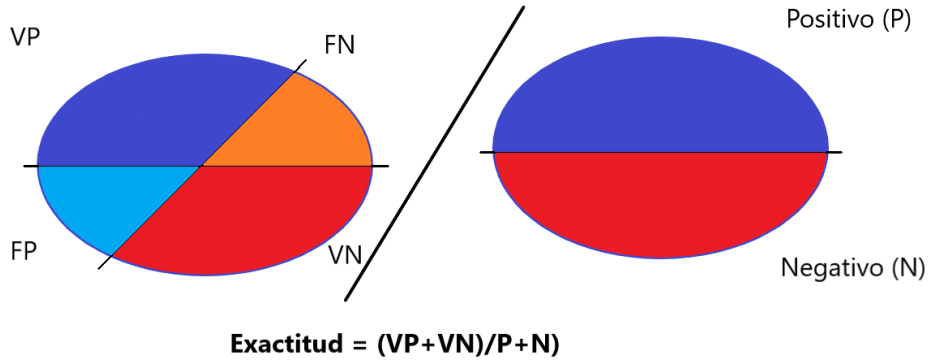

Figura 2.8. Exactitud de un clasificador

La tasa de verdaderos positivos es también llamada sensibilidad de un clasificador. Este término se origina en el campo médico, en el que la métrica es típicamente utilizada para estudiar la efectividad de una prueba clínica en detectar una enfermedad y es equivalente a investigar cuán sensible es una prueba para detectar la presencia de la enfermedad.

La métrica complementaria a esto es la especificidad de un algoritmo de aprendizaje y se enfocaría en la proporción de instancias negativas que son detectadas, por lo tanto, la especificidad es la tasa de casos negativos de un clasificador binario. Esto es la sensibilidad tiene en cuenta los casos positivos, mientras que cuando se la mide en los casos negativos se refiere a la especificidad (Japkowicz & Shah, 2011). En las Figuras 2.9 y 2.10 se ilustran los conceptos de sensibilidad y especificidad (Takaya & Rehmsmeier, 2015).

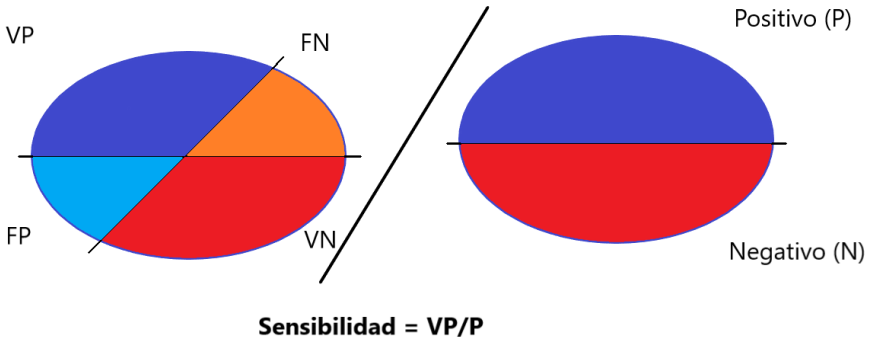

Figura 2.9. Sensibilidad de un clasificador

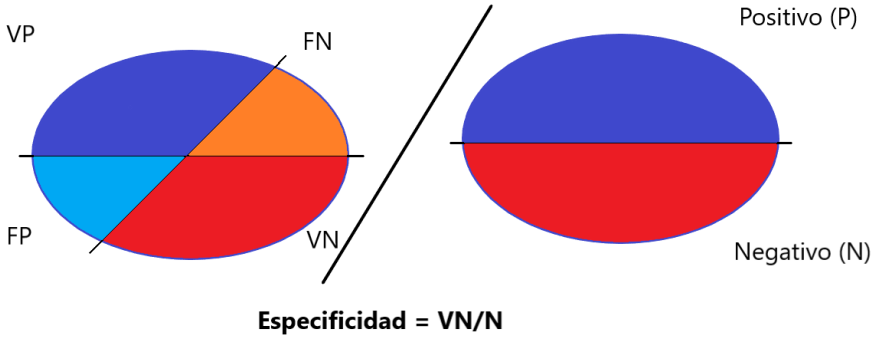

Figura 2.10. Especificidad de un clasificador

Otro aspecto para la evaluación es la pregunta de qué proporción de ejemplos, entre todos los asignados, realmente pertenecen a una determinada clase. Esto se determina mediante una métrica denominada porcentaje de acierto (Japkowicz & Shah, 2011). En la Figura 2.11 se ilustra el concepto de porcentaje de acierto (Takaya & Rehmsmeier, 2015).

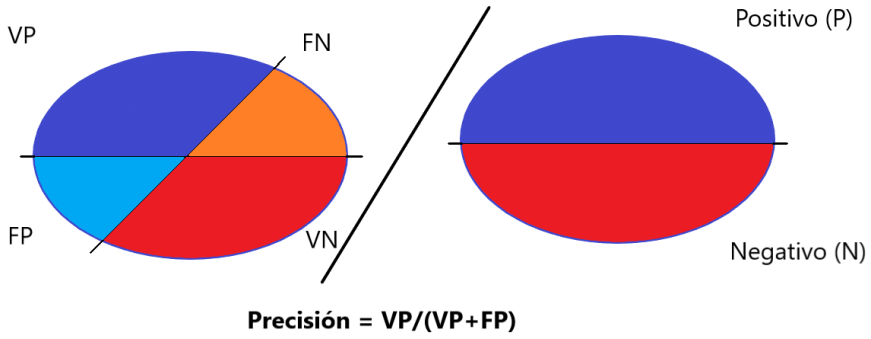

Figura 2.11. Precisión de un clasificador

2.5.2. Índice Kappa

El índice Kappa (κ), es un coeficiente propuesto originalmente por Cohen en 1960, que permite medir la concordancia entre los resultados de dos o más variables. Se utiliza comúnmente para cuantificar el grado de acuerdo entre diferentes clasificaciones con categorías idénticas (Yang & Zhou, 2015).

El índice calcula la diferencia entre la proporción de acuerdo observado y la proporción de acuerdo esperado por azar. Si es igual a cero, entonces el grado de acuerdo que se ha observado puede atribuirse enteramente al azar, alcanza el máximo de 1 sólo si hay acuerdo perfecto entre los observadores; si es positivo, indica que el grado de acuerdo es mayor que el que cabría esperar si solo estuviera operando el azar y viceversa: en el caso (ciertamente improbable) en que fuera negativo, los datos estarían exhibiendo menos acuerdo que el que se espera solo por concepto de azar.

El índice κ, aplicado a la tabla de confusión permite evaluar si la clasificación observada es similar (concordante) con la clasificación predicha por el clasificador. Para dos categorías se calcula:

$$\kappa = [(Po - Pe)/(1 - Pe)], 0 \leq \kappa \leq 1$$

Donde:

Po, es la proporción de aciertos $[(VP + VN)/T]$.

Pe, es la proporción de aciertos esperados bajo la hipótesis de independencia entre las dos variables $[(VP + FN) * (VP + FP) + (FP + VN) * (FN + VN)]/T2$

Hartling *et al.* (2012) propusieron una clasificación del índice κ como interpretación de la concordancia definida por sus valores, una adaptación de la misma se presenta en la Tabla 2.1.

Tabla 2.1. Escala de valoración de la concordancia por índice Kappa (adaptado de Hartling *et al*., 2012)

Valor κ	Interpretación de la concordancia
< 0,20	Pobre
0,21-0,40	Leve
0,41-0,60	Moderada
0,61-0,80	Buena
0,81-1,00	Muy buena

2.5.3. Métodos gráficos de evaluación de modelos de clasificación

Los gráficos de cajas (box-plot), tienen por objeto reflejar la forma de las distribuciones de las variables en estudio, dando en un mismo elemento gráfico información acerca de la mediana, la media, los cuantiles 0,05, 0,25, 0,75 y 0,95 y mostrando la presencia, si los hubiere, de valores extremos. Cuando se dispone de información sobre una variable medida en los individuos de una muestra, es posible calcular algunos valores, denominados estimadores o estadísticos, que permiten describir el comportamiento de dicha variable. Los estadísticos de posición se refieren a la ubicación de los datos sobre el campo de variación de la variable, los más utilizados son: los valores mínimo y máximo, la media aritmética, la mediana, la moda y los cuartiles 1 y 3. Los estimadores de variación o variabilidad hacen referencia a la forma en que se distribuyen los datos, es decir su mayor o menor heterogeneidad, los más empleados son: desviación estándar, variancia, coeficiente de variación y rango. Cuando se trata de evaluar modelos, en los gráficos de caja se representan los porcentajes de acierto para los diferentes modelos propuestos (Perelman *et al.*, 2019).

Los gráficos biplots permiten presentar en un mismo gráfico a las observaciones y las variables, de forma tal que se pueden hacer interpretaciones sobre las relaciones conjuntas. Las observaciones son generalmente graficadas como puntos y la configuración de estas es obtenida a partir de combinaciones lineales de las variables originales. Las variables son graficadas como vectores desde el origen y los ángulos entre ellos representan la correlación entre las variables (Di Rienzo *et al*., 2020).

2.6. Aplicaciones

En las últimas décadas se viene ensayando con la combinación de métodos para la resolución de problemas complejos de clasificación de muestras, así por ejemplo se utilizó el Análisis KNN (no supervisado) en la matriz generada luego de aplicar un Análisis Discriminante Linear LDA (método supervisado) para clasificar propóleos provenientes de Argentina (Cantarelli *et al.*, 2011). Otra posibilidad es la combinación de técnicas de reconocimiento de patrones y técnicas de regresión tal como el Análisis Discriminante Lineal por Cuadrados Mínimos Parciales (PLS-DA). Esta nueva metodología, así como el LDA, corresponden a métodos supervisados, por lo que su desarrollo es de gran utilidad (Barker & Rayens, 2003).

Arango *et al*. (2016), detectaron las zonas arables en un predio utilizando imágenes satelitales y técnicas de aprendizaje automático.

Diversas aplicaciones de estos métodos de inteligencia artificial se han encontrado para la clasificación de frutas y vegetales y el análisis de imágenes en la identificación o clasificación de frutos.

Hameed *et al*. (2018), realizaron una revisión de los métodos de aprendizaje automático (SVM, KNN, ANN y CNN – redes neuronales convolucionales) aplicados en los últimos años para clasificar frutas y vegetales. Describen como desafíos cuando se trabaja con imágenes, la selección de un sensor adecuado para la obtención de los datos; la definición de las características que permiten distinguir a los objetos a

clasificar; el conjunto de algoritmos usados para la clasificación y reconocimiento de imágenes.

Naik y Patel (2017) realizaron una revisión de la metodología utilizada para la clasificación y tipificación de frutas y discutieron el uso de KNN, SVM, ANN y CNN (redes neuronales convolucionales) y encontraron en los diferentes trabajos revisados valores de exactitud que variaban entre 97% (ANN), 95-100% (DA y SVM) en mango, 92% (PCA y SVM) en tomate, 86-98% (SVM) en uvas, >68% (Cluster Análisis y DA) en frutillas y 97% (ANN) en un mix de frutas. Alfanti *et al*. (2018) emplearon KNN, ANN y SVM para clasificar frutos de palma de aceite (*Elais guineensis*) por su grado de madurez y obtuvieron mayor exactitud con ANN (93%).

Gil *et al.* (2014) analizaron el comportamiento de diversas técnicas de inteligencia artificial para clasificar frutas a partir de imágenes. Para frutos de manzana, los clasificadores de mejor comportamiento fueron ANN, con 89,9% de exactitud, DA (lineal) y SVM con 90% de exactitud.

Zawbaa *et al.* (2014) emplearon algoritmos de aprendizaje automático para clasificar manzanas, frutillas y naranjas a partir de imágenes, encontraron que RF (exactitud de 85% en manzana, 87,50% en naranja y 90,91% en frutilla) presentó mejor comportamiento en la clasificación que KNN y SVM.

Pholpo *et al.* (2011) utilizaron PCA, PLS-DA (Análisis Discriminante por Mínimos Cuadrados Parciales) y Modelado Independiente por Analogía de Clases (SIMCA) para clasificar frutas longan (*Dimocarpus longan*) en golpeadas y no golpeadas; el PLS-DA presentó el mejor comportamiento en la clasificación con un 100% de tasa de éxito. Saeed *et al.* (2012) emplearon DA, SIMCA y KNN para caracterizar la maduración de los frutos de la palma de aceite; encontraron que el DA (con algoritmo cuadrático y distancia de Mahalanobis) presentó el mejor comportamiento, con una exactitud superior al 85%. Kim *et al.* (2009) buscaron desarrollar un método para detectar enfermedades en la cáscara de pomelos basado en características de la textura del color aplicando DA, lograron una exactitud general del 96,7%, mientras que mancha grasienta y melanosis, la exactitud fue del 90%.

Norasyikin *et al.* (2012), encontraron que el algoritmo de clasificación mediante un perceptrón multicapa fue adecuado para caracterizar la maduración de los frutos de la palma de aceite. Vijayarekha & Govindaraj (2006) utilizaron ANN para identificar mandarinas con y sin defectos, obteniendo un clasificador con 32 neuronas de entrada, 3 de salida y 10 capas ocultas; lograron un 84,21% de frutas bien clasificadas en el caso de frutas con picaduras, 50% en el caso de pudrición del tallo y 100% en caso de rajaduras. Alonso Salces *et al*. (2005) estudiaron los perfiles poli fenólicos de manzanas para sidra de acuerdo con su estado de maduración empleando DA, KNN y ANN a fin de definir reglas de decisión; las ANN presentaron un excelente comportamiento con éxitos en la predicción de 97% en la categoría de frutos inmaduros y el 99% para la de maduros. Astuti *et al.* (2018) comparan en uso de ANN y SVM para definir algoritmos a fin de realizar una clasificación automática de frutas a partir de imágenes, la exactitud alcanzada con SVM (100%) fue mejor que con ANN (50%) y con menor tiempo de entrenamiento. Sabanci *et al.* (2016) utilizaron KNN y ANN para clasificar diferentes variedades de manzana y encontraron que ANN presentó mejor comportamiento con un porcentaje de acierto de 98,89%.

Woo y Mirisaee (2009) aplicaron el método del vecino más cercano (KNN) para el reconocimiento de bananas, manzanas, frutillas, limones, sandías y basado en el color la forma y el tamaño, obteniendo un 90% de exactitud. Li *et al.* (2014) clasificaron arándanos en diferentes estadios de crecimiento mediante imágenes y encontraron que el KNN fue la técnica de mejor comportamiento con un acierto entre 85-98%.

Otros mecanismos de recolección de datos también han sido empleados, Buratti *et al*. (2004), utilizando nariz y lengua electrónica, clasificaron vinos italianos empleando PCA, DA y DT, obteniendo con el DA (lineal), el 100% de asignaciones correctas para un origen y errores de 3,77% para los otros orígenes; con DT obtuvieron errores de 13,21%.

También se han encontrado antecedentes de clasificación de origen geográfico de frutas. Pérez *et al.* (2006) estudiaron el origen de frutillas, arándano y pera a partir del perfil multielemental, aplicaron Análisis de la Variancia (ANOVA), AD con funciones lineal y cuadrática, ANN y redes neuronales genéticas (GNN). Todos los modelos

estudiados presentaron el 100% de exactitud en la clasificación de frutillas y arándanos, pero en el caso de las peras, con el DA lineal solamente obtuvieron entre 60-80% de acierto, con el DA cuadrático entre 85-100% y con ANN entre 80-90%, la técnica de mejor comportamiento fue GNN con 100% de acierto.

El grupo de trabajo que acompañó el desarrollo de la presente tesis doctoral posee probada experiencia en la combinación de estas técnicas quimiométricas con datos multielementales de distintos alimentos provenientes de nuestro país, tales como, cítricos (Pellerano *et al.*, 2008), mieles multiflorales (Pellerano *et al.*, 2012), hierbas medicinales (Cantarelli *et al.*, 2010) y porotos (Pérez Rodríguez *et al.*, 2019).

2.7. Referencias

Alfanti, MS; Shariff, ARM; Bejo, SK; Saaed, OMB; Mustapha, O. 2018. Real time oil palm FFB ripeness grading system based on ANN, KNN y SVM classifiers. 2018IOP Conf. Series: Earth and Environmental Science. 169. 012067.

Alonso Salces, RM; Herrero, C; Barranco, A; Berrueta, LA; Gallo, B; Vicente, F. 2005. Classification of apple fruits according to their maturity state by the pattern recognition analysis of their polyphenolic compositions. Food Chemistry. 93: 113-123.

Arango, RB; Díaz, I; Campos, AM; Canas, ER; Combarro, EF. 2016. Automatic arable land detection with supervised machine learning. Earth Sciences Informatics. 9(4): 535-545.

Astuti, W; Dewanto, S; Soebandrija, KEN; Tan, S. 2018. Automatic fruit classification using support vector machines: comparison with artificial neural network. 2nd International Conference on Eco Engineering Development 2018. IOP Conferences Series: Earth and Environmental Sciences 195 012047.

Ballabio, D; Consonni, V. 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods. 5 (16): 3790-3790.

Barker, M; Rayens, W. 2013. Partial Least Squares for Discrimination. Journal of Chemometrics. 17 (3): 166-173.

Breiman, L; Friedman, J; Stone, CJ; Olshen, RA. 1984. Classification and Regression Trees. Taylor and Francis, 368 pp.

Brereton, R. 2009. Chemometric for pattern recognition. John Wiley & Sons. 503 pp.

Buratti, S; Benedetti, S; Scampicchio, M; Pangerod, EC. 2004. Characterization and classification of Italian Barbera wines using an electronic nose and an amperometric electronic tongue. Analytica Chimica Acta. 525: 133-139.

Cantarelli, M; Pellerano, R; Del Vitto, L; Marchevsky, E; Camiña, J. 2010. Characterization of two south american food and medicinal plants by chemometric methods based on their multielemental composition. Phytochemical Analysis. 21 (6): 550–555.

Cantarelli, M; Camiña, J; Pettenati, E; Marchevsky, E; Pellerano, R. 2011. Trace mineral content in argentenean raw propolis by INNA. LWT - Food Science and Technology. 44: 256-260.

Cuadras, C.M. 2010. Métodos de Análisis Multivariante. CMC Editions. Manacor 30. 08023 Barcelona, España. 278 pp.

Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W. InfoStat versión 2020. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar

Gill, J; Sandhu, PS; Singh, T. 2014. A review of automatic fruit classification using soft computing techniques. International Conference on Computers, Systems and Electronic Engineering (ICSCEE 2014). 91-98.

Hair, JK Jr; Anderson, RE; Tatham, RL; Black, WC. 2005. Análisis Multivariante, 5ta. Ed. Prentice Hall Iberia, Madrid. ISBN:84-8322-035-0. 799 pp.

Hameed, K; Chai, D; Rassau, A. 2018. A comprehensive review of fruits and vegetable classification techniques. IMAVIS. Doi: 10.1016/j.imavis.2018.09.016.

Hartling, L; Hamm, M; Milne, A; VandermeerB; Santaguida, L; Ansari, M; Tsertsvadze, A; Hempel, S; Shekelle, P; Dryden, DM. 2012. Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 Mar. Table 2, Interpretation of Fleiss' kappa (κ) (from Landis and Koch 1977). Disponible en línea: https://www.ncbi.nlm.nih.gov/books/NBK92295/table/methods.t2/. Visita: 05/04/2020.

Hastie, T; Tibshirani, R; Friedman. J. 2017. The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2nd Edition. Springer Series. 446 pp.

James, G; Witten, D; Hastie, T; Tibshirani, R. 2017. An Introduction to Statistical Learning with Applications in R. Springer Science+Business Media New York 8th printing. 425 pp.

Japkowicz, N; Shah, M. 2011. Evaluating Learning Algorithms: A Classification Perspective (pp. I-Vi). Cambridge: Cambridge University Press.

Johnsson, DE. 2000. Métodos Multivariantes aplicados al análisis de datos. México: Internacional Thomson Editores.

Kim, DG; Burks, TF; Qin, J; Bulanon, DM. 2009. Classification of grapefruit peel diseases using color texture feature analysis. International Journal of Agriculture & Biological Engeeniering. 2 (3): 41-50.

Kumar, N; Bansal, A; Sarma, G; Rawal, R. 2014. Chemometrics tools used in analytical chemistry: an overview. Talanta.123:186-199.

Li, H; Lee, WS; Wang, K. 2014. Identifying blueberry fruits of different growth stages using natural outdoors color images. Computers and Electronics in Agriculture. 106: 91-101.

Mitchel, TM. 1997. Machine Learning. Ed. Mc Graw Hill. ISBN: 0070428077. 423 pp.

Naik, S; Patel, B. 2017. Machine vision-based fruit classification and grading – A review. International Journal of Computer Applications. 170 (9): 22-34.

Norasyikin, F; Junita, MS; Haidi, I; Zaini, AH. 2012. Oil palm fresh fruit bunch ripeness classification using artificial neural network. 2012 4th International Conferences on Intelligence and Advances Systems (ICIAS2012). 18-21.

Otto, M. 2007. Chemometrics, Statistics and Computers Application in Analytical Chemistry, 2nd edition. Wiley-VCH, Weinheim.

Pellerano, RG; Mazza, SM; Marigliano, RA; Marchevsky, EJ. 2008. Multielement Analysis of Argentinean Lemon Juices by Instrumental Neutronic Activation Analysis and Their Classification According to Geographical Origin. Journal of Agricultural and Food Chemistry. 56 (13): 5222-5225.

Pellerano, R; Uñates, M; Cantarelli, M; Camiña, J; Marchevsky, E. 2012. Analysis of trace elements in multifloral Argentine honeys and their classification according to provenance. Food Chemistry. 134 (1): 578-582.

Peña, D. 2002. Análisis de Datos Multivariantes. Madrid: Mc Graw Hills/ Interamericana de España. 540 pp.

Perelman, SB, Garibaldi, LA; Tognetti, PM. 2019. Experimentación y Modelos Estadísticos. Ed. Facultad de Agronomía. Universidad de Buenos Aires. 475 pp.

Pérez, AL; Smith, BW; Anderson, KA. 2006. Stable Isotope and trace element profiling combine with classification models to differentiate geographic growing origin for three fruits: effects of subregion and variety. Journal of Agricultural and Food Chemistry. 54: 4506-4516.

Pérez Rodríguez, M; Gaiad, J; Hidalgo, M; Avanza, M; Pellerano, R. 2019. Classification of cowpea beans using multielemental fingerprinting combined with supervised learning. Food Chemistry. (95): 232-241.

Pholpo, T; Pathaveerat; Sirisonboom, P. 2011. Classification of longan fruit bruising using visible spectroscopy. 2011. Journal of Food Engineering. 104: 169-172.

Quinlan, JR. 2014. C 4.5: Programs for Machine Learning. Morgan Kaufmann Pubishers. California. 301 pp.

Raschka, S. 2015. Python Machine Learning. UK: Packt Publishing Ltd. 454 pp.

Russel, S; Norving, P. 2004. Artificial Intelligence – A Modern Approach. 2nd Edition. Prentice Hall. Upper Saddle River, NJ. 1240 pp.

Sabanci, K; Ünlersen, MF; Polat, K. 2016. Classification of different forest types with machine learning algorithms. Research for Rural Development. 22nd Annual International Scientific Conference Research for Rural Development. 1, 254-260.

Saeed, OMB; Sankaran, S; Shariff, ARM; Shafri, HZM; Ehsani, R; Alfatni, MS; Hazir, MHM. 2012. Classification of oil palm fresh brunches based on their maturity using portable four-band sensor system. Computers and Electronics in Agriculture. 82: 55-60.

Takaya, S; Rehmsmeier, M. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Plos One. Disponible en línea: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432. https://doi.org/10.1371/journal.pone.0118432. Visita: 02/05/2020.

Varmuza, K; Filzmoser, P. 2016. Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton. 336 pp.

Vijayarekha, K; Govindaraj, R. 2006. Citrus fruit external defect classification using wavelet packed transform features and ANN. Conference: Industrial Technology, 2006. ICIT 2006. 2872-2877.

Williams, G. 2011. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery New York: Springer Science & Business Media.

Woo, CS; Mirisaee, SH. 2009. A new method for fruits recognition system. 2009 International Conference on Electrical Engineering and Informatics. AI-26: 130-134.

Yang, Z; Zhou, M. 2015. Weighted Kappa Statistic for Clustered Matched-Pair Ordinal Data. Computational Statistics and Data Analysis. 82: 1–18.

Zawbaa, HM; Hazman, M; Abbas, M; Hassanien, AE. 2014. Automatic fruit classification using Random Forest algorithm. 2014 International Conference on Hybrid Intelligent Systems (HIS). 164-168.

CAPÍTULO III

MATERIALES Y MÉTODOS

En este capítulo se describe la obtención de muestras, incluyendo: el área geográfica de procedencia, las especies y variedades, los métodos de muestreo, los procesos de pretratamiento y acondicionamiento y los procedimientos analíticos para la determinación de la composición multielemental de los jugos de fruta cítrica estudiados.

3.1. Área de estudio

En esta tesis el estudio se centró en las dos regiones de mayor importancia en la producción de citrus de la Argentina: el Noroeste Argentino (NOA) y el Noreste Argentino (NOA) (Figura 3.1).

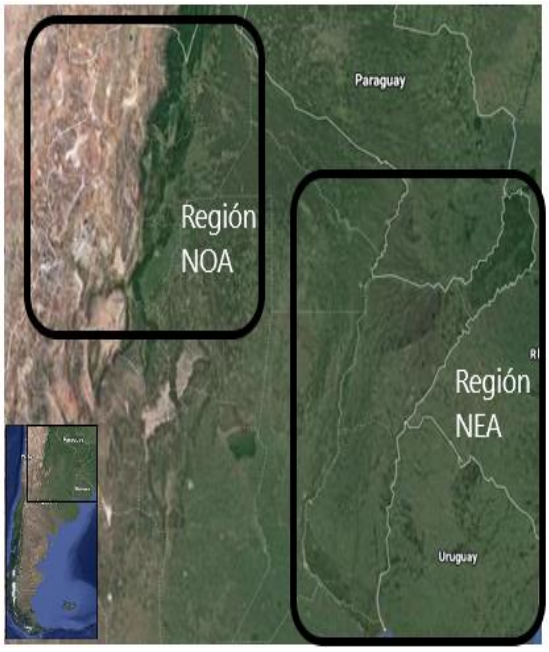

Figura 3.1. Regiones en estudio

3.1.1. Características de la Región NOA

El NOA abarca las provincias de Jujuy, Salta, Tucumán, Catamarca, La Rioja y Santiago del Estero, a los efectos de esta tesis se consideraron las provincias de Tucumán y Jujuy, por ser las de mayor concentración de la producción citrícola.

La provincia de Tucumán está integrada por diferentes unidades de relieve como son las zonas de llanuras, que cubren el este y centro de la provincia; las áreas de montañas, que abarcan el sector occidental y del noreste, limitando con la unidad anterior y una serie de cuencas y/o valles inter montanos distribuidos en diferentes sectores. Diversos factores, entre los que se encuentra el relieve, el clima, los materiales originarios y la cobertura vegetal, condicionan una gran variabilidad espacial en los suelos, de acuerdo con la clasificación de la USDA (Departamento de Agricultura de Estados Unidos) los suelos corresponden a cinco órdenes taxonómicos: Molisol, Entisol, Alfisol, Inceptisol y Aridisol (Puchulu y Fernández, 2014).

El clima de Tucumán presenta variedades debido al relieve de su superficie. Aunque mayormente posee un clima cálido subtropical con estación seca, con temperaturas templadas o calurosas en función de las variaciones de altitud. Las precipitaciones tienen un régimen con características monzónicas: torrenciales y estacionales. Durante los meses de abril a octubre tiene lugar la temporada húmeda o de lluvias, donde se acumula el 90% de las precipitaciones totales. La pluviosidad anual supera los 1000 mm (aproximadamente 100 – 105 días con precipitaciones al año). En la región occidental la acumulación de lluvias puede llegar a los 3000 mm. Las temperaturas máximas alcanzan los 30 a 31ºC durante los meses de diciembre y enero, con mínimas de 6 a 7ºC en los meses de julio y agosto[10].

La provincia de Jujuy presenta una acentuada continuidad de clima y relieve con la provincia de Salta, razón por la cual se encuentran en su territorio las mismas unidades de paisaje que en esta última. Sin embargo, Jujuy tiene todos esos paisajes en un área mucho menor (un tercio de la de Salta), lo que la hace más curiosa desde

[10] http://www.ora.gov.ar/informes/atlas_noa_precipitaciones.pdf

un punto de vista geográfico. Según el clima, la provincia de Jujuy puede ser dividido en: región templada, región subtropical o cálida, quebrada de Humahuaca y sierras subandinas. La región templada es la región agrícola-ganadera de la provincia, se caracteriza por suelos fértiles, presenta un régimen pluvial subtropical, es decir: veranos lluviosos e inviernos secos[11].

3.1.2. Características de la Región NEA

La región productora citrícola del Nordeste Argentino (NEA) está integrada por las provincias de Misiones, Corrientes, Entre Ríos, Chaco y Formosa. En esta tesis, se consideraron las tres provincias con mayor relevancia en la actividad citrícola: Misiones, Corrientes y Entre Ríos, correspondientes al NEA.

Corrientes, emplazada entre los paralelos 28 y 30°S, presentó, durante la campaña 2016/2017, una superficie de 25.508 ha citrícolas, de las cuales 54,3% se destinaron a la producción de naranjas, 33,3% a mandarinas, 10,6% a limones y 1,9% a pomelos (Federcitrus, 2018). Sus plantaciones más importantes se ubican sobre los márgenes y paleocauces de los ríos Paraná y Uruguay. La zona del río Paraná (centro oeste de la provincia), comprende a los departamentos de Bella Vista, Mburucuyá, Saladas, Concepción, General Paz, San Miguel y San Roque, en la que se destaca la superficie implantada con mandarinos, naranjos y limoneros. En la zona del río Uruguay (sudeste de la provincia), ubicada en las áreas cercanas a los departamentos de Monte Caseros y Mocoretá, la producción está orientada principalmente a variedades de naranjos y mandarinos (Molina, 2006; Palacios, 2013).

La provincia de Corrientes presenta una gran variedad fitogeográfica, aunque muy modificada por la actividad humana, producto del contacto de distintas formaciones vegetales: sabanas con herbazales, especies hidrófilas en las áreas

[11] http://www.ora.gov.ar/informes/atlas_noa_precipitaciones.pdf

pantanosas, bosques en galería en las riberas de los ríos y superficies aisladas de bosques en medio de pastizales (Carnevali, 1994).

La región presenta clima subtropical húmedo, muy cálido en verano, pero con presencia de heladas en invierno. Tiene características de clima húmedo, con frecuentes excesos hídricos en otoño y primavera, y moderados y eventuales déficit, principalmente en verano. La temperatura media anual varía entre 21°C al Norte y 19°C al Sur. Las temperaturas estivales e invernales relativamente elevadas y su variación anual definen al clima correntino. Las lluvias son abundantes y frecuentes, superando los 1.500 mm anuales en el NE descendiendo gradualmente hasta menos de 1.000 mm en el ángulo SO. La principal característica de este régimen es su irregularidad. La distribución anual de precipitaciones tiene dos máximos, en primavera y en otoño, y un mínimo en invierno. El número de días con precipitaciones varía muy poco entre los meses del año y se encuentra entre 6 y 8 días al mes, que totalizan alrededor de 80 a 100 días de ocurrencia de lluvias por año (Carnevali, 1994, Escobar *et al.*, 1996).

La humedad relativa media anual oscila en todas las localidades de la provincia entre 70 y 75%, siendo mínimos los valores en verano, y máximos en invierno, según la variación inversamente proporcional a la temperatura. Esta elevada humedad promedio es el resultado de la enorme cantidad de cuerpos de agua que caracterizan el territorio provincial, no solo por los ríos y arroyos, sino también por lagunas, esteros y cañadas de variada extensión (Escobar *et al.*, 1996).

La región en estudio presenta suelos bien drenados, lavados y ácidos, con características físicas óptimas para el cultivo de citrus. El tipo de suelo que predomina fue clasificado como Udipsament Álfico, perteneciente al orden Entisol, caracterizado por la predominancia en el contenido de arena, pH ácido, mínima fertilidad manifestada por su baja capacidad de intercambio catiónica, poco contenido de carbono orgánico y fósforo asimilable. Los Entisoles tanto de régimen údico como ácuico, que contienen no más de 7% de arcilla en los primeros 30 cm (Udipsamentes, Psamacuentes), presentan ausencia de estructura y su matriz es en general de grano simple (Escobar *et al.*, 1996).

Entre Ríos, situada entre los paralelos 30 y 32°S, es la primera provincia productora de naranjas y mandarinas del país. Los centros de producción más importantes son los departamentos de Federación, Concordia y una pequeña fracción del departamento de Colón, todas ubicadas sobre el río Uruguay (noreste de la provincia) (Palacios, 2013). La superficie plantada, de 36.386 ha se distribuye en un 54,0% destinada a la producción de naranjas, 42,2% a mandarinas, 1,7% a limones, 2,1% a pomelos (Federcitrus, 2018).

La provincia de Entre Ríos, junto con Santa Fe y Córdoba, casi toda Buenos Aires y el este de La Pampa, pertenecen a la región fitogeográfica Pampeana que ocupa las llanuras del este de la República Argentina entre los 31°S y 39°S, aproximadamente. Al norte, oeste y sur limita con la región del Espinal, al este y sudeste con el Océano Atlántico. Se extiende sobre llanuras horizontales o muy poco onduladas, con algunas serranías de poca altura. La vegetación dominante es la estepa de gramíneas. El clima es templado cálido, con precipitación anual de 600-1100 mm, en forma de lluvias durante todo el año que disminuyen de norte a sur y de este a oeste y temperaturas medias de 13-17°C. Esta región puede ser subdividida en cuatro subregiones, la provincia de Entre Ríos comprende la zona norte, donde es más húmeda y se caracteriza por la abundancia de gramíneas subtropicales que forman grandes flechillares dominados por los géneros *Paspalum*, *Axonopus* y *Digitaria*, entre otros (Apodaca *et al.*, 2015).

Misiones presenta actividad citrícola distribuidas en varias zonas de su territorio, pero la mayor concentración de producción se encuentra en el centro sur de la provincia. Durante la campaña 2016/2017 se registró una superficie de 6.198 ha citrícolas, de las cuales 30,5% se destinaron a la producción de naranjas, 47,9% a mandarinas y 12,7% a limones. La producción total de citrus estuvo en las 48.587 t, de las que corresponde 14.353 t a naranjas, 19.412 t a mandarinas y 7.411 t a limones (Federcitrus, 2018).

La provincia de Misiones se encuentra situada en el ángulo nordeste de la República Argentina, casi inmediatamente al sur del trópico de Capricornio, es decir, en la zona subtropical. Tiene la forma de un pentágono irregular, alargado en sentido

NE-SO, alcanzando en tal orientación un máximo largo de 375 kilómetros, mientras su ancho mínimo - entre los ríos Paraná y Uruguay - es de 70 kilómetros. Ocupa una superficie de 29.801 kilómetros cuadrados y su perímetro, de 1.200 Km de longitud, ofrece 1.080 Km como fronteras internacionales: unos 330 km sobre el Río Paraná con Paraguay y unos 750 Km sobre los ríos Uruguay, Pepirí Guazú, San Antonio e Iguazú con Brasil.

Misiones se localiza en el sector sudoccidental de la Gran Cuenca Sedimentaria del Paraná, y corresponde al Planalto Meridional del Brasil, región Alto Paraná-Alto Uruguay, actuando como una divisoria de aguas entre las cuencas de estos dos grandes ríos. Desde el punto de vista fitogeográfico, se ubica en el Dominio Amazónico y forma parte de las Selvas Subtropicales, con tendencia a su sustitución por bosques cultivados de pinos y agricultura (Cabrera, 1976). El mismo autor señala que la Provincia Paranaense incluye a dos distritos, el Distrito de las Selvas Mixtas y el Distrito de los Campos. Este último se extiende por el sudoeste de Misiones y nordeste de Corrientes, donde se funde en complejo ecotono con la Provincia Chaqueña. Los suelos del Distrito de los Campos son también lateríticos y el clima no difiere mucho del Distrito de las Selvas Mixtas, si bien la precipitación es ligeramente menor y la sequía invernal más marcada (Gunther *et al*, 2008).

En esta tesis se trabajó con muestras de frutas cítricas recolectadas en origen provenientes de las provincias de Tucumán, Jujuy y Corrientes en el caso de los limones; Entre Ríos, Corrientes y Misiones, las mandarinas y naranjas.

3.2. Variedades estudiadas

3.2.1. Variedades de limonero

Los frutos de limonero (*Citrus limón* L) estudiados corresponden a las tres variedades más producidas en la actualidad a nivel nacional: 'Eureka', 'Lisboa' y 'Génova'.

La variedad 'Eureka' es la más cultivada en el país, es originaria de los Ángeles (EE. UU.). Los árboles son medianamente grandes y muy productivos, las frutas son medianas, algo alargadas y con cáscara algo gruesa en la producción de invierno, mientras que en las producciones de verano toma una forma más redondeada y cáscara fina. De las tres variedades estudiadas es la que se caracteriza por presentar producciones importantes en verano e invierno (Palacios, 2013).

La variedad 'Lisboa' es originaria de California (EE. UU.), es la más productiva del NOA. Los árboles se caracterizan por presentar la producción dentro del follaje lo que le permite obtener fruta de mejor calidad y mayor tolerancia al frío. Concentra su floración en primavera (cerca del 90%), por lo que la cosecha se realiza en invierno, período más importante de exportación e industrialización de la fruta (Palacios, 2013).

La variedad 'Génova' tiene su origen posiblemente en Italia, es muy cultivada en Tucumán, donde se presentan diversas líneas. Las plantas son medianamente vigorosas con la producción hacia la periferia, el follaje es muy denso y las hojas oscuras. La mayor producción se presenta en invierno, aunque florece en varios períodos. Su principal característica es la precocidad. Las frutas son de tamaño mediano, con buen contenido de jugo y acidez (Palacios, 2013).

3.2.2. Variedades de mandarino

Entre las diferentes variedades de mandarino y sus híbridos que se producen en el NEA fueron seleccionados dos, 'Okitsu' (*Citrus unshiu* Marcovitch), de maduración extra temprana y el tangor 'Murcott' (*Citrus sinensis* L. x *Citrus deliciosa* Tenore), de maduración más tardía, con buen comportamiento comercial y posibilidades futuras.

El variedad 'Okitsu' integra el grupo de las denominadas mandarinas Satsuma, originaria de Japón, dentro de este grupo es posiblemente la variedad más difundida en Argentina. Los árboles son de tamaño medio, con hábito de crecimiento semiabierto y algo llorón, sus ramas tienen tendencia al crecimiento en zigzag, son

resistentes al frio. Los frutos son de buen tamaño, mediano, de forma redondeada ligeramente achatada y no presentan semillas. La corteza es espesa y rugosa, se separa fácilmente de los gajos y su color es amarillo naranja o naranja asalmonado con algunas tonalidades verdosas (CSC, 2008; Palacios, 2013).

El tangor 'Murcott' es un híbrido (*Citrus sinensis* L. x *Citrus deliciosa* Tenore), producto de un cruzamiento realizado en Florida. Las frutas se presentan en racimos, son de tamaño mediano, con muchas semillas y color amarillo anaranjado a la madurez. La pulpa es muy jugosa y dulce, lo que la hace muy apetecible. Si bien la planta tolera muy bien las heladas, la fruta no las resiste y cae (CSC, 2008; Palacios, 2013).

3.2.3. Variedades de naranjo

Entre las diferentes variedades de naranjo dulce (*C. sinensis* L) que se producen en el NEA fueron seleccionadas dos, 'Valencia late', de maduración tardía y 'Salustiana', de maduración intermedia. El criterio utilizado para esta selección fue, además de abarcar diferentes períodos de maduración, trabajar con una variedad representativa en cuanto a la superficie implantada, tradicional y de relevancia económica actual, por lo que se seleccionó a 'Valencia late', y con una emergente, por lo que se decidió trabajar con 'Salustiana' que, aunque presenta menor superficie cultivada, se encuentra presente en todas las zonas que son objeto de este estudio.

Dentro de las variedades de naranjo de maduración tardía, 'Valencia late' adquiere la mayor importancia en el país y en el mundo (el 60% de los naranjales del mundo pertenecen a esta variedad). Se originó como una mutación de naranjo dulce de China (probablemente Xuegan), y se introdujo a Valencia, España, a mediados del siglo XV. En 1870, fue importada por viveristas británicos desde las islas Azores a EE. UU., donde ha logrado un gran desarrollo y expansión. Sus principales características son: árbol medio-vigoroso, con hojas de 9-9,5 cm x 3,8-4 cm, de forma ovalo-elípticas, con bordes dentados y pecíolo alado pequeño. Flores de tamaño medio, polen y

megaspora parcialmente estériles. Frutos elíptico-esferoidales, de color naranja, con pocas semillas de 2 a 5 por fruto, y de fácil pelabilidad (CSC, 2008; Palacios, 2013).

El variedad 'Salustiana' pertenece a las naranjas llamadas “grupo de las blancas”, cuyo origen se encuentra en una mutación que se ha producido en la provincia de Valencia atendiendo a las condiciones climáticas y del suelo de esta tierra. El árbol es normalmente, vigoroso, de ramas fuertes, frondoso y presenta un tamaño medio incluso grande. Se muestra sensible al frío y los cambios de temperatura. Su recolección se realiza a fines del otoño o principios del invierno. Los frutos son redondeados o en ocasiones achatados, presentando una corteza fina, sin semillas, que ofrece abundante zumo dulce y de poca acidez, siendo además una variedad que soporta muy bien el paso del tiempo una vez recolectada manteniendo sus propiedades. Es una naranja clásica que lleva en el mercado más de 50 años y que satisface las papilas gustativas de la mayor parte de consumidores, por lo que es apta y recomendable tanto para el consumo directo como para consumirlas en jugos (Palacios, 2013).

3.3. Muestras

A continuación, se describen los procedimientos para la obtención de muestras, su pretratamiento y los procedimientos analíticos para la obtención de los contenidos de los diferentes elementos minerales estudiados en los jugos cítricos.

3.3.1. Obtención de las muestras de frutos

Se analizaron 74 muestras de frutos de limonero recolectados en diferentes cooperativas y productores agrícolas durante la campaña 2014/2015, mediante un muestreo al azar simple, sobre un padrón de productores con un nivel tecnológico

medio. Los frutos obtenidos corresponden a tres variedades: 'Eureka', 'Lisboa' y 'Génova'. Se seleccionaron cuatro lugares diferentes en la región norte de Argentina para la recolección de frutas (Figura 3.2.). Tres sitios ubicados en dos de las provincias correspondientes a la región NOA: Tucumán (TN-I: Tafí Viejo y TN-II: Famaillá), y Jujuy (JY: Santa Clara) y un sitio correspondiente a la región NEA: Corrientes (CTE: Bella Vista). Cada muestra estuvo conformada por el jugo de diez frutos de una misma planta, los que fueron procesados conjuntamente. Luego de recolectados los frutos, fueron identificados y transportados en bolsas de red plásticas.

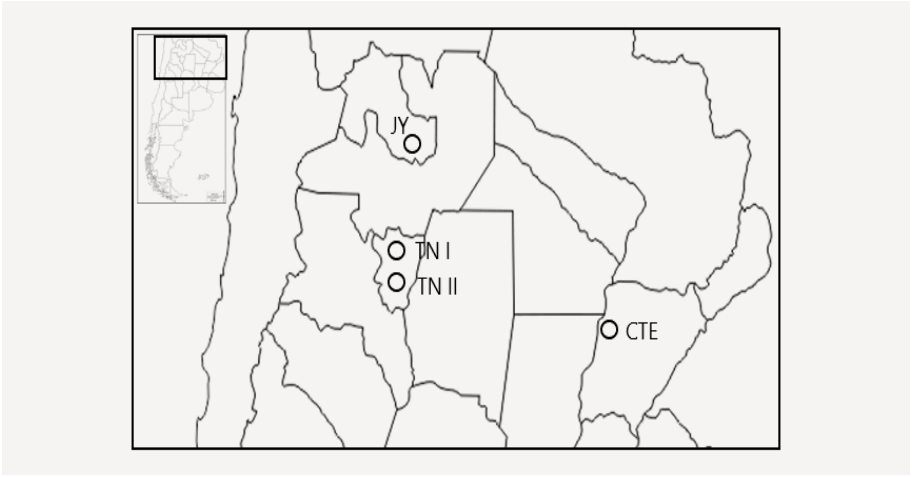

Figura 3.2. Zonas de muestreo de frutos del norte argentino en las provincias de Jujuy (JY), Tucumán (TN-I y TN-II) y Corrientes (CTE)

En la Tabla 3.1 se describe el esquema de muestreo utilizado.

Tabla 3.1. Esquema de muestreo de limones

Campaña	Variedad	Sitios	Muestras
2014/2015	Eureka	CTE	6
		JY	6
		TN-I	7
		TN-II	7
	Genova	CTE	6
		JY	6
		TN-I	6
		TN-II	6
	Lisboa	CTE	6
		JY	6
		TN-I	6
		TN-II	6

Las muestras de naranjas y mandarinas fueron recogidas durante las campañas 2015/2016 a 2017/2018 en huertos comerciales ubicados en las zonas más representativas de la producción citrícola del NEA y en las que se encuentra la mayor concentración de productores: centro sur de Misiones (CSMN), centro oeste de Corrientes (COCR), sudeste de Corrientes (SECR) y noreste de Entre Ríos (NEER) (Figura 3.3).

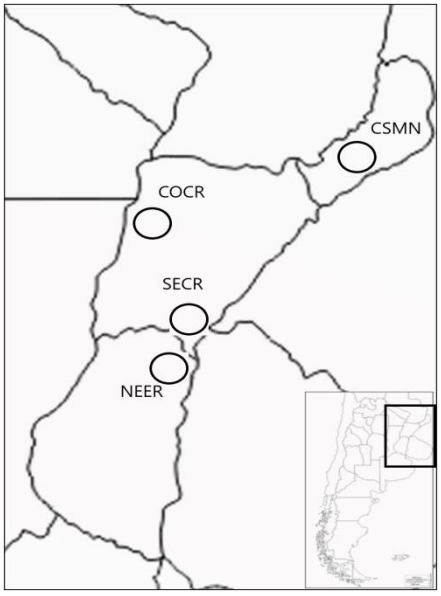

Figura 3.3. Zonas de producción donde se realizaron los muestreos de frutos del nordeste argentino, en el centro sur de Misiones (CSMN), centro oeste de Corrientes (COCR), sudeste de Corrientes (SECR) y noreste de Entre Ríos (NEER)

Para la selección de las muestras de frutos de naranjo y mandarino se utilizó un padrón integrado por productores que presentan un nivel tecnológico medio (representativo de la mayoría de los productores), emplean técnicas de cultivo comunes en la región, atienden a la diversidad de calidad de su fruta y comercializan su producción en distintos mercados (interno de frutos frescos, de exportación de frutos frescos y de industria para elaborar jugos o derivados).

Sobre la base de dicho padrón, para cada variedad se siguió un muestreo en etapas:

En una primera etapa se dividió por zona de producción, en cuatro estratos: noreste de la provincia de Entre Ríos, en las localidades de Villa del Rosario y Chajarí (NEER); sureste de la provincia de Corrientes, en las localidades de Mocoretá y Monte

Caseros (SECR); centro oeste de la provincia de Corrientes, en las localidades de Bella Vista, Tabay y Santa Rosa (COCR); centro-sur de la provincia de Misiones, en las localidades de Alem y 2 de Mayo (CSMN). En cada uno de los estratos, se utilizó un muestreo aleatorio simple para seleccionar 5 lotes.

En una segunda etapa, en cada uno de los lotes escogidos se empleó un método al azar sistemático para seleccionar 5 plantas, ubicadas a lo largo de una línea que recorría el lote de NO a SE, de las que se extrajeron 10 frutos por planta. Los frutos de una misma planta conformaron una unidad muestral y se procesaron conjuntamente.

A partir de dicho procedimiento, se obtuvieron 200 muestras de mandarino (100 muestras de cada variedad), y 200 muestras de naranjo dulce (100 de cada variedad), que se utilizaron para las determinación del contenido multielemental mediante MP-AES. Para la determinación del contenido multielemental mediante FAAS, se trabajó con 120 muestras de naranjo dulce (60 muestras de cada variedad), obtenidas durante las dos primeras campañas siguiendo el mismo procedimiento.

En la Tabla 3.2 se presenta el esquema de muestreo utilizado para mandarinas y naranjas.

Tabla 3.2. Esquema de muestreo de mandarinas y naranjas.

Campaña	Zona productora	Variedades de mandarina	Lotes	Muestras	Variedades de naranjas	Lotes	Muestras	
2015/2016	COCR	Murcott	2	10	Salustiana	2	10	*
	CSMN		2	10		2	10	*
	NEER		2	10		2	10	*
	SECR		2	10		2	10	*
	COCR	Okitsu	2	10	Valencia late	2	10	*
	CSMN		2	10		2	10	*
	NEER		2	10		2	10	*
	SECR		2	10		2	10	*
2016/2017	COCR	Murcott	1	5	Salustiana	1	5	*
	CSMN		1	5		1	5	*
	NEER		1	5		1	5	*
	SECR		1	5		1	5	*
	COCR	Okitsu	1	5	Valencia late	1	5	*
	CSMN		1	5		1	5	*
	NEER		1	5		1	5	*
	SECR		1	5		1	5	*
2017/2018	COCR	Murcott	2	10	Salustiana	2	10	
	CSMN		2	10		2	10	
	NEER		2	10		2	10	
	SECR		2	10		2	10	
	COCR	Okitsu	2	10	Valencia late	2	10	
	CSMN		2	10		2	10	
	NEER		2	10		2	10	
	SECR		2	10		2	10	

* Muestras utilizadas para la determinación con espectroscopía de absorción atómica por llama.

Los frutos fueron recolectados en el momento en que alcanzaron el punto de maduración comercial, identificados y transportados en bolsas de red plásticas.

La maduración organoléptica comercial hace referencia al proceso por el cual las frutas adquieren las características sensoriales que las definen como comestibles o aptas para su procesamiento posterior, se dispone de índices para determinar el momento óptimo de recolección. En este trabajo se evaluó el punto de maduración de los frutos teniendo en cuenta parámetros anatómicos (color, tamaño y textura), y parámetros fisicoquímicos tales como sólidos solubles (refractometría, °Brix), acidez titulable, ratio (sólidos solubles/acidez titulable) y porcentaje de jugo. Todas estas mediciones se realizaron en laboratorio y fueron utilizadas para decidir el momento de inicio de la cosecha.

De acuerdo con lo establecido en el Protocolo de Calidad para fruta fresca cítrica, se consideraron los siguientes requerimientos de calidad en los frutos muestreados (Código Alimentario Argentino, 2020):

Requerimientos generales para los cítricos:

Las frutas deben estar enteras, firmes, libre de manchas, libre de lesiones de distinto origen, libres de podredumbres, limpias, bien desarrolladas, libres de enfermedades, libres de cochinillas, exentas de humedad externa anormal (salvo la condensación consiguiente a su remoción de una cámara frigorífica), exentas de cualquier olor y/o sabor extraño.

Las frutas deberán presentar un desarrollo óptimo al ser cosechadas, y un estado tal que les permita soportar el transporte y la manipulación, y llegar en estado satisfactorio al lugar de destino.

Madurez: se establecerá sobre la base de la cantidad de jugo (porcentaje) y sobre la relación sólidos solubles totales-acidez (índice de madurez), que se determinará de acuerdo con las técnicas operativas previstas en el apartado Nº130 de la Resolución SAG N°145/83, Fruticultura - Frutas frescas cítricas.

El grado de color deberá ser tal que, después de un desarrollo normal los frutos tengan el color normal típico de la especie y variedad de que se trate, cubrirá como mínimo el 70% de la superficie total de cada unidad.

El calibre se determina por el diámetro máximo de la sección ecuatorial del fruto.

Requerimientos particulares por especie:

Los limones deben presentar un mínimo de 35% de jugo para exportación y un mínimo del 30% para el mercado nacional.

Las mandarinas (independientemente de su destino final) contendrán como mínimo, entre 30 y 35 % de jugo y una relación sólidos solubles/acidez de 7 a 1.

Las naranjas para exportación deben presentar un mínimo de 40% de jugo y una relación sólidos solubles/acidez de 6 a 1 y para el mercado nacional un mínimo del 35% de jugo y una relación sólidos solubles/acidez de 6 a 1.

La toma de muestras fue posible en virtud de que el grupo de trabajo de la Cátedra de Fruticultura de la Facultad de Ciencias Agrarias de la Universidad Nacional del Nordeste cuenta con los acuerdos y antecedentes que aseguraron la colaboración de los productores y el acceso a las frutas en las distintas épocas de cosecha.

3.3.2. Obtención de jugos

El pretratamiento y posterior acondicionamiento de las muestras de limones, y los procedimientos analíticos para la determinación de la composición multielemental de los jugos de limón se realizó en el Instituto de Química de San Luis (INQUISAL).

Una vez en el laboratorio, los frutos fueron limpiados y lavados con agua desionizada. El jugo se extrajo con un extractor doméstico, de plástico, y fue colado para eliminar las semillas. Todas las muestras se liofilizaron durante un mínimo de 48 h a una presión de cámara de 0,05 mbar, se homogeneizaron y se almacenaron en bolsas de cremallera de polietileno, etiquetadas.

Todos los productos químicos utilizados fueron de la más alta pureza disponible y los materiales de vidrio fueron lavados con ácido nítrico (HNO_3) ultrapuro grado 65% (m/m) (Sigma) y enjuagados con agua ultrapura (se utilizó exclusivamente agua desionizada ultra pura con una resistividad de 18,1 M cm^{-1}). Se trabajó con soluciones estándar mono y multielemento de grado de análisis de trazas (Sigma-Aldrich y Agilent).

En las muestras de naranja y mandarina, el pretratamiento y posterior acondicionamiento se realizó en el Instituto de Química Básica y Aplicada del Nordeste (IQUIBA-NEA). Los procedimientos analíticos para la determinación de la composición multielemental de los jugos de naranja se realizaron, mediante espectroscopía de

absorción atómica por llama en la Facultad de Ingeniería de la UNNE y mediante espectroscopía de emisión atómica de plasma de microondas en el IQUIBA-NEA.

Una vez arribados al laboratorio, los frutos fueron procesados a la brevedad posible, generalmente en un plazo no mayor a 48 horas. Se lavaron con agua destilada y cepillo de cerdas blandas, para eliminar cualquier tipo de contaminación por tierra u otro residuo. Luego se cortaron utilizando un cuchillo plástico y se extrajo el jugo interior usando un exprimidor de frutas convencional tipo doméstico de plástico. Los jugos obtenidos se colocaron en bolsas de polietileno termocontraíble (250 ml) y a continuación se cerraron las mismas usando una termoselladora de polietileno tipo guillotina. Los envases de jugo cerrados se congelaron en freezer convencional (-18ºC) hasta su análisis.

3.4. Análisis multielemental

3.4.1. Digestión de muestras

Las muestras fueron pretratadas para realizar la determinación multielemental siguiendo distintos protocolos según el instrumento analítico de medida debido a los distintos requerimientos de cada técnica.

3.4.1.1. Pretratamiento de muestras para espectroscopía atómica de masas

Para analizar las muestras por espectroscopía de masas acoplado a plasma inductivo (ICP-MS) resulta necesario colocar las muestras en una forma soluble, libre de materia orgánica, para lograrlo se realiza un pretratamiento denominado *digestión de muestras*. La digestión de muestras para ICP-MS se realizó en recipiente cerrado asistida por microondas. El pretratamiento de las muestras para medición por ICP-MS

se realizó en el Instituto de Química San Luis (INQUISAL) dependiente de la UNSL y CONICET.

Se utilizó un horno de digestión de microondas de alto rendimiento (marca Milestone, modelo Ethos One), capaz de controlar presión y temperatura de digestión individualmente en cada bomba de digestión, provisto con carrusel para digerir 8 (ocho) muestras por cada tanda.

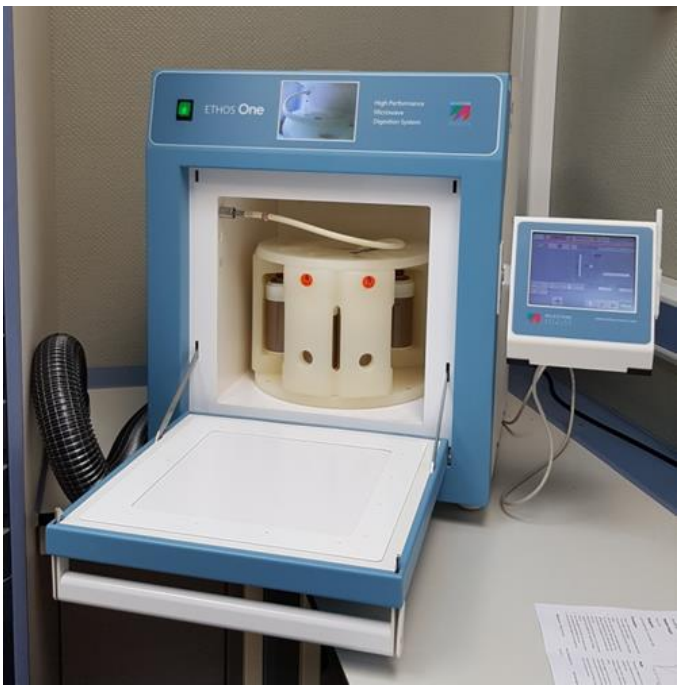

Figura 3.4. Horno de digestión de microondas de alto rendimiento (Milestone, modelo Ethos One)

Para la digestión se pesó en balanza analítica 0,50 g de muestra seca. Por tratarse de jugos, los mismos fueron secados en baño de vapor previamente a la digestión. A continuación, se colocó la muestra en una bomba de digestión de teflón, con cierre de seguridad para microondas. Se añadió a cada muestra 2,0 ml de H_2O_2 30% (m/m) y 6,0 ml de HNO_3 65% (m/m) purificado, y se mantuvo en reposo durante 10 minutos a temperatura ambiente para permitir que se estabilice la mezcla. El programa de digestión de microondas aplicado incluyó las siguientes etapas de temperatura: (1) 25-200°C durante 15 min, (2) 200°C durante 15 min y (3) 200-110°C durante 15 min, seguido inmediatamente por ventilación a temperatura ambiente en campana (20 min). Finalmente, los digestos se diluyeron a volumen final de 10 mL

utilizando una solución de HNO_3 (Altundag & Tuzen, 2011). Las soluciones blanco fueron preparadas de la misma manera que la muestras.

3.4.1.2. Pretratamiento de muestras para espectroscopía óptica de absorción y emisión atómica

Los requisitos de las muestras para poder ser analizadas por espectroscopía óptica son similares a los descritos en la sección anterior, sin embargo, dado que estas técnicas poseen rangos de sensibilidad menores a las espectroscopías de masa en esta tesis se utilizaron métodos clásicos de digestión de muestras. El pretratamiento de las muestras para medición por técnicas de espectroscopía atómica óptica se realizó en el Instituto de Química Básica y Aplicada del Nordeste (IQUIBA-NEA) dependiente de la UNNE y CONICET.

Para la digestión de muestras, se colocan 50 mL de muestra descongelada en capsulas de porcelana (Figura 3.5 a), en las que se realizó la digestión por vía seca. En primer término, se secaron las muestras en estufa a 65°C durante 24-48 horas, una vez eliminada la presencia de líquidos en la muestra de jugos, se completó el secado a 110°C en estufa durante aproximadamente 12 horas, hasta obtener constancia de peso (Figura 3.5 b). Las muestras secas se colocaron en horno mufla 550-600°C (Figura 3.5 c) durante aproximadamente 4 horas (hasta obtención de cenizas blancas, Figura 3.5 d), realizando una rampa escalonada de aumento de la temperatura para evitar la combustión. Finalmente, las cenizas obtenidas se disolvieron con 5 mL de ácido nítrico para análisis (65%), y se llevaron a volumen final de 30 mL con agua desionizada (Figura 3.5 e).

Figura 3.5. Pretratamiento de las muestras de jugos cítricos

3.4.2. Determinación multielemental

Según la disponibilidad se utilizaron diferentes técnicas analíticas para la determinación multielemental de los jugos cítricos bajo estudio. Dichas técnicas se describen a continuación:

3.4.2.1. Espectroscopía atómica de masas con plasma acoplado inductivamente

En el caso de los jugos obtenidos a partir de frutos del limonero, las determinaciones de oligoelementos en muestras digeridas se llevaron a cabo mediante espectroscopía de masas con plasma acoplado inductivamente (ICP-MS). Esta es una técnica analítica que combina las propiedades analíticas de dos metodologías de alto desempeño, por un lado, la elevada sensibilidad de la espectroscopía de masas, y por el otro la gran capacidad del plasma para generar iones atómicos debido a sus elevadas temperaturas (6000 a 8000 ºK). Esta es una técnica altamente sensible y capaz de determinar de forma cuantitativa casi todos los elementos presentes en la tabla periódica que tengan un potencial de ionización menor que el potencial de ionización del argón (gas utilizado como transportador – *Carrier*) a concentraciones muy bajas (ng/L o ppt) (Aceto, 2016).

Entre sus principales ventajas se pueden nombrar:

- Bajos límites de detección (del orden de las ppt).
- Capacidad de determinación simultánea en lapsos de tiempo reducidos.
- Capacidad de determinar la mayoría de los elementos de aparición frecuente en muestras naturales.
- Capacidad de medir isótopos individuales de cada elemento, posibilitando realizar mediciones de abundancia isotópica.

Sin embargo, esta técnica a pesar de sus grandes virtudes, no se encuentra exenta de presentar algunas limitaciones que se resumen a continuación:

- Presencia de interferencias espectrales derivadas de la matriz, se modifican en función de la composición de las muestras. Generalmente son variables y desconocidas.

- Rango lineal limitado, especialmente cuando se trabaja con niveles disímiles de concentración (configuración de alta sensibilidad para elementos ultra traza, podrían saturar la señal de mayoritarios).

- Tolerancia del plasma: supresión de la ionización o problemas en la interfase (bloqueo de los conos y deriva de la señal, requieren revisión y mantenimiento constante) debido a muestras con alto contenido de matriz (sólidos disueltos > 0.3%).

La primera etapa en un ICP-MS consiste en obtener los iones atómicos a partir del aerosol de la muestra líquida, obtenida mediante el uso de nebulizadores, donde se mezcla la muestra con el gas transportador (argón) para obtener una niebla de finas partículas de muestra líquida. A continuación, ocurren los siguientes procesos sucesivos al introducir la muestra al plasma (Figura 3.6): *Desolvatación:* ocurre cuando se evapora el disolvente hasta producir un aerosol molecular sólido finamente dividido. *Atomización:* la mayoría de las moléculas se disocia para formar un gas atómico. *Ionización:* los átomos generan iones y electrones libres por el elevado aporte de energía del plasma.

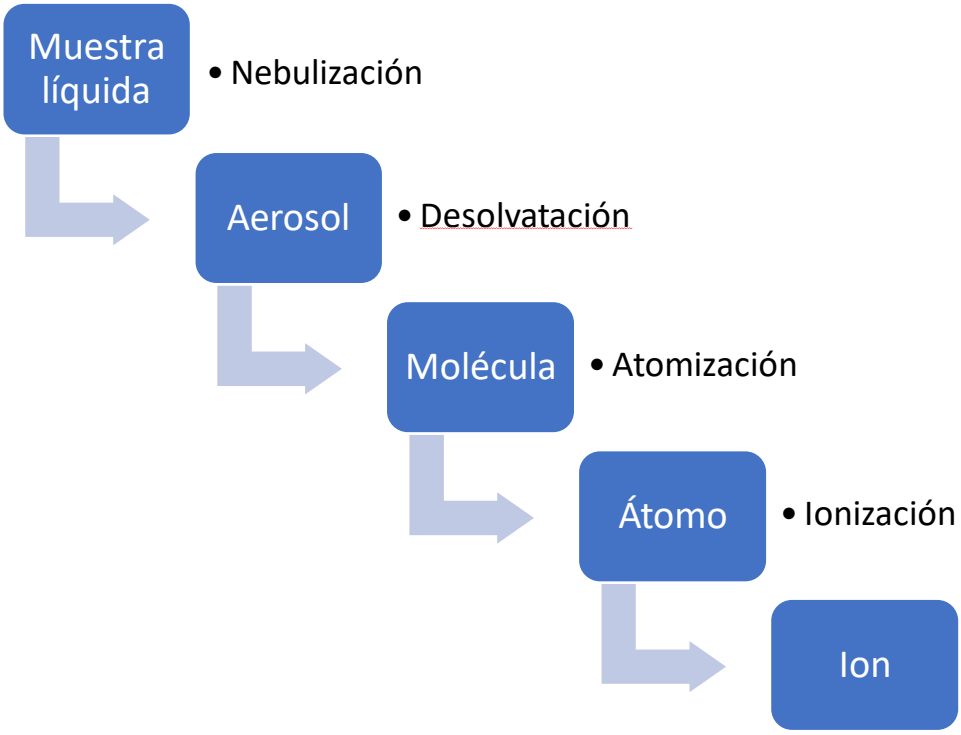

Figura 3.6. Etapas de la ionización de la muestra líquida en plasma.

El equipo utilizado en esta tesis fue espectrómetro ICP-MS de marca Agilent modelo 7700 disponible en la UNSL. Dicho instrumento está equipado con un nebulizador concéntrico de vidrio MicroMist® combinado con una cámara de nebulización de cuarzo de doble paso. Para suprimir las interferencias espectrales poliatómicas presentes en el plasma, el haz iónico se hizo atravesar por una celda de colisión de helio localizada antes de la entrada al cuadrupolo analizador. Se inyectó en la celda de colisión gas helio (5 ml/min) a presión constante. Las etapas a continuación del plasma del ICP-MS se describen a seguir (Figura 3.7): *Interfase:* este dispositivo permite ajustar la presión atmosférica del plasma al alto vacío. *Sistema de vacío:* proporciona el alto vacío para la óptica iónica, el cuadrupolo y el detector. *Lentes iónicos:* enfocan el haz iónico para su correcta introducción dentro del cuadrupolo. *Cuadrupolo:* actúa como un filtro de masa para ordenar los iones de acuerdo con su relación de masa-carga (m/z). *Detector:* contabiliza los iones individuales separados en la etapa anterior. *Sistema de control y manejo de datos:* controla todos los parámetros del instrumento y el manejo de obtención de datos.

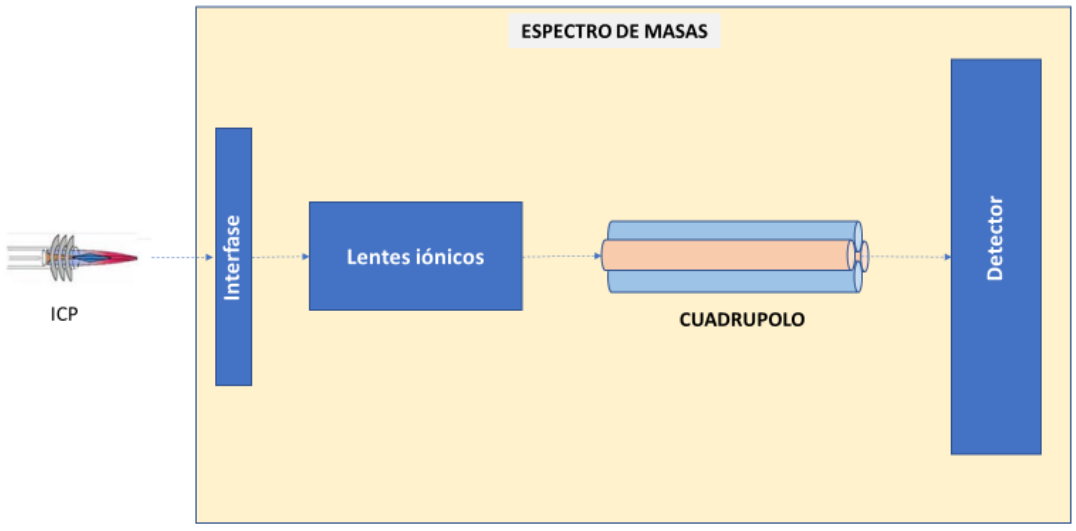

Figura 3.7. Esquema de un espectrómetro tipo ICP-MS.

En esta tesis todos los parámetros instrumentales fueron ajustados utilizando solución estándar multielemental en ácido nítrico diluido para ICP-MS número VI (30 elementos) marca Merck. Los isótopos seleccionados para la medición fueron Ag^{107},

Al^{27}, As^{75}, Ba^{137}, Bi^{209}, Co^{59}, Cr^{53}, Cu^{63}, Fe^{56}, Ga^{71}, La^{139}, Li^{7}, Mn^{55}, Mo^{95}, Ni^{60}, Rb^{85}, Sb^{121}, Sc^{45}, Se^{78}, Sn^{118}, Sr^{88}, Tl^{205}, V^{51} y Zn^{66}. Las muestras digeridas, soluciones estándar y los blancos correspondientes, se midieron mediante nebulización directa. La selección de isótopos para cada elemento objetivo se llevó a cabo comprobando la ausencia de interferencias poliatómicas, isobáricas y físicas. Las curvas de calibración se obtuvieron a cinco niveles de concentración diferentes en triplicado. Los valores de los coeficientes de determinación (R^2) para comprobación del ajuste lineal oscilaron entre 0,9980 y 0,9997. Para corregir el efecto matriz de las muestras digeridas, se añadió a cada muestra una solución estándar de 100 µg/L Rh^{103} y Y^{89} (Agilent), como patrón interno.

Para la validación del método de determinación multielemental en muestras digeridas de jugo de limón, se utilizó un material vegetal certificado, SRM 1570a (NIST-hojas de espinaca) digerido de la misma manera que las muestras. Los porcentajes de recuperación variaron entre 99 y 107% para los elementos Al, As, Co, Cu, Mn, Ni, Se, Sr, V y Zn. Debido a que no todos los elementos estudiados se consideraron en el material certificado, se llevaron a cabo estudios de recuperación en muestras con sobreagregado de estándar en jugo de limón, seleccionadas al azar, previo a su digestión, a niveles de 10 y 100 µg/kg. Las soluciones se analizaron en condiciones instrumentales optimizadas por triplicado. Los valores de recuperación se encontraron en el rango de 98 y 104%. Las recuperaciones obtenidas confirmaron que no se produjeron pérdidas significativas de elementos durante el procedimiento de digestión

3.4.2.2. Espectroscopía de absorción atómica por llama

Las concentraciones de los elementos K, Ca, Fe, Mg, Zn y Mn, cuya concentración se puede anticipar, se encuentran en niveles de los mg por litro, fueron medidos por FAAS, en función de la disponibilidad de dicho equipamiento. Las mediciones de muestras digeridas se realizaron en el Laboratorio de Química de la Facultad de Ingeniería de la Universidad Nacional del Nordeste.

Una vez procesadas las muestras de jugo (ver sección 3.2.1.2), se analizaron por FAAS. Esta técnica se fundamenta en el hecho de que al suministrar una determinada cantidad de energía a un átomo cualquiera en estado fundamental (E_o), ésta es absorbida por el átomo de tal forma que se incrementará el radio de giro de sus electrones de la capa externa llevando al átomo a un nuevo estado energético (E_1) que llamamos excitado. Cuando éste vuelve a su estado fundamental, cede una cantidad de energía cuantitativamente idéntica a su energía de excitación, emitiendo radiaciones a longitudes de onda determinadas. Cuando los átomos en estado fundamental se encuentran con las mismas radiaciones que ellos mismos son capaces de emitir, se produce una absorción de estas, desplazándose el equilibrio del estado fundamental al excitado. El fenómeno de absorción de radiaciones, a determinadas longitudes de onda, en el caso particular en que el medio absorbente sean los átomos en estado fundamental se conoce como Espectroscopía de Absorción Atómica (EAA). En EAA se emplean lámparas específicas dependiendo del elemento que se va a determinar. Estas son capaces de emitir una línea atómica característica, son las denominadas lámparas de cátodo hueco (Akman *et al.*, 2007).

Figura 3.8. Espectrómetro de absorción atómica marca Agilent 240 AA (fuente: agilent.com)

En esta tesis se utilizó un espectrómetro de absorción atómica marca Agilent® modelo 240 AA (Figura 3.8). Este sistema posee como principal característica el hecho

de poseer el sistema óptico completamente sellado, siendo apto para ser usado en ambientes polvorientos o corrosivos. Posee espejos cubiertos de cuarzo y sistema de purga de aire para limpiar continuamente el interior del equipo eliminando vapores corrosivos. Es un equipo de fácil mantenimiento de la lámpara: la lámpara es accesible de inmediato para su fácil ajuste o sustitución sin necesidad de retirar las cubiertas de los instrumentos.

Esta es una técnica de análisis instrumental, que, si bien posee capacidades de determinación inferiores a las técnicas de espectroscopía de masas, brinda resultados robustos y es generalmente la técnica de elección para la determinación de elementos mayoritarios y minoritarios en muestras agrícolas, en laboratorios de mediana complejidad. Entre sus principales ventajas se puede destacar (Caroli, 1992):

- Rangos de concentración más elevados, por lo que resultan aptas para la determinación de elementos mayoritarios.
- Bajo costo de instalación relativo a las técnicas de emisión atómica.
- Amplia disponibilidad en laboratorios de mediana complejidad del medio.

Presenta también una serie de limitaciones con respecto a las técnicas de emisión atómica. Entre las principales se puede nombrar:

- Menores temperaturas de excitación de la materia, siendo una técnica apta para la cuantificación de un número limitado de elementos.
- No permiten la determinación simultánea de más de un elemento por vez.
- Requieren disponer de lámparas específicas para cada elemento que se quiere determinar.
- Utiliza gases combustibles y oxidantes que requieren cuidados especiales para su manejo.

- La sensibilidad de esta técnica es menor varios órdenes de magnitud con respecto a las técnicas de emisión atómica (del orden de los mg/L - ppm).

Una vez que se obtuvieron las muestras listas para ser analizadas en el espectrofotómetro de absorción atómica con llama, se realizó una curva de calibración para cada elemento utilizando cinco soluciones estándar en concentraciones crecientes. Las curvas obtenidas se evaluaron determinando sus parámetros estadísticos característicos (R^2 > 0,998). Los resultados experimentales obtenidos se evaluaron mediante el método de sobreagregado de estándar (Pohl *et al.*, 2018, Mohd Fairulnizal *et al.*, 2019).

3.4.2.3. Espectroscopía óptica de emisión atómica por plasma de microondas

Adicionalmente, las muestras de jugos de naranja y mandarina fueron analizadas también por espectroscopía de emisión atómica por plasma de microondas (MP-AES). Estos análisis en base a la disponibilidad local se realizaron en el Instituto de Química Básica y Aplicada del Nordeste Argentino (IQUIBA-NEA) dependiente de la UNNE y el CONICET.

La espectroscopía de emisión atómica por plasma de microondas (MP-AES), consiste en un plasma inducido por microondas conectado a un espectrofotómetro de emisión atómica (AES). Se utiliza para la determinación simultánea de múltiples analitos de elementos mayoritarios y minoritarios. La MP-AES emplea energía de microondas para producir una descarga de plasma utilizando nitrógeno suministrado desde un generador de nitrógeno que lo extrae del aire ambiente, lo que elimina la necesidad de abastecimiento de gases combustibles de difícil manejo (Figura 3.9). El plasma de nitrógeno es considerablemente más caliente (hasta 5.000 ºK) que la llama de aire producida por acetileno utilizada en FAAS (Balaram, 2020).

Para la determinación, las muestras son nebulizadas antes de la interacción con el plasma. La muestra atomizada pasa a través del plasma y los electrones son

promovidos al estado excitado. Los electrones promovidos emiten luz al regresar al estado fundamental, que se separa en un espectro y la intensidad de cada línea de emisión se mide en el detector. Los elementos más comúnmente determinados se pueden medir con una sensibilidad de hasta partes por millón (ppm) (Thirumdas, *et al.* 2019). Las ventajas potenciales del MP-AES incluyen un menor costo de operación que otras técnicas espectroscópicas y la eliminación del requerimiento de gases inflamables. En esta tesis se utilizó un espectrómetro MP-AES marca Agilent 4200 (Figura 3.10).

Figura 3.9. Generador de nitrógeno marca Peak (Genius 5000)

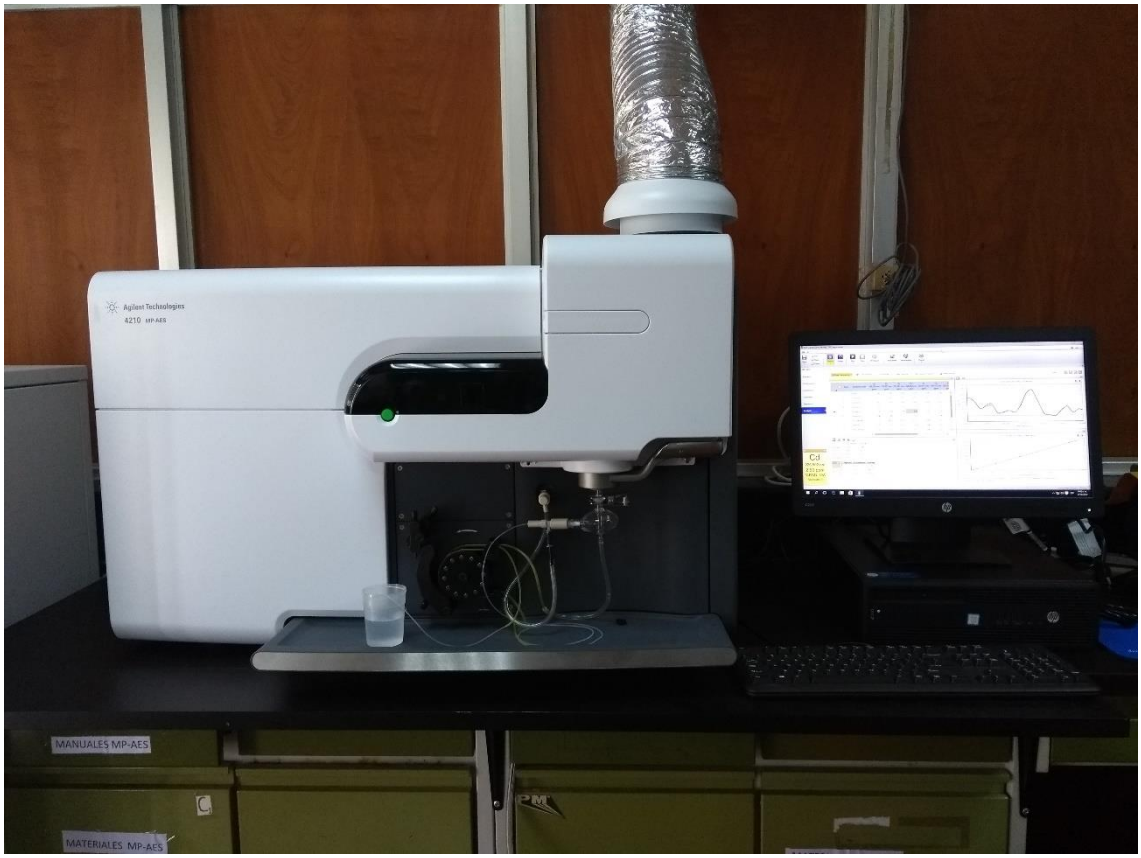

Figura 3.10. Espectrómetro MP-AES marca Agilent 4200.

En muestras digeridas se determinó la concentración de 20 elementos, Al, As, Ba, Cd, Co, Cu, Cr, Fe, K, Mg, Mn, Mo, Ni, Pb, Sb, Se, Sr, Tl y Zn. Se confeccionaron rectas de calibración para cada elemento a partir de patrones triplicados y empleando 5 niveles de concentración. Las diluciones se prepararon a partir de soluciones estándar certificados para espectrometría atómica (TraceCERT®). Se utilizaron patrones monoelementales de 100 mg/L de Al, As, Ba, Mg, Mo, Pb, Ti, Se, Sr, Sb y Tl; además de un patrón multielemental conteniendo 100 mg/L de Cr, Mn, Fe, Co, V, Ni, Cu, Zn, Ag y Cd. Los coeficientes de regresión obtenidos para cada recta de calibración tuvieron un R^2 ajustado comprendido entre 0,9986 a 0,9998. Estas rectas de calibración fueron luego utilizadas para la cuantificación de estos elementos en muestras reales. Adicionalmente, se determinaron los LODs. Los LOD están definidos como 3 veces la desviación estándar (σ) de la medida de 11 blancos de procedimiento. Estos blancos han sido preparados y analizados en diferentes días y de manera no consecutiva.

Por otro lado, se realizaron lecturas replicadas de una solución estándar conteniendo todos los elementos dentro de un mismo día de trabajo y en distintos días

de trabajo, con nuevas condiciones de sintonización del equipo. Estas mediciones se realizaron con el objeto de evaluar la precisión del método, teniendo en cuenta la repetibilidad y reproducibilidad de este. Valores inferiores a 3,62% de SD se obtuvieron para las mediciones realizadas en un mismo día (repetibilidad) y valores inferiores a 7,25% se alcanzaron para mediciones replicadas en distintos días (reproducibilidad), por lo que se puede afirmar que la precisión de la metodología propuesta es adecuada.

Las muestras fueron pretratadas para realizar la determinación multielemental siguiendo distintos protocolos según el instrumento analítico de medida.

3.5. Datos

El conjunto de datos con el que se trabajó estuvo formado por variables de clasificación que identificaron los sitios de origen de las muestras y las variables clasificadoras que corresponden a los contenidos de elementos minerales en jugos de frutos cítricos, determinados con diferentes instrumentos de análisis químico.

En el caso de los jugos de limones (74 muestras), las variables estudiadas fueron: Ag, Al, As, Ba, Bi, Co, Cr, Cu, Fe, Ga, La, Li, Mn, Mo, Ni, Rb, Sb, Sc, Se, Sn, Sr, Tl, V y Zn determinados mediante ICP-MS. En los jugos de mandarinas (200 muestras) y naranjas (200 muestras) se determinaron las concentraciones de: Al, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Mo, Ni, Pb, Sb, Se, Sr y Zn determinados mediante MP-AES. Para los jugos de naranjas se determinaron, además (120 muestras): K, Ca, Fe, Mg, Zn y Mn mediante FAAS.

3.6. Análisis de datos

En primer lugar, se realizó un Análisis Exploratorio de Datos empleando herramientas gráficas y técnicas de Estadística Descriptiva, Análisis de Varianza

(ANOVA) para cada uno de los elementos, con las correspondientes pruebas F combinadas con la prueba de Duncan para separación de promedios (Perelman *et al*., 2019).

Ante la necesidad de estudio simultáneo de varias variables, se utilizaron Métodos de Análisis Estadísticos Multivariantes y de Aprendizaje Automático, para explorar y proponer modelos matemáticos que permitan clasificar las distintas muestras de acuerdo con criterios preestablecidos y/o establecer criterios de identidad de muestras de acuerdo con su composición química.

La caracterización multivariada básica de las muestras de jugo de limón investigadas se realizó mediante un Análisis de Componentes Principales (PCA). Esta técnica representa la agrupación natural de las muestras estudiadas, así como las variables en un espacio multidimensional. Además, el PCA reduce el número de variables utilizadas para describir datos (Bro & Smilde, 2014). Posteriormente al PCA, se investigaron cinco técnicas de clasificación: dos derivadas del Análisis Discriminante (DA), que son el Análisis Discriminante Lineal (LDA) y el Análisis Discriminante Mínimo Parcial (PLS-DA), el Vecino Más Cercano (KNN), los Bosques Aleatorios (RF) y la Máquina de Vectores de Soporte (SVM), a fin de lograr la identificación de la procedencia del jugo de limón (James *et al.*, 2017).

Para la clasificación de muestras de jugos de mandarinas y naranjas, dentro de los Métodos Multivariantes se aplicaron técnicas de Análisis de Componentes Principales (PCA), Análisis de la Varianza Multivariado con posterior Prueba de Hottelling y Análisis Discriminante Lineal (LDA) (Peña, 2002). Entre las técnicas de Aprendizaje Automático se utilizaron Árboles de Decisión (DT), con los algoritmos C5.0 y CART; métodos vagos de K-Vecino más Cercano (KNN), buscando la mejor configuración; Redes Neuronales Artificiales (ANN), con datos normalizados, empleando el algoritmo MLP (perceptrón multicapa) y probando diferente cantidad de capas y número de neuronas por capa, seleccionando la cantidad de iteraciones en que se estabiliza el porcentaje de acierto; los Bosques Aleatorios (RF); y Máquinas de Vectores Soporte (SVM) con diferentes funciones *kernel* (James et al., 2017).

Para DT, KNN y ANN se probaron los métodos de *bootsting, live one out*, *live group out* y *cross validation*, de los que finalmente se seleccionó por sus mejores resultados el de *cross validation* con 10 cajas, se utilizó la función *train* del paquete Caret de R para construir un modelo y con esa función se probaron diferentes valores para los distintos parámetros.

El PCA se empleó para realizar una caracterización multivariada básica de las muestras de jugo de naranja dado que esta técnica permite reducir la dimensión del espacio vectorial y representar la agrupación natural de las observaciones, así como las variables, en un espacio de dimensión reducida con la menor pérdida de información (Bro & Smilde, 2014).

El LDA se efectuó con el propósito de discriminar entre clases, maximizando la variancia entre clases y minimizando la varianza dentro de cada clase. Este análisis define funciones canónicas o discriminantes, que son combinaciones lineales de las variables originales que optimizan esa separación. El resultado de la predicción, para el conjunto de prueba, se obtiene mediante la proyección de las observaciones de acuerdo con la distancia mínima al centroide de cada clase (Moncayo *et al.*, 2015).

El PLS-DA es un método de clasificación lineal basado en un algoritmo de regresión parcial de mínimos cuadrados (PLS), para construir modelos predictivos cuando los factores (variables independientes) son muchos y altamente colineales. El algoritmo PLS busca variables latentes con la máxima covarianza con las variables dependientes que representan la pertenencia a clases. Como tal, PLS tiene en cuenta la variable dependiente al definir variables latentes. Esa técnica se convirtió en una herramienta establecida en el modelado quimiométrico, ya que a menudo es posible interpretar el factor extraído en términos del sistema físico subyacente. En general, el método PLS-DA a menudo se considera que funciona bien en la práctica (Ballabio & Consonni, 2013).

Los DT se utilizaron por ser herramientas de clasificación poderosas y flexibles, en las que la regla de clasificación final tiene una forma simple, fácil de interpretar y usar en futuras clasificaciones. Toman en cuenta el hecho de que pueden ser

construidas diferentes relaciones entre variables en distintas regiones de los datos. Realizan una selección automática paso a paso de variables y calculan el rango de importancia de estas (Huang *et al*., 2014). Se utilizó la función *train* del paquete Caret de R para construir un modelo por cada algoritmo y con esa función se probaron diferentes valores para los distintos parámetros. Para el algoritmo C5.0 se optimizó el tipo de salida (árbol o regla), el filtrado de variables (si o no) y el número de experimentos; para el algoritmo CART para optimizó el parámetro *cp* que establece la necesidad de introducir o no un nuevo nodo (Rpart y Rpart 2).

El KNN es una técnica discriminante no lineal, que se centra en las distancias entre los objetos y, en particular, en los objetos más cercanos. Este método se utilizó para clasificar muestras desconocidas por proyección en el espacio multivariante y asignación a la clase de su vecino más cercano en el conjunto de entrenamiento. El parámetro *k* se debe optimizar y representa el número de vecinos que se tiene en cuenta para decidir por mayoría de votos la clase de muestras desconocidas (Huang *et al*., 2014). Se trabajó con el paradigma de los métodos vagos, en particular el del vecino más cercano variando el valor del parámetro *k*.

Las ANN son definidas como estructuras que comprenden elementos de procesamiento simples, adaptativos, densamente interconectados, denominados neuronas artificiales (nodos), que pueden realizar cálculos paralelos masivamente para el procesamiento de datos y la representación del conocimiento (Berrueta *et al*., 2007). Se realizó un preprocesamiento de los datos en el que se centraron y escalaron los mismos, luego se probaron los métodos ya mencionados. En el caso de las ANN el paquete Caret permite configurar el número de capas ocultas y el número de neuronas por capa oculta, construido con el método MLP mediante el parámetro *size*.

Los RF son un método de aprendizaje de conjunto que combina una agregación de arranque (bagging) para formar un conjunto de submuestras y predictores de decisión de árbol para la clasificación. Este método utiliza el promedio en la respuesta para mejorar el porcentaje de acierto predictiva y el control de sobreajuste (Kuhn & Johnson, 2013). Una de las principales ventajas de este método es que generalmente alcanza altos niveles de acierto, usualmente mucho más alto que el obtenido con un

solo árbol de decisión (Batista *et al*., 2012). Además, este método requiere un bajo costo computacional para conjuntos de datos grandes y proporciona estimaciones de las variables más importantes en la clasificación.

Las SVM constituyen un método de máquina de aprendizaje que crea una asignación de los datos de entrenamiento en un espacio de alta dimensión, calculan un hiperplano de separación óptimo mediante un algoritmo iterativo que aprende la distribución de la muestra en los límites de cada clase considerada (Li *et al*., 2009). La complejidad del modelo se controla mediante una función de error de penalización para evitar un ajuste excesivo (Zeng y Lu, 2012). Cuando los grupos no son linealmente separables, una forma de abordar ese problema es proyectar los ejemplos en un espacio de dimensión superior (o eventualmente inferior) donde el conjunto sea linealmente separable y resolver el problema en ese espacio; para realizar esta transformación se necesita una función llamada *kernel* que transforma los ejemplos de un espacio a otro, añadiendo dimensiones. Para Las SVM dotadas de *kernel* son clasificadores muy poderosos con las ventajas de que no son sensibles al ruido ni proclives al sobreajuste, pero también tienen un puntos débiles, fundamentalmente que son como cajas negras, es decir no se puede conocer internamente su proceso y no se sabe a priori cual es el *kernel* más adecuado. Se probaron funciones *kernel* polinomial y radial, no lineales y de uso más común, asignando diferentes valores a los parámetros correspondientes a cada modelo.

Cuando se trabajó con métodos supervisados, a fin de evaluar los resultados de los métodos de clasificación, para conocer su rendimiento, se dividió el conjunto de datos en dos subconjuntos disjuntos, un subconjunto de entrenamiento y uno de prueba. El subconjunto de entrenamiento se utilizó para la definición de los modelos y la optimización de los parámetros, luego, estos modelos se probaron sobre el subconjunto de prueba, con datos no utilizados en la definición de los modelos y se aplican los conceptos de optimización de la complejidad de los modelos, utilizando la técnica de validación cruzada. La matriz de datos se dividió aleatoriamente en un conjunto de entrenamiento, con miembros de clase conocida utilizados exclusivamente para optimizar los parámetros que son necesarios para cada método; y

el conjunto de prueba que contenía los objetos restantes no incluidos en el entrenamiento, utilizado exclusivamente para evaluar el desempeño de cada modelo contra un conjunto de muestras desconocidas para el modelo. En la validación cruzada, el conjunto de datos se dividió aleatoriamente en k subconjuntos mutuamente excluyentes de aproximadamente el mismo tamaño. El clasificador fue entrenado y probado k veces, cada vez. La validación cruzada se repitió *n* veces para cada modelo de clasificación y la estimación final de precisión es la media de todas las estimaciones calculadas.

En el caso de jugos limón, el subconjunto de entrenamiento estuvo constituido por el 70% y el de prueba por el 30% restante, para los jugos de mandarinas y naranjas se contaba con mayor cantidad de datos, por lo que se empleó el 60% de los datos en el subconjunto de entrenamiento y el 40% restante en el subconjunto de prueba.

Para la comparación de los métodos y la selección final del modelo se emplearon los criterios de sensibilidad, especificidad, porcentaje de acierto e índice κ.

Los análisis fueron realizados con los Software Infostat (Di Rienzo *et al.*, 2020) y el Software libre R 3.2.1. (R Core Team, 2020), con la metodología descripta por Kuhn (2008).

3.7. Referencias

Aceto, M. 2016. 8 - The Use of ICP-MS in Food Traceability. Woodhead Publishing Series in Food Science, Technology and Nutrition, Advances in Food Traceability Techniques and Technologies. 137-164 pp.

Akman, S; Demirata-Ozturk, B; Tokman, N. 2007. Chapter 17 - Atomic Absorption Spectroscopy, (Eds): Picó, Y. Food Toxicants Analysis, Elsevier. 637-665 pp.

Altundag, H; Tuzen, M. 2011. Comparison of dry, wet and microwave digestion methods for the multi element determination in some dried fruit samples by ICP-OES.

Food and chemical toxicology: an international journal published for the British Industrial Biological Research Association. 49: 2800-2807.

Apodaca, M; Crisci, J; Katinas, L. 2015. Las provincias fitogeográficas de la República Argentina: definición y sus principales áreas protegidas. SN - 978-950-9149-39-7. 79-101.

Balaram, V. 2020. Microwave plasma atomic emission spectrometry (MP-AES) and its applications – A critical review, Microchemical Journal, 159: 105483. ISSN 0026-265X.

Ballabio, D; Consonni, V. 2013. Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods. 5 (16): 3790-3790.

Batista, BL; da Silva, LRS; Rocha, BA; Rodrigues, JL; Berretta Silva, AA; Bonates, TO; Gomes, VSD; Barbosa, RM; Barbosa, F. 2012. Multi-element determination in Brazilian honey samples by inductively coupledcoupled plasma mass spectrometry and estimation of geographic origin techniques. International Food Research. 49 (1): 209-215.

Berrueta, LA; Alonso Salces, RM; Hérberger, K. 2007. Supervised pattern recognition in food analysis. J. Chromatography A. 1158 (1-2): 196-214.

Bro, R; Smilde, AK. 2014. Principal Component Analysis, Analytical Methods. 6 (9): 2812-2831.

Cabrera, AL. 1976. Regiones fitogeografías argentinas. Enciclopedia Argentina de Agricultura y Jardinería. 85 pp.

Carnevali, R. 1994. Fitogeografía de la Provincia de Corrientes. Gobierno de la Provincia de Corrientes. 1ra edición. 324 pp.

Caroli, S. 1992. Atomic absorption spectrometry: a look into the future, (Eds): Minoia, C; Caroli, S. Applications of Zeeman Graphite Furnace Atomic Absorption Spectrometry in the Chemical Laboratory and in Toxicology, Pergamon. 647-667 pp. ISBN 9780080410197.

Código Alimentario Argentino. 2020. Ley 18284. Disponible en línea: https://www.argentina.gob.ar/anmat/codigoalimentario. Visita 12/02/2020.

CSC. Chinese Society of Citriculture. 2008. Citrus varieties in China. China agriculture press. ISBN: 9787109126053. 203 pp.

Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W. InfoStat versión 2020. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.infostat.com.ar.

Escobar, E; Ligier, H; Melgar, R; Matteio, H; Vallejos, O. 1996. Mapa de suelos de provincia de Corrientes 1:500.000 INTA. Centro Regional Corrientes.

Federcitrus. 2018. La actividad citrícola Argentina. Disponible en línea: https://www.federcitrus.org/. Visita: 04/09/2019.

Gunther, D; Correa de Temchuk, M; Lysiak, E. 2008. Zonas agroeconómicas homogéneas y sistemas de producción predominantes de la provincia de Misiones. Boletín Técnico INTA. 87 pp.

Huang, X; Teye, E; Owusu-Sekyere, JD; Takrama, J; Sam-Amoah, LK; Yao, L; Firempong, CK. 2014. Simultaneous Measurement of Titratable Acidity and Fermentation Index in Cocoa Beans by Electronic Tongue Together with Linear and Non-linear Multivariate Technique. Food Analytical Methods. 7 (10): 2137-2144.

Kuhn, M. 2008. Caret package. Journal of Statistical Software. 28 (5): 1-26.

Kuhn, M; Johnson, K. 2013. Nonlinear Classification Models, in: M. Kuhn, K. Johnson (Eds.) Applied Predictive Modeling, Springer New York, New York, NY. 329-367 pp.

James, G; Witten, D; Hastie, T; Tibshirani, R. 2017. An Introduction to Statistical Learning with Applications in R. Springer Science+Business Media New York 8th printing. 425 pp.

Mohd Fairulnizal, MN; Vimala, B; Rathi, DN; Mohd Naeem, MN. 2019. 9 - Atomic absorption spectroscopy for food quality evaluation, (Eds): Zhong, J; Wang, X. In Woodhead Publishing Series in Food Science, Technology and Nutrition, Evaluation Technologies for Food Quality, Woodhead Publishing. 145-173 pp.

Moncayo, S; Manzoor, S; Navarro Villoslada, F; Cáceres, JO. 2015. Evaluation of supervised chemometric methods for sample classification by Laser Induced Breakdown Spectroscopy. Chemometrics and Intelligent Laboratory Systems. 146: 354-364.

Palacios, J. 2013. Citricultura. Talleres Gráficos Alfa Beta S.A. ISBN: 9789874383266. 518 pp.

Peña, D. 2002. Análisis de Datos Multivariantes. Madrid: Mc Graw Hills/ Interamericana de España. 540 pp.

Perelman, SB, Garibaldi, LA; Tognetti, PM. 2019. Experimentación y Modelos Estadísticos. Ed. Facultad de Agronomía. Universidad de Buenos Aires. 475 pp.

Pintus, PV; Mazal, YA; Woloszyn, J. 2010. Plan de competitividad, conglomerado citrícola de Misiones. Ministerio de Economía y Finanzas Públicas, Secretaría de Política Económica, Programa de Competitividad del Norte Grande.

Pohl, P; Jedryczko, D; Dzimitrowicz, A; Szymczycha-Madeja, A; Welna, M; Jamroz, P. 2018. Fruit Juices, Academic Press. (Eds): Rajauria, G; Tiwari, BK. 739-761 pp.

Puchulu, ME; Fernández, DS. 2014. Características y distribución espacial de los suelos de la provincia de Tucumán. En: Moyano, S.; Puchulu, M. E.; Fernández, D.; Aceñolaza, G.; Vides, M. E.; Nieva, S. (Eds.), Geología de Tucumán. Colegio de Graduados en Ciencias Geológicas de Tucumán.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Thirumdas, R; Janve, M; Siliveru, K; Kothakota, A. 2019. 10 - Determination of food quality using atomic emission spectroscopy, In Woodhead Publishing Series in Food Science, Technology and Nutrition, Evaluation Technologies for Food Quality, Woodhead Publishing. 175-192 pp.

Zheng, H; Lu, H. 2012. A least-squares support vector machine (LS-SVM) based on fractal analysis and CIELab parameters for the detection of browning degree on mango (*Mangifera indica* L.). Computers and Electronics in Agriculture. (83): 47-51.

CAPÍTULO IV

DETERMINACIÓN DEL ORIGEN GEOGRÁFICO DE JUGOS DE LIMÓN BASADO EN PERFILES DE OLIGOELEMENTOS

4.1. Introducción

En general, los frutos de limonero (*Citrus limon* L., Osbeck) se pueden comercializar en el mercado internacional como frutas frescas, jugos o aceites. Hoy en día, Argentina produce alrededor de 1.675.851 t al año, principalmente de las variedades 'Génova' y 'Eureka'. Tomando la producción mundial (alrededor de 8.815.000 t al año), la participación de Argentina es del 20 % aproximadamente, dado que la producción anual puede variar entre países debido a las variaciones climáticas (Federcitrus, 2018).

Los jugos naturales de limón son fuentes de nutrientes, antioxidantes como flavonoides, carotenoides y vitamina C, elementos esenciales, K, Cu, Fe, Mg y Zn y fibra dietética soluble e insoluble. Juntos, estos nutrientes promueven varios beneficios para la salud y proporcionan protección contra diversas enfermedades (Llorente *et al*., 2014; Matheyambath *et al*., 2016).

Los productores, comerciantes y consumidores están especialmente interesados en el correcto etiquetado del origen y la trazabilidad de los frutos y del zumo de limón. La determinación de la autenticidad del origen geográfico es una cuestión importante para la creciente industria alimentaria en el control de calidad y la inocuidad de los alimentos (Amenta *et al*., 2015).

En este contexto, el uso de técnicas analíticas multielementales se está generalizando cada vez más para determinar el origen geográfico de los alimentos (Kelly *et al*., 2005). En general, las técnicas útiles para obtener la huella dactilar

elemental de los alimentos son aquellas con capacidad de detección de múltiples elementos, como aquellas basadas en espectrometría de masas con plasma de acoplamiento inductivo (ICP-MS, o Inductively Coupled Plasm Mass Spectrometry) (Gonzálvez *et al*., 2009). El análisis quimiométrico de los datos de composición de elementos complejos obtenidos por estos métodos instrumentales proporciona una mejor interpretación y la posibilidad de adquirir información relevante sobre la autenticidad de estos alimentos (Forina *et al.*, 2009).

Con la finalidad de realizar un adiestramiento en el funcionamiento de las técnicas de aprendizaje automático, con información química de contenido de oligoelementos en muestras de jugo de limón de las variedades 'Eureka', 'Lisboa' y 'Génova', el objetivo de este capítulo fue obtener modelos de clasificación adecuados a efectos de analizar orígenes de jugo de limón de las regiones noroeste (NOA) y noreste de Argentina (NEA).

Se ha propuesto un método ICP-MS y se ha determinado el contenido de 24 oligoelementos (Ag, Al, As, Ba, Bi, Co, Cr, Cu, Fe, Ga, La, Li, Mn, Mo, Ni, Rb, Sb, Sc, Se, Sn, Sr, Tl, V y Zn) en jugo de limón obtenido a partir de frutos cultivados en Argentina y derivados de las dos regiones productoras de limón más importantes del país: NOA (provincias de Jujuy y Tucumán) y NEA (provincia de Corrientes).

4.2. Materiales y Métodos

4.2.1. Muestras y procedimiento analítico

El conjunto de datos con el que se trabajó estuvo formado por variables de clasificación que identificaron los sitios de origen de las muestras (Corrientes: CTE, Jujuy: JY, Tucumán sitio 1: TN-I y Tucumán sitio 2: TN-II) y las variables clasificadoras que corresponden a los contenidos de elementos minerales en jugos de limón, determinados con ICP-MS.

Mediante un muestreo al azar simple sobre un padrón de productores con características tecnológicas medias, se obtuvieron 74 muestras, cada una de ellas compuesta por el jugo de 10 frutos (unidad muestral). Se determinaron en cada muestra la composición multielemental, conformada por la concentración de 24 (veinticuatro) elementos a nivel de vestigios mediante ICP-MS.

4.2.2. Datos

El conjunto de datos con el que se trabajó estuvo formado por una variable que identificaba los sitios de origen de las muestras (TN-I, TN-II, JY y CTE) y 24 variables que corresponden a los contenidos en, jugos de limón, de los siguientes elementos minerales: Ag, Al, As, Ba, Bi, Co, Cr, Cu, Fe, Ga, La, Li, Mn, Mo, Ni, Rb, Sb, Sc, Se, Sn, Sr, Tl, V y Zn.

4.2.3. Análisis de datos

Para abordar el análisis de los datos se realizó en primer lugar un análisis exploratorio de datos mediante la determinación de parámetros estadísticos distribucionales de cada variable y análisis de componentes principales (PCA) que permitió obtener una primera aproximación en un espacio reducido de variables.

Con los resultados obtenidos en la primera etapa, se propuso la elaboración de un modelo predictivo de origen geográfico de las muestras, basado en su composición multielemental. Se ensayaron, optimizaron y compararon los desempeños de cinco técnicas supervisadas clasificatorias, adecuadas para el análisis del tipo de datos problema: Análisis Discriminante Lineal (LDA), Análisis Discriminante por Mínimos Cuadrados Parciales (PLS-DA), Método del Vecino más Cercano (KNN), Máquinas de Vectores Soporte (SVM) y Bosques Aleatorios (RF).

4.3. Resultados y discusión

4.3.1. Caracterización multielemental

Los resultados de las concentraciones totales de elementos a nivel de vestigios en jugos de limón se presentan en la Tabla 4.1. Los resultados se expresan como valores medios para 5 (cinco) réplicas acompañados de sus respectivas desviaciones estándar (SD). Las concentraciones de los elementos Ag, As, Ga, Sb y Tl, no se incluyeron en la citada tabla, por encontrarse por debajo de los límites de detección (LOD) en todas las muestras.

Respecto del orden de abundancia elemental, todas las muestras presentaron perfiles de concentración similares. Las concentraciones de oligoelementos (contenidos por encima de 1,0 μg/g) se presentaron en el siguiente orden de concentración decreciente: Fe > Zn > Rb > Cu > Sr > Mn > Ba > Al > Ni. Teniendo en cuenta la concentración media de todas las muestras, el oligoelemento más abundante fue Fe seguido de Zn y Rb, todos ellos se encontraron en contenidos medios superiores a 10 μg/g. Al, Ba, Cu, Mn, Ni y Sr presentaron contenidos que oscilaban entre 1 y 10 μg/g. Por otro lado, se observaron las mayores variaciones entre diferentes muestras para los siguientes elementos ultra-traza (concentraciones inferiores a 1,0 μg/g): La, Cr, Se, Li, Mo, Co, Sn, Sc, V y Bi.

Niveles similares de concentración para los elementos Co, Cr, Fe, La, Sc y Zn fueron reportados en jugos de limón de otras procedencias (USDA, 2020; Potortí *et al.*, 2018, Tufour *et al.*, 2011). En comparación con otros jugos de frutas cítricas, como naranjas o pomelos, los contenidos reportados de Fe y Zn fueron superiores (Barros *et al.*, 2012), Al, Co, Cu, Fe, Li, Mn, Ni, Sr, V y Zn estaban en niveles similares (Simpkins *et al.*, 2000; Szymczycha Madeja & Welna, 2013). En general, el contenido de Rb resultó más bajo, mientras que Ba, Mo y Cr se encontraron en niveles más elevados que los reportados para el jugo de naranja comercial (Szymczycha Madeja & Welna, 2013).

Varios factores pueden contribuir a variaciones en los niveles de oligoelementos en el jugo de limón según su origen geográfico. En primer lugar, la disponibilidad del elemento en el suelo para su extracción por la planta. Esta disponibilidad depende principalmente de la capacidad de intercambio catiónico de los suelos, que puede variar entre los tipos de suelo, el pH y la composición de la matriz mineral. Otros factores, como las prácticas agrícolas, las aplicaciones de fertilizantes, el riego artificial o la madurez de los frutos en la cosecha, pueden influir en las concentraciones de oligoelementos (Szymczycha Madeja *et al*., 2014).

Tabla 4.1. Concentraciones de elementos minerales en muestras de jugo de limón (μg/g). Promedios (Me) y desviaciones estándares (DE)

Elemento	Me	DE
Al	3,20	0,02
Ba	5,40	0,64
Bi	0,15	0,05
Co	0,025	0,15
Cr	0,35	0,05
Cu	9,40	1,22
Fe	15,60	1,18
La	0,40	0,02
Li	0,30	0,01
Mn	7,50	0,44
Mo	0,25	0,01
Ni	1,40	0,24
Rb	10,20	1,57
Sc	0,02	0,01
Se	0,35	0,05
Sn	0,20	0,02
Sr	8,70	0,84
V	0,15	0,02
Zn	11,60	0,23

4.3.2. Análisis de componentes principales

Como etapa preliminar, antes del modelado de clasificación, se utilizó un PCA para el análisis exploratorio de los datos. Se aplicó a la matriz de datos auto escalados, que permitió un estudio de la estructura de datos en un espacio con dimensionalidad reducida, manteniendo la máxima cantidad de información presente en los mismos. Se extrajeron cuatro componentes principales con valores propios superiores a uno explicando el 56,3% de la variación total (CP1: 25,3%, CP2: 14,9%, CP3: 9,2% y CP4:

6,9%). Estos valores bajos de varianza resumida por cada componente permiten anticipar la presencia de elevadas correlaciones entre las concentraciones elementales. Asimismo, el bajo porcentaje de varianza resumida por las primeras componentes no representa un serio inconveniente, dado que este análisis se realizó con el objeto de explorar las posibles correlaciones entre variables y posibles agrupamientos o similitudes entre muestras (Bro *et al*. 2014).

La Figura 4.1 muestra los resultados obtenidos en el espacio formado por las dos primeras componentes principales.

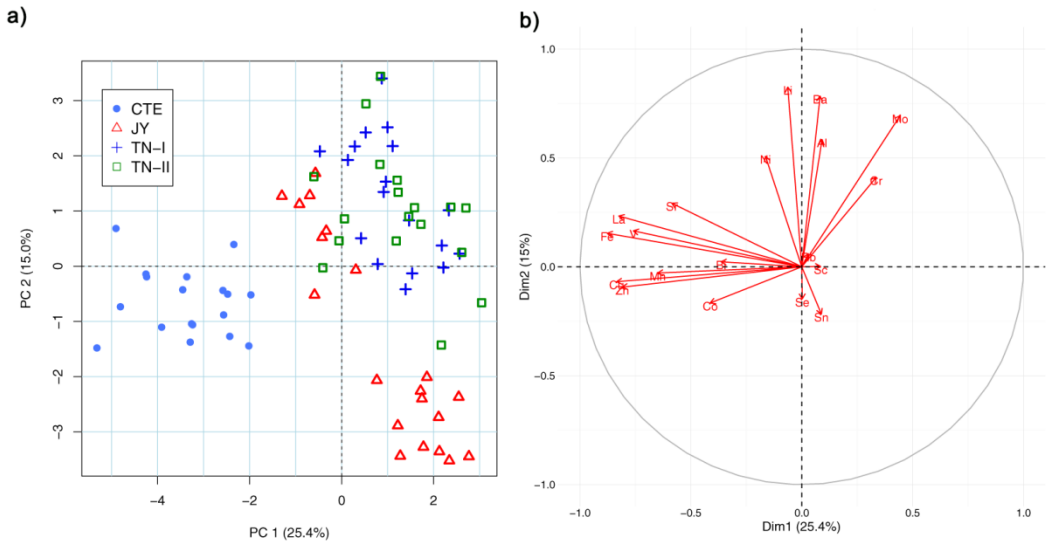

Figura 4.1. Distribución de observaciones (a) y variables (b) en el espacio formado por las dos primeras componentes principales

En la Figura 4.1a se puede observar una tendencia a agrupaciones entre muestras del mismo origen. Esta figura muestra claramente el agrupamiento de muestras en dos grupos principales según CP1, el primer grupo formado por muestras CTE (región NEA) y el segundo grupo de muestras JY, TN-I y TN-II (región NOA). Esta distribución se puede interpretar a partir de la gráfica de *loadings* (Figura 4.1b) que indica que las concentraciones de Fe, La, V, Cu y Zn son mayores para las muestras de CTE. Se puede obtener información adicional de estos gráficos, que sugieren que Mo y Cr tienen valores más altos para muestras TN, mientras que las muestras JY aparecen divididas en dos subgrupos, más próximas a las muestras de TN, pero asociadas a

valores de Ni y Sr y proyecciones positivas en PC2, y otro subgrupo mejor resuelto, por puntuaciones negativas en PC2, que corresponden a bajas concentraciones de Ni y Li.

En resumen, el PCA mostró la presencia de agrupación natural en las muestras según su origen geográfico. Por lo tanto, a efectos de obtener modelos de clasificación adecuados y eficaces para diferenciar muestras de jugo de limón, se aplicaron métodos supervisados de aprendizaje.

4.3.3. Análisis supervisado de muestras

Con el fin de construir los diferentes modelos clasificatorios de muestras de acuerdo al origen geográfico de las mismas, la matriz de datos se subdividió en dos subconjuntos: un conjunto de *entrenamiento* (70% de los casos) para optimizar los modelos y un conjunto de *pruebas* (30%) para calcular el desempeño como modelo predictivo de clasificación. A continuación, se utilizó el método de validación cruzada *k-fold,* repetido *10* veces en el conjunto de entrenamiento para construir los diferentes clasificadores. En el método de validación cruzada *k-fold,* un conjunto de datos se divide en *k* subconjuntos de igual tamaño, uno de los subconjuntos se selecciona para las pruebas y los otros se utilizan para el entrenamiento. Este procedimiento se repite *n* veces, por lo que cada subconjunto se utiliza para probar al menos una vez, en este caso se utilizó $k = 10$ y $n = 5$. Para la generación de los subconjuntos se utilizó un muestreo estratificado que permitió mantener la representación de cada grupo en los subconjuntos de validación. Los casos incluidos en cada conjunto se cambian aleatoriamente para cada modelo reproducido.

Como ya se anticipó, en este capítulo, se probaron cinco técnicas diferentes (LDA, PLS-DA, KNN, SVM y RF) con el fin de clasificar el jugo de limón según su origen geográfico. PLS-DA, KNN, RF y SVM necesitan optimizar varios parámetros de manera tal que se seleccione un número adecuado de éstos para construir el modelo: selección del número de componentes significativos para PLS-DA; número *k* de vecinos para KNN; número de variables probadas en cada división y número de árboles para RF;

factor de penalización C y ε de la función *kernel* para SVM. Todos ellos se calcularon mediante la técnica de validación cruzada tipo subgrupos (k-fold), y el porcentaje de acierto global se utilizó como criterio para la optimización.

Una vez seleccionados los valores óptimos para cada modelo, se analizaron su aplicación a los respectivos conjuntos de prueba. Para comparar los desempeños obtenidos, aparte del porcentaje de acierto global, se tuvieron en cuenta la sensibilidad (muestras pertenecientes a la clase y clasificadas correctamente en esta clase) y la especificidad (muestras que no pertenecen a la clase modelada y clasificadas correctamente como no pertenecientes) En la Tabla 4.2. se resumen los resultados obtenidos después de la aplicación de los diferentes modelos de clasificación.

Tabla 4.2. Valores de sensibilidad y especificidad obtenidos con diferentes modelos aplicados en un mismo conjunto de datos de contenidos minerales en jugos de limón

	LDA		PLS-LDA		KNN		SVM		RF	
Sitio	Sen (%)	Esp (%)	Sen (%)	Esp (%)	Sen (%)	Esp (%)	Sen (%)	Esp (%)	Sen (%)	Esp (%)
CTE	100	100	100	87	100	100	100	100	100	100
JY	100	93	67	87	83	93	100	100	100	80
TN-I	60	75	80	80	60	81	60	81	60	87
TN-II	--	87	20	100	20	81	40	88	20	93
% de acierto	66,7		66,7		66,7		**76,2**		71,4	

Como se observa, el orden de las tasas de identificación exitosas fue el siguiente: SVM 76% > RF 71% > LDA 67% = PLS-DA 67% = KNN 67%. Los modelos lineales, como LDA y PLS-DA, junto con KNN presentaron un rendimiento similar desde el punto de vista del porcentaje de acierto global, sin embargo, no resuelven completamente el problema de clasificación.

En general, las muestras de la región NE (CTE) pudieron clasificarse correctamente por los cinco modelos propuestos, excepto por PLS-DA (que clasificó erróneamente dos muestras JY como CTE). RF presentó un mayor rendimiento para

clasificar las muestras CTE y JY, pero no pudo resolver las muestras de TN-I y TN-II. La mejor clasificación para muestras de jugo de limón se logró utilizando el método SVM con la función kernel tipo RBF, con los parámetros C = 32 y épsilon = 0,38. El porcentaje de acierto general de la clasificación fue del 76,2%, debido a las similitudes entre los grupos TN-I (3 muestras mal clasificadas) y TN-II (2 muestras mal clasificadas) en términos de contenido de oligoelementos. La clasificación fue 100% correcta para las otras muestras (grupos CTE y JY). La mejor clasificación lograda por SVM, que es un modelo no lineal, generalmente se debe a la flexibilidad y capacidad del algoritmo para crear un modelo generalizado, incluso para grupos de entrenamiento pequeños.

En líneas generales, todas las técnicas utilizadas fueron capaces de clasificar correctamente, con adecuada sensibilidad y especificidad, las muestras provenientes de Corrientes y Jujuy, no obstante, las muestras de las dos zonas de Tucumán no se han podido diferenciar entre sí de la misma manera, lo que afecta en gran medida el porcentaje de acierto general. Por este motivo se decidió agrupar las muestras de acuerdo con la provincia de origen, obteniéndose 3 zonas: CTE, JY y TN (TN-I + TN-II), para las cuales se repitió el procedimiento de análisis explicado anteriormente.

4.3.4. Comparación de modelos

Por último, en la búsqueda de mejorar los porcentajes de acierto de los modelos propuestos, se compararon los rendimientos de clasificación teniendo en cuenta únicamente las provincias de origen como factor de agrupamiento. Las distribuciones obtenidas se muestran en la Tabla 4.3.

Tabla 4.3. Porcentaje de acierto y coeficientes Kappa para los diferentes métodos estudiados

Método	Porcentaje de acierto	κ
LDA	95	0,90
PLS	95	0,90
KNN	95	0,90
RF	95	0,90
SVM	**100**	**0,98**

Los resultados obtenidos confirman que el mejor método para este conjunto de datos fue SVM, con un acierto medio del 100% y un índice κ de 0,98.

Estos valores de acierto obtenidos con los diferentes métodos de análisis de datos se encontraron en el mismo orden a los obtenidos por Potortí *et al.* (2018) al clasificar, limones con indicación geográfica.

4.4. Resumen de resultados

La técnica de ICP MS ha permitido detectar, en muestras de jugo de limón, las concentraciones de Al, Ba, Bi, Co, Cr, Cu, Fe, La, Li, Mn, Mo, Ni, Rb, Sc, Se, Sn, Sr, V y Zn, las concentraciones de 5 oligoelementos (Ag, As, Ga, Sb y Tl) no pudieron determinarse dado que se encontraron por debajo de los límites de detección (LOD) en todas las muestras.

Las concentraciones de oligoelementos (contenidos por encima de 1,0 μg/g) se presentaron en el siguiente orden decreciente: Fe > Zn > Rb > Cu > Sr > Mn > Ba > Al > Ni. Teniendo en cuenta la concentración media de todas las muestras, el oligoelemento más abundante fue Fe seguido de Zn y Rb, todos ellos se encontraron en contenidos medios superiores a 10 μg/g, Al, Ba, Cu, Mn, Ni y Sr presentaron contenidos que oscilaban entre 1 y 10 μg/g, Por otro lado, se observaron las mayores discrepancias entre diferentes muestras para los siguientes elementos ultra traza (concentraciones inferiores a 1,0 μg/g): La, Cr, Se, Li, Mo, Co, Sn, Sc, V y Bi.

El análisis exploratorio mostró la existencia de patrones para separar las muestras según el origen geográfico, entre regiones basados en las concentraciones de Fe, La, V, Cu, Zn y dentro del NOA por los contenidos de Mo, Cr, Ni, Li y Sr. Las muestras de la región NEA se clasificaron correctamente con respecto a las muestras NOA, estas últimas presentaron dificultades en su resolución de acuerdo con el lugar de origen. Los mejores resultados se logran cuando se considera la provincia de origen de las muestras como factor de clasificación. La diferenciación de los jugos entre

regiones se puede establecer en base a los contenidos de Fe, La, V, Cu y Zn, en función de los resultados obtenidos en el PCA.

En general, las muestras de la región NEA (CTE) pudieron clasificarse correctamente por los cinco modelos propuestos, excepto por PLS-DA (que clasificó erróneamente dos muestras JY como CTE).

Cuando se trabajó con los 4 sitios (CTE, JY, TN-I y TN-II) las técnicas utilizadas fueron capaces de diferenciar las muestras de CTE y JY, no obstante, las muestras de TN-I y TN-II no se pudieron diferenciar de la misma manera, lo que afecta en gran medida el porcentaje de acierto general (SVM 76% > RF 71% > LDA = PLS-DA = KNN 67%).

Los mejores resultados se logran cuando se reduce el factor de agrupamiento a la provincia de origen de las muestras (CTE, JY y TN). El desempeño de los métodos de clasificación propuestos generó resultados en el siguiente orden según sus porcentajes globales de identificación exitosas: SVM 100% > RF = LDA = PLS-DA = KNN 95%.

4.5. Referencias

Amenta, M; Ballistreri, G; Fabroni, S; Romeo, FV; Spina, A; Rapisarda, P. 2015. Qualitative and nutraceutical aspects of lemon fruits grown on the mountainside of the Mount Etna: A first step for a protected designation of origin or protected geographical indication application of the brand 'Limone dell'Etna. International Food Research. (74): 250-259.

Barros, HMR; Ferreira, TAPC; Genovese, MI. 2012. Antioxidant capacity and mineral content of pulp and peel from commercial cultivars of citrus from Brazil. Food Chemistry. 134: 1892-1898.

Bro R; Smilde A. 2014. Principal component analysis. Anal. Methods. 6: 2812-2831 pp.

Federcitrus. 2018. La actividad citrícola Argentina. Disponible de: https://www.federcitrus.org/. Visita: 04/09/2019.

Forina, M; Casale, M; Oliveri, P. 2009. Application of Chemometrics to Food Chemistry, in: S.D. Brown, R. Tauler, B. Walczak (Eds.) Comprehensive Chemometrics. Elsevier, Oxford. 75-128.

Gonzálvez, A; Armenta, S; de la Guardia, M. 2009. Trace-element composition and stable-isotope ratio for discrimination of foods with Protected Designation of Origin. TrAC Trends in Analytical Chemistry. 28 (11): 1295-1311.

Kelly, S; Heaton, K; Hoogewerff, J. 2005. Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. Trends in Food Science & Technology. 16 (12): 555-567.

Lorente, J; Vegara, S; Martí, N; Ibarz, A; Coll, L; Hernández, J; Valero, M; Saura, D. 2014. Chemical guide parameters for Spanish lemon (*Citrus limon* (L.) Burm.) juices. Food Chemistry. (162): 186-191.

Matheyambath, AC; Padmanabhan, P; Paliyath, G. 2016. Citrus Fruits. Encyclopedia of Food and Health. Academic Press, Oxford. 136-140.

Potortí, AG; Di Bella, G; Mottese, AF; Bua, GD; Fede, MR; Sabatino, G; Salvo, A; Somma, R; Dugo, G; Lo Turco, V. 2018. Traceability and Protect Geographic Indication (PGI) Interdonato lemon pulps by chemometric analysis of the mineral composition. Journal of Food Composition and Analysis. Accepted manuscript. https://doi.org/10.1016/j.jfca.2018.03.001.

Simpkins, WA; Louie, H; Wu, M; Harrison, M; Goldberg, D. 2000. Trace elements in Australian orange juice and other products. Food Chemistry. 71(4): 423-433.

Szymczycha Madeja, A; Welna, M. 2013. Evaluation of a simple and fast method for the multi-elemental analysis in commercial fruit juice samples using atomic emission spectrometry. Food Chemistry. 141(4): 3466-3472.

Szymczycha Madeja, A; Welna, M; Jedryczko, D; Pohl, P. 2014. Developments and strategies in the spectrochemical elemental analysis of fruit juices. TrAC Trends in Analytical Chemistry. (55): 68-80.

Tufour, J; Bentum, J; Essumang, D; Koranteng Addo, J. 2011. Analysis of heavy metals in citrus juice from the Abura-Asebu-Kwamankese District, Ghana. Journal of Chemical and Pharmaceutical Research. 3 (2): 397-402.

USDA. 2020. United States Departament of Agriculture, Citrus Fruits USDA Foreign Agricultural Service. Disponible en línea: http://www.fas.usda.gov/commodities/citrus-fruit. Visita: 12/02/2020.

CAPÍTULO V

DETERMINACIÓN DEL ORIGEN GEOGRÁFICO DE JUGOS DE MANDARINAS PRODUCIDAS EN EL NORDESTE DE ARGENTINA

5.1. Introducción

El término mandarinos agrupa a los árboles de varias especies (*Citrus unshiu* Marcovitch (Satsumas), *C. nobilis* Loureiro (King), *C. deliciosa* Tenore (Común), *C. Clementina* Hort. ex Tanaka (Clementinas), *C. reticulata* Blanco (Dancy, Ponkan), *C. reshni* Hort. ex Tanaka (Cleopatra), *C. sunki* Hort. ex Tanaka (Sunki y Suenkat)). La envergadura de los árboles y las características de sus hojas, flores y frutos los convierten en especies ornamentales para jardines, pero su principal destino productivo son los frutos, denominados mandarinas, de alto valor alimenticio (Palacios, 2013).

En el año 2017 la Argentina se ubicó octava a nivel mundial en cuanto a la producción de frutos cítricos, con un total de 3.272.771 t. Alrededor del 27% de la producción se destinó al consumo interno como fruta fresca, cerca del 46% a la industrialización (incluye la producción de jugos y esencias, entre otros), y aproximadamente un 11% fue exportado a diferentes mercados. Los principales compradores de los cítricos argentinos son: España (19%), Rusia (18%) y Holanda (11%). El nordeste argentino (NEA), que comprende la llamada Mesopotamia Argentina, es la tradicional zona citrícola, en la que las importantes producciones de las provincias de Entre Ríos, Corrientes y Misiones contribuyen con el 37% de la producción total del país. El 15% de los frutos cítricos producidos en Argentina corresponde a mandarinas, de las cuales el 88% es aportado por el NEA (SENASA, 2014; Federcitrus, 2018; FAO, 2017).

En la actualidad, los integrantes de la cadena agroalimentaria se interesan por conocer y garantizar el origen geográfico e identidad de los productos agropecuarios y, en especial, de aquellos que intervienen en la producción de alimentos derivados. La certificación de origen puede ser un elemento esencial para asegurar la autenticidad de un determinado producto alimentario dado que protege, sobre todo, a los productos regionales y confirma características de calidad relacionadas con su origen (Drivelos & Georgiou, 2012; Luykx & van Ruth, 2008).

El SITC®-NEA es el primer sistema de información de argentino, y quizás de otros países, que permite conocer todo el proceso de un producto, en este caso de los cítricos, desde el campo hasta el destino final. Si bien logra la trazabilidad de los cítricos producidos, se basa en información documental y no contempla mecanismos para comprobar la identidad física de las muestras en cualquier etapa de la cadena productiva, por lo que puede ser vulnerable a pérdida de información o maniobras de adulteración o contaminación. Se plantea, entonces, la necesidad de definir estrategias para establecer la trazabilidad y rastreabilidad de frutos cítricos, aunque no se disponga de información de toda la cadena de producción, procesamiento y comercialización, o complementariamente a ella.

Si bien la composición mineral de los vegetales obedece a patrones generales definidos para las diferentes especies y variedades, existe cierto grado de variabilidad que se debe, en gran medida, a condiciones de los sitios donde las plantas crecen (Weil & Brady, 2017). Esta influencia de las condiciones locales en la composición de los tejidos vegetales permite utilizarla para caracterizar el origen específico de productos vegetales, debido a la correlación entre elementos a nivel de vestigios, el tipo de suelo, las condiciones de cultivo y ambientales en origen (Gonzálvez & de la Guardia, 2013).

El análisis de la composición mineral ha sido utilizado por diversos autores para determinar la procedencia geográfica de diferentes alimentos, entre los que se puede destacar la determinación de la procedencia geográfica de vino (Dutra *et al.*, 2013; Geana *et al.*, 2013) mieles (Batista *et al.*, 2012; Di Bella *et al.*, 2015), arroz (Cheajesadagul *et al.*, 2013; Chung *et al.*, 2015; Maione *et al.*, 2016) y leche (Magdas *et al.*, 2016), entre otros.

Los espectrómetros de emisión atómica de plasma por microondas (MP-AES) ofrecen alta sensibilidad, límites de detección inferiores a las partes por billón (ppb), velocidad superior a la de la absorción atómica de llama y no necesita gases combustibles. La nueva generación de espectrómetros MP-AES funciona con aire, lo que reduce enormemente los costos operativos.

El objetivo de este capítulo fue la utilización de información de contenido de elementos minerales en muestras de jugo de mandarina de las variedades 'Okitsu' (*C. unshiu* Marcovitch) y tangor 'Murcott' (*C. sinensis* L. x *C. deliciosa* Tenore), para obtener modelos de clasificación adecuados a efectos de analizar orígenes de jugo de mandarina de diferentes zonas productoras del NEA.

5.2. Materiales y Métodos

5.2.1. Obtención y procesamiento de las muestras

Mediante un muestreo en etapas sobre un padrón de productores con características tecnológicas medias, en cuatro zonas productivas del NEA (centro sur de Misiones: CSMN, centro oeste de Corrientes: COCR, sudeste de Corrientes: SECR y noreste de Entre Ríos: NEER) se seleccionaron 5 lotes al azar simple, en cada lote 5 plantas por azar sistemático ubicadas a lo largo de una línea que recorría el lote de NO a SE, de las que se extrajeron 10 frutos por planta de los que se obtuvo el jugo (unidad muestral), se realizó la digestión por vía seca y la determinación multielemental.

En este capítulo se evalúa la eficacia de la información obtenida mediante espectroscopía de emisión atómica de plasma de microondas (MP-AES), para detectar la presencia de elementos que permitan conocer el origen de las muestras de jugos de mandarina de diferentes zonas de la región NEA.

5.2.2. Datos

Mediante la técnica de MP-AES permitió determinar en los jugos de mandarina las concentraciones de Al, Ca, Cd, Cr, Cu, Fe, K, Mn, Mg, Sr y Zn; las concentraciones de Ba, Co, Mo, Ni, Pb, Se y Sb no se informan, dado que se encontraron por debajo de los límites de detección (LOD) en todas las muestras.

El conjunto de datos con el que se trabajó estuvo formado por una variable que identificaba los sitios de origen de las muestras (COCR, CSMN, NEER y SECR) y 11 variables clasificadoras que corresponden a los contenidos, en jugos de mandarinas, de los siguientes elementos: Al, Ca, Cd, Cr, Cu, Fe, K, Mn, Mg, Sr y Zn.

5.2.3. Análisis de datos

En virtud de la dificultad que representa el manejo de grandes cantidades de datos, se utilizaron Métodos de Análisis Estadístico Multivariante y de Aprendizaje Automático, para explorar y proponer modelos matemáticos que permitan clasificar las distintas muestras de acuerdo con criterios preestablecidos y/o establecer criterios de identidad de muestras de acuerdo con su composición química.

Dentro de los Métodos Multivariantes se aplicaron técnicas de Análisis de Componentes Principales (PCA) y Análisis Discriminante Lineal (LDA) (Peña, 2002). Entre las técnicas de Aprendizaje Automático se utilizaron Árboles de Decisión (DT), empleando los algoritmos C5.0, CART; K-vecino más cercano (KNN), buscando la mejor configuración; Redes Neuronales Artificiales (ANN), con datos normalizados, utilizando el algoritmo MLP, probando diferente cantidad de capas y número de neuronas por capa, seleccionando la cantidad de iteraciones en que se estabiliza el porcentaje de acierto; y Máquinas de Vectores Soporte (SVM) (James *et al.*, 2017).

5.3. Resultados y Discusión

Con la información de contenido de los diferentes elementos minerales en los jugos de mandarinas se realizó en primer término una caracterización de los jugos en función de su contenido multielemental y posteriormente un análisis de las diferentes zonas productivas a fin de establecer una diferenciación geográfica.

5.3.1. Caracterización multielemental

Los promedios, desviaciones estándares, mínimos y máximos de las concentraciones de los diferentes elementos estudiados se presentan en la Tabla 5.1. Las concentraciones de Ba, Co, Mo, Ni, Pb, Se y Sb no se muestran porque se encontraban por debajo de los límites de detección (LOD) en todas las muestras.

Tabla 5.1. Concentraciones de elementos minerales en muestras de jugo de mandarina (μg/g). Promedios (Me), desviaciones estándares (DE), mínimos (Mín) y máximos (Máx), por variedad

	Variedad							
	MUR				OKI			
Elemento	Me	DE	Mín	Máx	Me	DE	Mín	Máx
Al	44,34	68,85	4,69	339,04	130,4	355,6	1,55	1649,93
Ca	15,53	5,5	8,35	31,22	14,77	9,01	5,88	37,23
Cd	0,54	0,6	0	1,59	2,09	8,26	0	39,91
Cr	0,19	0,26	0	0,87	0,22	0,3	0	1,08
Cu	2,45	4,65	0,2	16,7	9,13	22,38	0,1	98,34
Fe	0,32	0,21	0,12	0,92	0,35	0,26	0,08	0,88
K	937,69	334,26	277,69	1715,2	920,5	368,89	91,37	1822,66
Mg	115,26	43,15	69,23	222,74	109,11	48,68	45,94	205,44
Mn	3,17	4,64	0,3	13,49	4,3	7,95	0,15	27,92
Sr	1,02	1,98	0,13	7,12	2,2	4,73	0,1	19,78
Zn	1,87	2,13	0,09	6,99	3,69	7,64	0,13	31,3

Las muestras de ambas variedades presentan perfiles de concentración diferentes. Teniendo en cuenta la concentración media de todas las muestras, el elemento más abundante fue K con contenidos medios cercanos a los 1000 μg/g, seguido de Mg con contenidos medios superiores a 100 μg/g. En orden de importancia

por sus contenidos medios le sigue Al, con valores medios ubicados entre 44 y 130 μg/g, pero valores máximos superiores a los del Mg; Ca presentó valores medios de 15 μg/g; los contenidos de Mn, Cu, Zn y Sr se encontraron entre 1 y 5 μg/g; y Cd, Cr y Fe, con contenidos medios inferiores a 1 μg/g.

En comparación con otros jugos de frutas cítricas, como naranjas o pomelos, los contenidos reportados de K, Mg, Ca, Mn, Cu, Zn, Sr, Fe y Cr se encontraban en niveles similares (Simpkins *et al*., 2000). En general, los contenidos de Al y Cd estaban en niveles más altos de lo reportado para el jugo de naranja comercial de Polonia (Szymczycha Madeja & Welna, 2013).

Si bien son varios los factores que pueden contribuir a variaciones en los niveles de los diferentes elementos en los jugos de mandarina según su origen geográfico se deben tener en cuenta la disponibilidad del elemento en el suelo para su extracción por la planta, las prácticas agrícolas, las aplicaciones de fertilizantes, el riego artificial o la madurez de los frutos en la cosecha, que pueden influir en las concentraciones de elementos (Szymczycha Madeja *et al*., 2014).

5.3.2. Análisis exploratorio

En una etapa previa al modelado de clasificación se realizó un PCA para el análisis exploratorio de los datos, lo que permitió visualizar la estructura de datos en una dimensión reducida, manteniendo la máxima cantidad información presente en los mismos. Se extrajeron dos componentes principales que explican el 94,9% de la variabilidad total (CP1: 75,3% y CP2: 19,6%). La Figura 5.1 presenta un Biplot con los resultados obtenidos en el espacio formado por las dos primeras componentes principales.

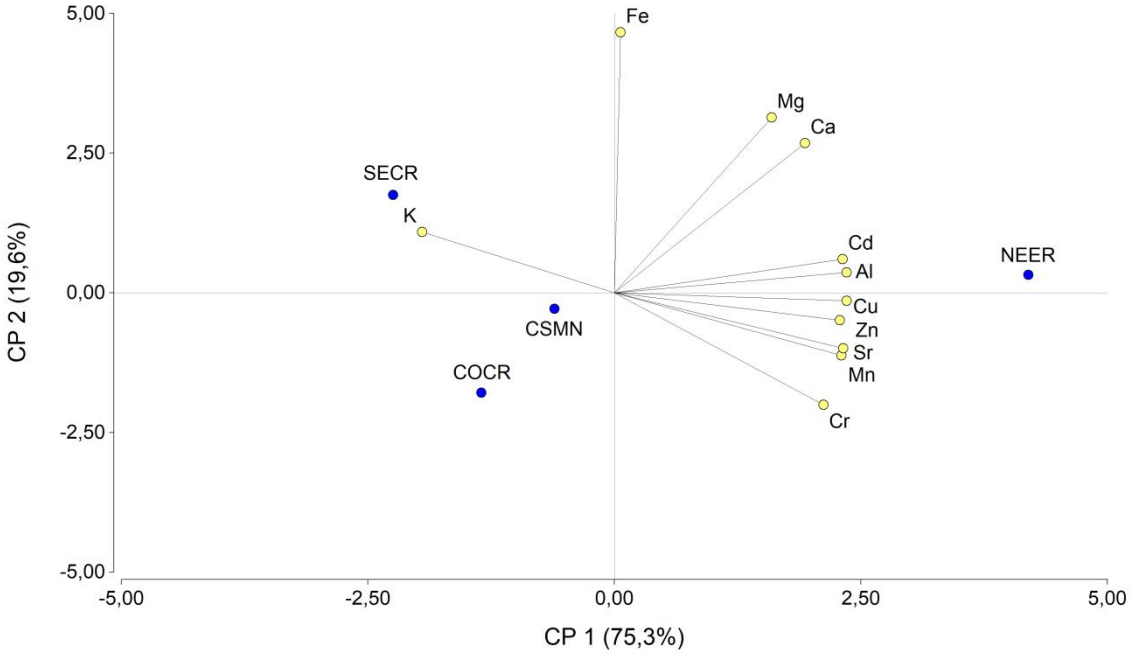

Figura 5.1. Biplot que representa los contenidos de elementos minerales en jugo de mandarina y las zonas productoras (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) sobre las dos primeras coordenadas resultantes del Análisis de Componentes Principales

Esta figura muestra la separación de muestras en dos grupos principales según CP1. El primer grupo formado por muestras del nordeste de Entre Ríos y el segundo grupo de muestras correspondientes a Corrientes y Misiones. Esta distribución se puede interpretar en función de las concentraciones de K, que son mayores para las muestras de sudeste y centro oeste de Corrientes y centro sur de Misiones, mientras que los demás elementos presentan contenidos superiores en el nordeste de Entre Ríos. Sobre la segunda coordenada se podría diferenciar la zona de sureste de Corrientes, asociada a mayores contenidos de Ca, Fe y Mg, de las del centro sur de Misiones y centro oeste de Corrientes, estas últimas zonas presentan perfiles muy similares y no logran separarse.

5.3.3. Selección de modelos

Debido a que el PCA mostró la presencia de agrupación natural en las muestras según su origen geográfico, a efectos de obtener modelos de clasificación para

diferenciar muestras de jugo de mandarinas se aplicaron métodos de reconocimiento de patrones de aprendizaje supervisado, los que se presentan a continuación.

Una vez seleccionados los valores óptimos para cada modelo, la clasificación alcanzada con los diferentes métodos en el conjunto de pruebas fue evaluada considerando los criterios de sensibilidad, y porcentaje de acierto general (Marcelo *et al*., 2014). También se empleó el índice Kappa (κ) utilizado anteriormente, que aplicado a la matriz de confusión permite evaluar si la clasificación observada es similar (concordante) con la clasificación predicha por el clasificador (Warrens, 2020).

5.3.3.1. Análisis Discriminante Lineal

En la Figura 5.2 se observa la distribución de la proyección de las muestras en el espacio definido por las 2 primeras variables discriminantes identificadas de acuerdo con las zonas productoras. Se puede observar una tendencia en las muestras a agruparse de acuerdo con la zona de origen. Las observaciones del noreste de Entre Ríos se ubican entre el 1ro y 4to cuadrante, perfectamente separadas de las otras tres zonas. El grupo de observaciones de las otras tres zonas presentan una elevada superposición entre ellas, aunque las muestras del sureste de Corrientes muestran una tendencia a proyectarse en el 3er cuadrante.

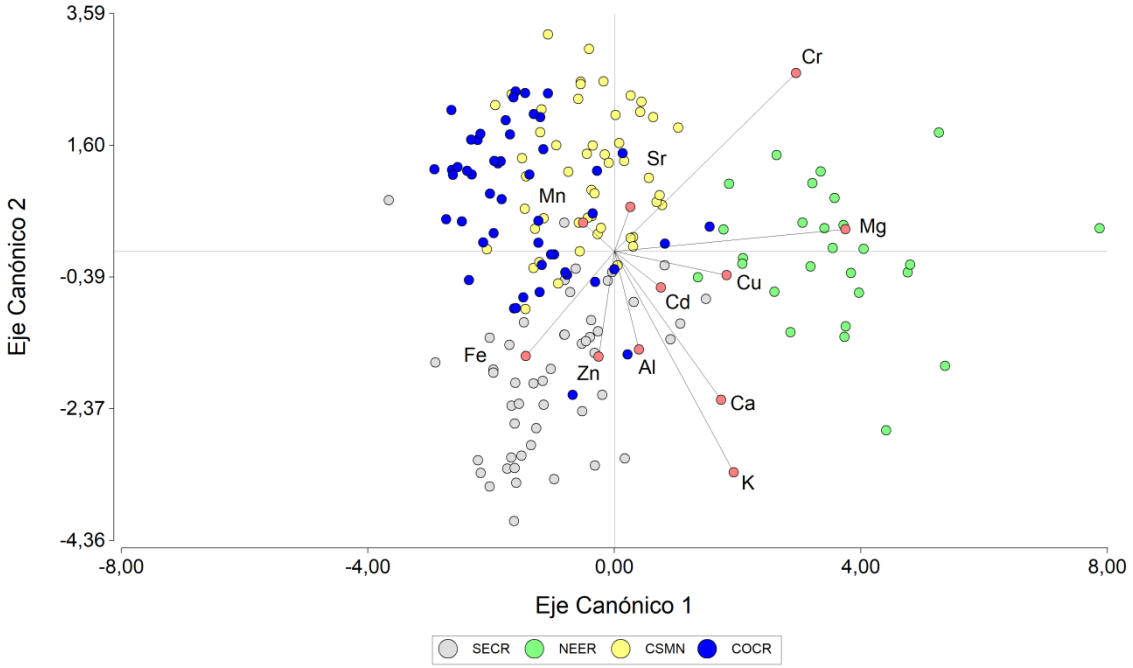

Figura 5.2. Biplot que representa los contenidos de elementos minerales en jugo de mandarina y las zonas productoras (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) sobre los dos primeros ejes canónicos del Análisis Discriminante Lineal

Las funciones discriminantes canónicas son presentadas en la Tabla 5.2.

Tabla 5.2. Funciones Discriminantes Canónicas

	1	2
Al	0,11	-0,40
Ca	0,47	**-0,61**
Cd	0,21	-0,15
Cr	**0,80**	**0,73**
Cu	**0,50**	-0,10
Fe	-0,39	**-0,43**
K	**0,53**	**-0,91**
Mg	**1,02**	0,09
Mn	-0,14	0,12
Sr	0,07	0,18
Zn	-0,07	**-0,43**

La primera función discriminante, que permite separar los jugos provenientes del nordeste de Entre Ríos, se construye, principalmente, con los contenidos de Mg, Cr, K y Cu. La segunda función discriminante se construye principalmente con los contenidos de K, Cr, Ca, Fe y Zn.

A continuación, se presentan los resultados de interés de la clasificación con Análisis Discriminante Lineal para evaluar el desempeño del método (Tabla 5.3).

Tabla 5.3. Criterios de selección de modelos para el Análisis Discriminante Lineal por zona de producción (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona	Acierto (%)	Sensibilidad	Especificidad	κ
COCR	66	0,66	0,99	0,72
CSMN	92	0,92	0,90	0,76
NEER	100	1,00	0,97	0,95
SECR	82	0,82	0,94	0,76

El comportamiento general del método fue bueno, con un porcentaje de acierto general de 86% e índice κ = 0,80. Al analizar por zona productiva, como se observa en la tabla 5.3 y la figura 5.2, el método no logra el mismo comportamiento para clasificar muestras de todas las zonas, en el caso de COCR la sensibilidad fue del 66% muy por debajo de las demás, lo que indica que no sería adecuado su uso para clasificar estas muestras.

Estos resultados obtenidos con jugos de mandarina son inferiores a los obtenidos por Benabdelkamel *et al.* (2012), logrando con LDA porcentajes de acierto del 96,6%, al clasificar mandarinas clementinas según su origen. Asimismo, son levemente superiores a los hallados con la aplicación del LDA en jugos de limón clasificados según 4 sitios de muestreo y menores a los obtenidos cuando se clasificaron los jugos de limón según las tres provincias, que se presentan en el Capítulo IV de esta tesis (Gaiad *et al.*, 2016), donde el método demostró correspondencia en los valores de sensibilidad y especificidad por zona, que es lo que se espera de un buen método de clasificación (Takaya & Rehmsmeier, 2015).

5.3.3.2. Árboles de Decisión

Se generaron Árboles de Decisión utilizando diferentes algoritmos, los que han presentado comportamientos similares en relación con los criterios de decisión empleados y para las diferentes zonas productoras. En la Tabla 5.4 se presentan los

valores de los criterios de selección de modelos para los casos de mejor comportamiento.

Tabla 5.4. Criterios de selección de modelos por algoritmo usado en la generación de Árboles de Decisión

Algoritmo	Porcentaje de acierto	Índice κ
C5.0	**0,91**	**0,88**
CART (Rpart)	0,83	0,77
CART (Rpart 2)	0,83	0,77

Dentro de los Árboles de Decisión, el algoritmo C5.0 presentó mejor comportamiento que los CART (Rpart y Rpart2) en relación con el porcentaje de acierto y el índice κ, el clasificador obtenido es de tipo regla, sin filtrado de variables y 20 iteraciones. El algoritmo C5.0 arrojó un porcentaje de acierto general de 91% y un índice κ global de 0,88 (muy buena concordancia). En la tabla 5.5, se presentan los valores de sensibilidad y especificidad, logradas por zona productiva con la aplicación de este algoritmo.

Tabla 5.5. Sensibilidad y especificidad logradas con Árboles de Decisión generados con el algoritmo C5.0 por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona productiva	Sensibilidad	Especificidad
COCR	0,95	0,98
CSMN	0,70	1,00
NEER	1,00	0,95
SECR	1,00	0,95

En la Tabla 5.6 se presentan las reglas de clasificación obtenidas mediante el algoritmo C5.0 considerando cada zona de producción.

Tabla 5.6. Reglas de decisión para la clasificación de los jugos de mandarina procedentes de las diferentes zonas productivas (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) obtenidas mediante Árboles de Decisión mediante algoritmo C5.0

Reglas	Zona productiva	Acierto (%)
Cr ≤ 0,10, Mn > 2,97	COCR	76,1
Cr > 0,10	COCR	62,4
Cr ≤ 0,55, Fe ≤ 0,38, K ≤ 846,18	CSMN	95,3
Al ≤ 12.37, Mn ≤ 2.97	CSMN	89,1
Cr > 0,55	NEER	89,9
Cr > 0,10, K > 846,18, Mg > 123,48	NEER	79,8
Al > 12,37, Cr ≤ 0,10, Mn ≤ 2,97	SECR	91,9

Al analizar por zona de producción (Tabla 5.5), para el noreste de Entre Ríos, sudeste y centro oeste de Corrientes se consigue alta sensibilidad y especificidad, mientras que en el centro sur de Misiones se presenta la mayor especificidad, pero con baja sensibilidad.

Como se puede ver en la Tabla 5.6, la separación de los jugos de mandarina procedentes del centro oeste de Corrientes fue la que menores porcentajes de acierto obtuvo con valores entre 62,4 y 76,1 %, debido a que muestras de centro oeste de Corrientes fueron clasificadas como procedentes de centro sur de Misiones, afectando la sensibilidad de dicha zona. Estos resultados son similares a los informados por Maione *et al*. (2016), quienes demostraron que los DT son modelos de clasificación eficientes, aún con valores de sensibilidad menores al 100%.

Debido a que una adecuada sensibilidad se presenta juntamente con una buena especificidad, de acuerdo con lo establecido por Takaya & Rehmsmeier (2015), se considera que la técnica de Árboles de Decisión con algoritmo C5.0, presenta muy buen comportamiento para clasificar jugos de mandarinas en función de su contenido mineral en tres de las cuatro zonas estudiadas, no así en la zona centro sur de Misiones.

La siguiente figura refleja la importancia de las distintas variables para determinar la trazabilidad de los jugos de mandarina.

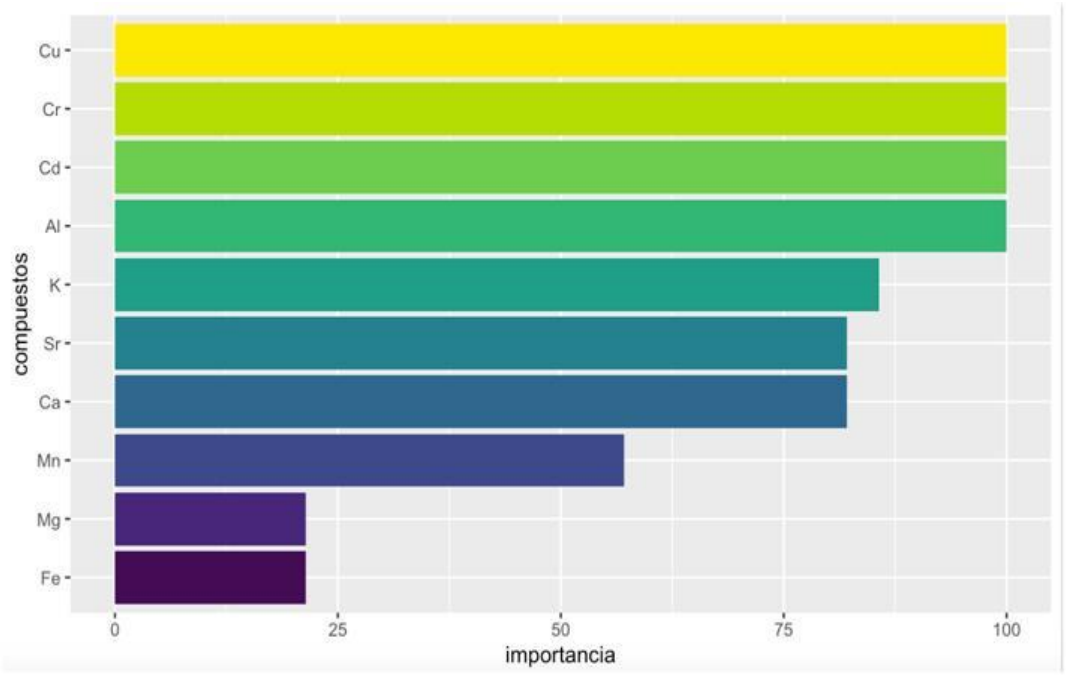

Figura 5.3. Importancia de las distintas variables para determinar la trazabilidad de los jugos de mandarina.

Podemos observar en la Figura 5.3 que el Zn, no resultó importante en la clasificación por zona de los jugos de mandarina, Fe y Mg tienen baja participación (menos del 25%) y los demás elementos presentan porcentajes de importancia como clasificadores de origen de estos jugos mayores al 50.

5.3.3.3 K-Vecino más Cercano

Con el método KNN se han probado diferentes opciones del parámetro *k* (número de vecinos que se tiene en cuenta para predecir). En la Tabla 5.7 se presentan los porcentajes de acierto e índice κ.

Tabla 5.7. Criterios de selección de modelos para diferentes valores del parámetro k usando métodos vagos de K-Vecinos más Cercanos

Valor de k	Porcentaje de acierto	κ
5	**0,85**	**0,80**
7	0,83	0,77
9	0,82	0,76
11	0,79	0,72
13	0,77	0,68
15	0,76	0,68
17	0,72	0,62
19	0,69	0,59
21	0,68	0,57
23	0,67	0,56

Los mejores resultados se obtuvieron con 5 vecinos más cercanos, no obstante, los valores obtenidos de acierto e índice κ, son relativamente bajos, lo que indica que la técnica de K-Vecinos más Cercanos no sería la mejor opción para clasificar jugos de mandarina en función de su contenido mineral.

Los porcentajes de acierto obtenidos con el método KNN en este capítulo son inferiores a los obtenidos por Li *et al.* (2014), quienes clasificaron arándanos en diferentes estadios de crecimiento y encontraron que el KNN fue la técnica de mejor comportamiento con un porcentaje de acierto entre 85 y 98%.

5.3.3.4. Redes Neuronales Artificiales

Variando el número de capas ocultas, la cantidad de neuronas por capa y de iteraciones, con la técnica de Redes Neuronales Artificiales (ANN) se han podido generar diferentes modelos. En la Tabla 5.8 se presentan los criterios de selección para los diferentes modelos definidos mediante ANN.

Tabla 5.8. Criterios de selección de modelos por algoritmo usado en la generación de Redes Neuronales Artificiales

Nro de capas ocultas	Neuronas por capa oculta	Acierto (%)	Índice κ
1	10	0,92	0,90
1	**25**	**0,96**	**0,94**
1	35	0,93	0,91
2	16-13	0,93	0,91
3	14-10-5	0,91	0,89

Los mejores resultados se obtuvieron con el algoritmo MLP con 1 capa oculta de 25 neuronas, lográndose un acierto del 96% y un índice κ = 0,94. En la Tabla 5.9 se presentan la sensibilidad y especificidad por zona productiva para el modelo escogido. Estos resultados son similares a los informados por Sabanci *et al.* (2016), quienes obtuvieron un 98,89% de acierto al clasificar diferentes variedades de manzana utilizando ANN.

Tabla 5.9. Sensibilidad y especificidad logradas con Redes Neuronales con 1 capa de 35 neuronas por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona productiva	Sensibilidad	Especificidad
COCR	0,85	1,00
CSMN	1,00	0,95
NEER	1,00	1,00
SECR	0,95	0,98

Con este algoritmo se logra la mayor sensibilidad y especificidad para la zona productora del nordeste de Entre Ríos; en el centro oeste de Corrientes se logró el valor máximo de especificidad, pero con menor sensibilidad; en el centro sur de Misiones la mayor sensibilidad, pero menor especificidad; y para el sudeste de Corrientes se registraron menores valores de ambos indicadores.

A continuación, se presentan gráficos que indican la importancia de los diferentes elementos en la construcción del modelo seleccionado por zona productora, en los que se puede observar que para las distintas zonas la incidencia de los elementos en la clasificación de las muestras es diferente.

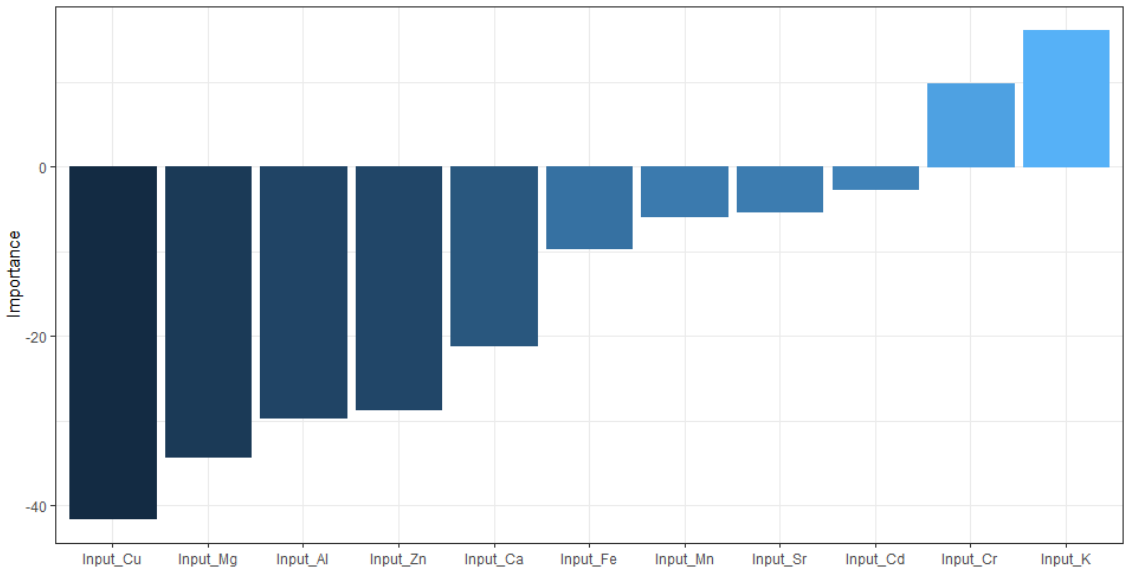

Figura 5.4. Importancia de los elementos en la construcción del modelo ANN para la zona productora del centro oeste de Corrientes

Para lograr una clasificación de los jugos de mandarinas en la zona del centro oeste de Corrientes el método considera principalmente los contenidos de Cu y Mg, siendo los demás elementos analizados menos importantes, Cr y K se presentan en sentido positivo, opuesto a los demás.

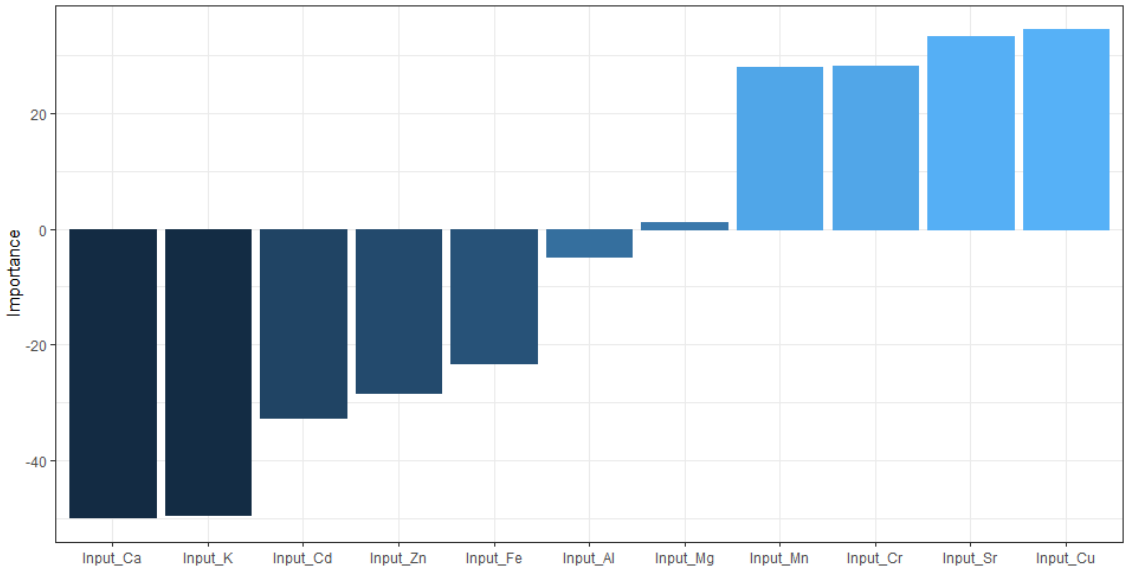

Figura 5.5. Importancia de los elementos en la construcción del modelo ANN para la zona productora del centro sur de Misiones

Se puede identificar al Ca y K con sentido negativo y Cu y Sr positivo, como los elementos más importantes en la clasificación de jugos provenientes del centro sur de Misiones.

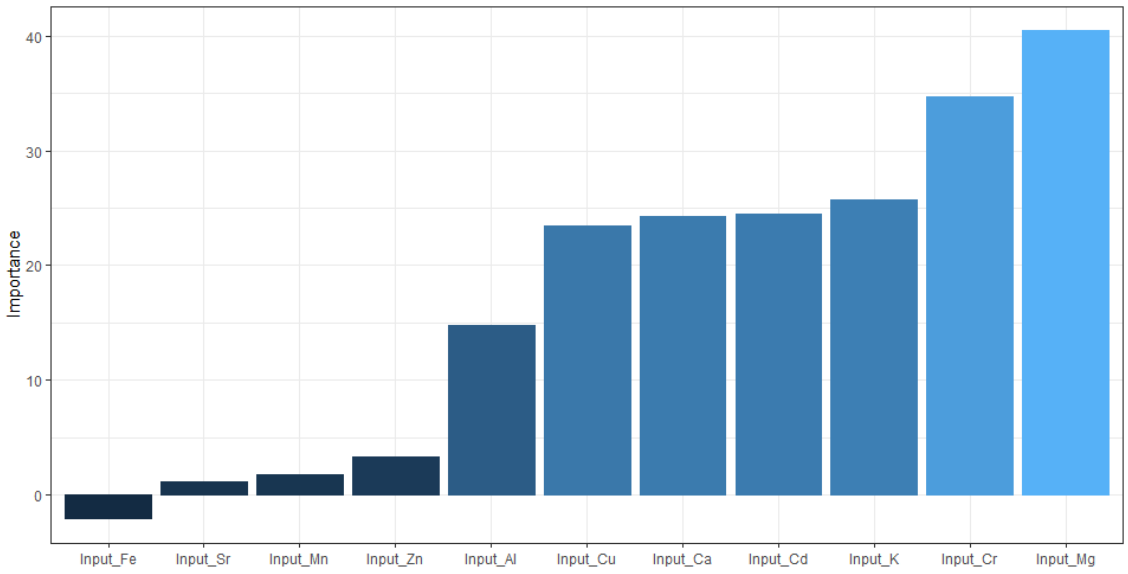

Figura 5.6. Importancia de los elementos en la construcción del modelo ANN para la zona productora del noreste de Entre Ríos

En la clasificación de las muestras provenientes del noreste de Entre Ríos, se observa en primer lugar que, a excepción del Fe, los demás elementos analizados se presentan con sentido positivo, siendo los más importantes Cr y Mg.

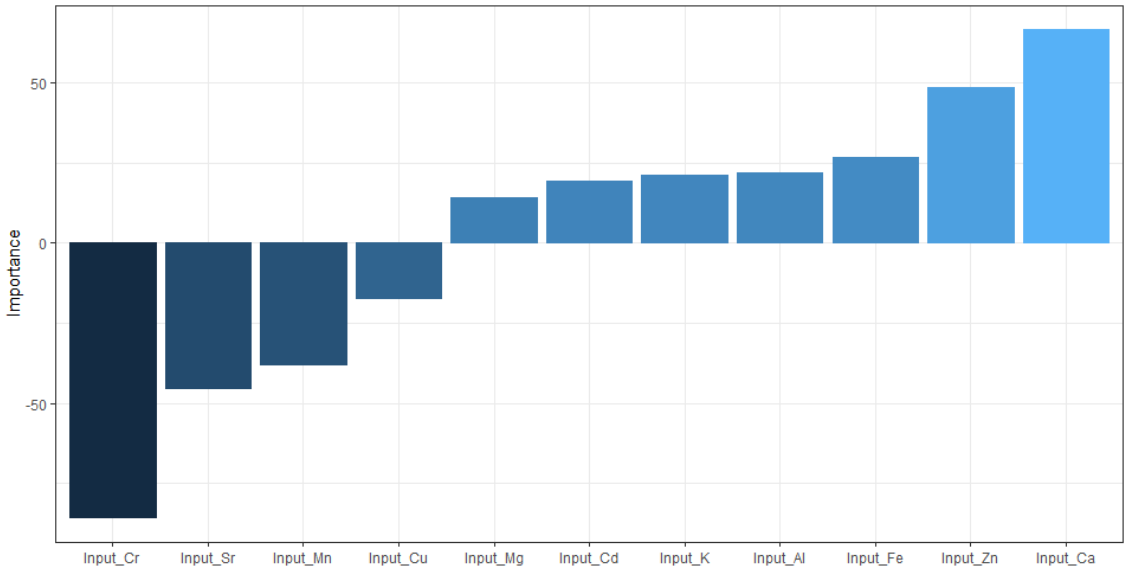

Figura 5.7. Importancia de los elementos en la construcción del modelo ANN para la zona productora del sudeste de Corrientes

Para la clasificación de los jugos del sudeste de Corrientes los elementos más importantes son Ca y Zn en sentido positivo y Cr, Sr y Mn en sentido negativo.

Analizando las Figuras 5.4 a 5.7 en conjunto se puede establecer que, dentro de los elementos con importancia en la discriminación, los que se encuentran presentes en al menos dos zonas productivas son Ca, Cr, Cu, Mg y Sr.

5.3.3.5. Máquinas de Vectores Soporte

Las SVM calculan un hiperplano de separación óptimo mediante un algoritmo iterativo que aprende la distribución de la muestra en los límites de cada clase considerada. Para evitar el sobreajuste se utilizaron las funciones *kernel* radial y polinomial para la clasificación. En la Tabla 5.8 se presentan los valores de acierto e índice κ para cada modelo.

Tabla 5.8. Criterios de selección de modelos para cada función *kernel* definida en Máquinas de Vectores Soporte

Kernel	Parámetros	Porcentaje de acierto	Índice κ
Radial	σ = 1 y C = 1	0,93	0,91
Radial	σ = 1 y C = 5	0,94	0,91
Radial	σ = 3 y C = 1	0,72	0,63
Polinomial	**Grado = 2 y c= 1**	**0,94**	**0,92**
Polinomial	Grado = 2 y c= 5	0,92	0,90
Polinomial	Grado = 3 y c= 1	0,92	0,89

El modelo con un *kernel* polinomial de grado 2 y c = 1 fue seleccionado por ser el de mejor comportamiento, con un acierto general de 94% y un índice κ de 0,92.

Tabla 5.9. Sensibilidad y especificidad logradas con Máquinas de Vectores Soporte con un *kernel* grado 2 y *c* = 1 por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona productiva	Sensibilidad	Especificidad
COCR	0,85	1,00
CSMN	1,00	0,92
NEER	1,00	1,00
SECR	0,90	1,00

Como se puede ver en la Tabla 5.9 con el algoritmo seleccionado se logra la mayor sensibilidad y especificidad para la zona productora del nordeste de Entre Ríos (100% para ambos valores); en el centro oeste y sudeste de Corrientes se logró el valor

máximo de especificidad, pero con menor sensibilidad; en el centro sur de Misiones la mayor sensibilidad, pero menor especificidad.

Los resultados obtenidos con SVM en esta tesis son similares a los de Astuti *et al.* (2018), quienes desarrollaron un sistema para clasificación automática de frutos aplicando ANN y SVM. Con SVM basado en una clasificación binaria y aplicando una función *kernel* polinomial obtuvieron una tasa de clasificación correcta del 100% para el conjunto de entrenamiento y el de prueba.

5.3.4. Comparación de modelos

A los efectos de comparar los diferentes métodos de clasificación, en la Tabla 5.10 se presentan los criterios de porcentaje de acierto e índice κ.

Tabla 5.10. Criterios de selección de modelos de clasificación

Método	Porcentaje de acierto	κ
LDA	86	0,80
DT	91	0,88
KNN	85	0,80
ANN	**96**	**0,94**
SVM	94	0,92

Comparando estos resultados con los obtenidos en el Capítulo IV (publicado en Gaiad *et al*., 2016), los porcentajes de acierto para los distintos métodos son mayores a los obtenidos al clasificar de acuerdo con 4 sitios (CTE, JY, TN-I y TN-II), pero inferiores a los obtenidos al clasificar de acuerdo con la provincia de origen de las muestras (Corrientes, Tucumán, Jujuy).

Benabdelkamel *et al*. (2012) estudiaron jugos de mandarinas clementinas de diferentes procedencias por su composición multielemental mediante PLS-DA, LDA y SIMCA (modelado independientes de analogías de clase) y obtuvieron resultados superiores a los logrados en este capítulo con el uso de LDA (96,6% de acierto).

Estos resultados son similares a los informados por Maione *et al*. (2016), quienes demostraron que los DT son modelos de clasificación eficientes.

Teniendo en cuenta que los valores de Kappa cercanos a uno indican un perfecto acuerdo entre la predicción del modelo y los valores verdaderos (Lantz, 2015) y mayores a 0,81 indican una muy buena concordancia (Hartling *et al.*, 2012), se puede establecer para las técnicas de mejor comportamiento en la clasificación de jugos de mandarina el siguiente orden, en función del porcentaje de acierto general: ANN 96% > SVM 94% > DT 91% > LDA 86% > KNN 85%.

Debido a los altos porcentajes de acierto obtenidos en todos los métodos, y aquellos que tienen niveles de concordancia muy buenos (ANN, SVM y DT), se debería considerar también la regla de Occam, que propone como más plausible al modelo más sencillo que se ajusta a los datos (Abu Mostafa *et al.*, 2012).

5.3.5. Marcadores químicos de identidad

Algunos de los métodos estudiados permiten detectar las variables de mayor peso en la discriminación. En la Tabla 5.11 se presentan esos clasificadores por método.

Tabla 5.11. Clasificadores con mayor poder discriminante por método usado en la clasificación por origen de jugos de mandarina del noreste argentino, por zona productora (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Método	Clasificadores	Zona productora
PCA	K	NEER
	Ca, Fe y Mg	SECR
LDA	Cr, Cu, K y Mg	NEER
	Ca, Cr, Fe, K y Zn	SECR
DT	Cr y Mn	COCR
	Al, Cr, Fe, K y Mn	CSMN
	Cr, K y Mn	NEER
	Al, Cr y Mn	SECR
ANN	Cu y Mg	COCR
	Ca, Cu, K y Sr	CSMN
	Cr y Mg	NEER
	Ca, Cr, Mn, Sr y Zn	SECR

Dentro de los elementos definidos como marcadores químicos de identidad presentes en los jugos de mandarina analizados, Al, Ca, Mg, Mn, y Sr concuerdan con los hallados por García Ruiz *et al.* (2007) en la clasificación de cidras de acuerdo con su origen, mediante el uso de LDA. Zhang *et al.* (2018) definieron algunos elementos como buenos marcadores para identificar el origen geográfico de muestras de té, utilizando PCA y LDA, entre ellos, Cr y K, coinciden con los definidos en este Capítulo para clasificar jugos de mandarina. Veljkovic *et al.* (2016) emplearon las concentraciones de 14 elementos para separar cuatro grupos de té, mediante el uso de PCA y Análisis Cluster, entre ellos Al, Ca, Cr, Cu, Mg, Mn, Sr y Zn, que concuerdan con los encontrados para jugos de mandarina. Van der Linde (2008) clasificó vinos por su origen utilizando los contenidos de 4 elementos entre ellos, Cu y Mn coinciden con los resultados de este Capítulo como buenos marcadores de identidad para jugos de mandarina.

5.4. Resumen de resultados

Mediante la técnica de MP-AES se pudieron estudiar en los jugos de mandarina las concentraciones de Al, Ca, Cd, Cr, Cu, Fe, K, Mn, Mg, Sr y Zn; las concentraciones de Ba, Co, Mo, Ni, Pb, Se y Sb se encontraban por debajo de los límites de detección (LOD) en todas las muestras.

Teniendo en cuenta la concentración media de todas las muestras, el elemento más abundante fue K con contenidos medios cercanos a los 1000 μg/g, seguido de Mg con contenidos medios superiores a 100 μg/g. En orden de importancia por sus contenidos medios le sigue Al, con valores medios ubicados entre 44 y 130 μg/g, pero valores máximos superiores a los del Mg; Ca presentó valores medios de 15 μg/g; los contenidos de Mn, Cu, Zn y Sr se encontraron entre 1 y 5 μg/g; y Cd, Cr y Fe, con contenidos medios inferiores a 1 μg/g.

Mediante PCA se ha podido establecer que las muestras del nordeste de Entre Ríos se diferencian de las correspondientes a Corrientes (sudeste y centro oeste) y Misiones, por sus menores contenidos de K, la zona de sureste de Corrientes se asocia a mayores contenidos de Ca, Fe y Mg y las del centro sur de Misiones y centro oeste de Corrientes presentan perfiles muy similares y no logran separarse.

El LDA permite clasificar los jugos de mandarina con un acierto de 86% e índice κ = 0,80. Las funciones discriminantes se construyeron, principalmente, con los contenidos de Cr, Cu, K y Mg (primera función) y de Ca, Cr, Fe, K y Zn (segunda función). El método no tiene el mismo comportamiento para clasificar muestras de todas las zonas, en el caso de COCR la sensibilidad fue del 66% muy por debajo de las demás, lo que indica que no sería adecuado su uso para clasificar estas muestras.

Con DT se obtuvo un acierto general de 91% y un índice κ global de 0,88. la separación de los jugos de mandarina procedentes del centro oeste de Corrientes fue la que menores porcentajes de acierto obtuvo con valores entre 62,4 y 76,1 %, debido a que muestras de centro oeste de Corrientes fueron clasificadas como procedentes de

centro sur de Misiones, afectando la sensibilidad de dicha zona. El Zn, no resultó importante en la clasificación por zona de los jugos de mandarina, Fe y Mg tienen baja participación (menos del 25%) y los demás elementos estudiados presentaron porcentajes mayores al 50.

Entre los Métodos de Clasificación por Vecindad (KNN), los mejores resultados se obtuvieron con 5 vecinos más cercanos, no obstante, los valores obtenidos de porcentaje de acierto e índice κ, son relativamente bajos, lo que indica que la técnica de K-Vecinos más Cercanos no sería la mejor opción para clasificar jugos de mandarina en función de su contenido mineral.

Al aplicar ANN se logró un acierto del 96% y un índice κ = 0,94, la sensibilidad y especificidad en todas las zonas superaron el 85%. Dentro de los elementos con mayor importancia en la discriminación, los que se encuentran presentes en al menos dos zonas productivas son Ca, Cr, Cu, Mg y Sr.

Las SVM permitieron clasificar los jugos de mandarinas con un acierto general de 94% y un índice κ de 0,92 la sensibilidad y especificidad en todas las zonas superaron el 85%.

Teniendo en cuenta los valores de porcentaje de acierto y de índice Kappa, se puede establecer para las técnicas de mejor comportamiento en la clasificación de jugos de mandarina el siguiente orden: ANN 96% > SVM 94% > DT 91% > LDA 86% > KNN 85%.

Los elementos considerados marcadores químicos de identidad para determinar el origen geográfico de jugos de mandarina del noreste argentino son: Al, Ca, Cr, Cu, Fe, K, Mg, Mn, Sr y Zn.

5.5. Referencias

Abu Mostafa, YS; Magdon Ismail, M; Lin HT. 2012. Learning from Data: A Short Course: AMLBook.com.

Astuti, W; Dewanto, S; Soebandrija, KEN; Tan, S. 2018. Automatic fruit classification using support vector machines: a comparison with artificial neural networks. The 2nd International Conference on Eco Engineering Development (ICEED 2018). IOP Conference Series: Earth and Environmental Science 195 012047.

Batista, L; da Silva, L; Rocha, B; Rodríguez, J; Berretta-Silva, A; Bonates, T; Gomes, V; Barbosa, R; Barbosa, F. 2012. Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. Food Research International. 49: 209-215.

Benabdelkamel, H; Di Donna, L; Mazzotti, F; Naccarato, A; Sindona, G; Tagarelli, A; Taverna, D. 2012. Authenticity of PGI “Clementine of Calabria” by multielemento fingerprint. Journal of Agricultural and Food Chemistry. 60: 3717-1726

Cheajesadagul, P; Arnaudguilhem, C; Shiowatana, J; Siripinyanond, A; Szpunar, J. 2013. Discrimination of geographical origin of rice based on multi-element fingerprinting by high resolution inductively coupled plasma mass spectrometry. Food Chemistry. 141: 3504-3509.

Chung, Y; Kim, J; Lee, J; Kim, S. 2015. Discrimination of geographical origin of rice (Oryza sativa L.) by multielement analysis using inductively coupled plasma atomic emission spectroscopy and multivariate analysis. Journal of Cereal Science. 65: 252-259.

Di Bella, G; Lo Turco, V; Potortí, A; Bua, G; Fede, M; Dugo, G. 2015. Geographical discrimination of Italian honey by multi-element analysis with a chemometric approach. Jouurnal of Food Composition and Analysis. 44: 25-35.

Drivelos, S. A & Georgiou, C. A. 2012. Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. TrAC Trends in Analytical Chemistry. 40:38-51.

Dutra, S; Adami, L; Marcon, A; Carnieli, G; Roani, C; Spinelli, F; Leonardelli, S; Vanderlinde, R. 2013. Characterization of wines according to the geographical origin by analysis of isotopes and minerals and the influence of harvest on the isotope values. Food Chemistry. 141: 2148-2153.

FAO. 2017. Citrus fruits statistics 2017. Disponible en línea: http://www.fao.org/economic/est/est-commodities/citricos/es/. Visita: 04/09/2019.

Federcitrus. 2018. La actividad citrícola Argentina. Disponible en línea: https://www.federcitrus.org/. Visita: 04/09/2019.

Gaiad, JE; Hidalgo, MJ; Villafañe, RN; Marchevsky, EJ; Pellerano, RG. 2016. Tracing the geographical origin of Argentinean lemon juices based on trace element profiles using advanced chemometric techniques. Microchemical Journal. 129: 243-248.

García Ruiz, S; Modovan, M; Fortunato, G; Wunderli, S; García Alonso, JI. 2007. Evaluation of strontium isotope abundance ratios in combination with multi-elemental analysis as a possible tool to study the geographical origin of ciders. Analytica Chimica Acta. 590: 55-66.

Geana, Y; Iordache, A; Ionete, R; Marinescu, A; Ranca, A; Culea, M. 2013. Geographical origin identification of Romanian wines by ICP-MS elemental analysis. Food Chemistry. 138: 1125-1134.

Gonzálvez, A; de la Guardia, M. 2013. Basic Chemometric Tools. In: de la Guardia, M; Gonzálvez, A. Comprehensive Analytical Chemistry. 60 (12): 299-315.

Hartling, L; Hamm, M; Milne, A; VandermeerB; Santaguida, L; Ansari, M; Tsertsvadze, A; Hempel, S; Shekelle, P; Dryden, DM. 2012. Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments [Internet]. Rockville (MD):

Agency for Healthcare Research and Quality (US); 2012 Mar. Table 2, Interpretation of Fleiss' kappa (κ) (from Landis and Koch 1977). Disponible en línea: https://www.ncbi.nlm.nih.gov/books/NBK92295/table/methods.t2/. Visita: 05/04/2020.

Lantz, B. 2015. Machine Learning with R: Packt Publishing.

Li, H; Lee, WS; Wang, K. 2014. Identifying blueberry fruit on different growth stages using natural outdoor color images. Computers and Electronics in Agriculture. 106: 91-101.

Luykx, D & Van Ruth S. 2008. An overview of analytical methods for determining the geographical origin of food products. Food Chemistry.107:897-911.

Magdas, D; Dehelean, A; Feher, Y; Cristea, G; Puscas, R; Dan, S; Cordea, D. 2016. Discrimination markers for the geographical and species origin of raw milk within Romania. International Dairy Journal. 61: 135-141.

Maione, C; Batista, B; Campiglia, A; Barbosa, F; Barbosa, R. 2016. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. Computers and Electronics in Agriculture. 121: 101-107.

Maione, C; Silva de Paula, E; Gallimberti, M; Batista, B; Campiglia, A; Barbosa, F: Barbosa, R. 2016. Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. Expert Systems with Applications. 49: 60-73.

Marcelo, MCA; Martins, CA; Pozebon, D; Dressler, VL; Ferro, MF. 2014. Classification of yerba mate (Ilex paraguariensis) according to the country of origin based on element concentrations. Microchemical Journal. 117: 164-171.

Moncayo, S; Manzoor, S; Navarro Villoslada, F; Cáceres, JO. 2015. Evaluation of supervised chemometric methods for sample classification by Laser Induced Breakdown Spectroscopy. Chemometrics and Intelligent Laboratory Systems. 146: 354-364.

Palacios, J. 2013. Citricultura. Talleres Gráficos Alfa Beta S.A. ISBN: 9789874383266. 518 pp.

SENASA. 2014. Escenarios y Tendencias. Informe Estadístico Cítricos Argentinos de Excelencia. SENASA. Ministerio de Agricultura, Ganadería y Pesca, Presidencia de la Nación.

Simpkins, WA; Louie, H; Wu, M; Harrison, M; Goldberg, D. 2000. Trace elements in Australian orange juice and other products. Food Chemistry. 71(4): 423-433.

Szymczycha Madeja, A; Welna, M. 2013. Evaluation of a simple and fast method for the multi-elemental analysis in commercial fruit juice samples using atomic emission spectrometry. Food chemistry. 141(4): 3466-3472.

Szymczycha Madeja, A; Welna, M; Jedryczko, D; Pohl, P. 2014. Developments and strategies in the spectrochemical elemental analysis of fruit juices. TrAC Trends in Analytical Chemistry. 55: 68-80.

Takaya, S; Rehmsmeier, M. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Plos One. Disponible en línea: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432. https://doi.org/10.1371/journal.pone.0118432. Visita: 02/05/2020.

Van der Linde, G. 2008. Multi-element analysis of South African wines un their provenance soils by ICP-MS and their classification according to geographical origin using multivariate Statistics. Tesis Magister Scientiae of Chemistry in the Faculty of Science at the University of Johannesburg.

Veljkovic, JN; Pavlovic, AN; Brcanovic, JM; Mitic, SS, Tosik, SB; Pecev-Marinkovic, ET; Mitic, MN. 2016. Differentiation of black, green, herbal and fruit bagged teas based on multi-element analysis using inductively coupled plasma atomic emission spectrometry. Chemical papers. 70 (4): 488-494.

Warrens, MJ. 2020. Kappa coefficients for dichotomous-nominal classifications. Advances in Data Analysis and Classification. Disponible en línea:

https://link.springer.com/content/pdf/10.1007/s11634-020-00394-8.pdf. Visita: 20/05/2020.

Weil, B; Brady, R. 2017. The nature and properties of soils. Pearson Education. 15th edition.

Zhang, J; Yang, R; Chen, R; Li, YC; Peng, Y; Liu, C. 2018. Multielemental analysis associated with chemometric techniques for geographical origin discrimination of tea lives (*Camelia sinensis*) in Ghizou Province, SW China. Molecules. 23: 1-16.

CAPÍTULO VI

COMPOSICIÓN MINERAL Y MODELOS PARA DETERMINAR IDENTIDAD DE JUGOS DE NARANJAS PRODUCIDAS EN LA REGIÓN NORDESTE ARGENTINA

6.1. Introducción

Se conoce como naranjas a los frutos del naranjo dulce (*Citrus sinensis* L.), que es el cítrico más cultivado en el mundo. Estas frutas son de tamaño mediano, con o sin semillas, jugo agridulce con 0,8 a 1% de acidez, hipocalóricas e hiposódicas compuestas principalmente por agua. Su contenido de grasa, proteínas y fibra es muy bajo, los hidratos de carbono son el segundo componente de mayor presencia. Asimismo, se caracterizan por su alto contenido de vitamina C y aportes de vitaminas B_1, B_2 y provitamina A. En general, por su época de maduración, a las variedades de naranja se las clasifica en tempranas, intermedias y tardías, de ello dependerá el momento de cosecha (Agustí, 2010; Palacios, 2013).

La actividad citrícola en la Argentina se destaca por su producción de limones, naranjas, mandarinas y, en menor medida, pomelos. La fruta producida permite abastecer la demanda del país durante todo el año y es muy oportuna para las exportaciones en contra estación a los países del hemisferio norte. Estos mercados internacionales atienden a consumidores más exigentes, por lo que la certificación de origen, que asegura la autenticidad de un determinado producto alimentario, mejora las condiciones de comercialización (Drivelos & Georgiou, 2012; Luykx & van Ruth, 2008).

La producción argentina para el año 2017 fue de 3.272.771 t y representa el 2,23% de la producción mundial de cítricos, el 31,22% corresponde a naranjas, el

51,22% a limones y limas, el 14,02% a mandarinas, tangerinas, clementinas y satsumas, el 3,42% a pomelos y toronjas (Federcitrus, 2018; FAO, 2017).

Dadas las exigencias de los mercados de exportación, en nuestro país se ha implementado el sistema registral SITC® con el fin de lograr la trazabilidad de los cítricos producidos. El mismo se basa en información documental y no contempla mecanismos que permitan comprobar la identidad física de las muestras en cualquier etapa de la cadena productiva, por lo que éstas pueden ser vulnerables a contaminación o adulteración. Resolver este problema implica la necesidad de contar con un mecanismo de identificación de las muestras físicas, que podría estar basado en la composición química de las mismas.

Si bien la composición mineral de los vegetales obedece a patrones generales definidos para las diferentes especies y variedades, existe cierto grado de variabilidad, que se debe en gran medida, a condiciones de los sitios donde las plantas crecen. El suelo es uno de los elementos clave en todos los ecosistemas terrestres y es el factor principal que controla el flujo del agua en el ciclo hidrológico, así como de las especies químicas dentro de los ciclos biogeoquímicos (Weil & Brady, 2017). Esta influencia de las condiciones locales en la composición de los tejidos vegetales ha permitido que la determinación de los contenidos de elementos a nivel de vestigio haya sido propuesta para determinar el origen geográfico de las muestras.

El término técnicas de huella dactilar describe a una variedad de métodos analíticos que pueden medir la composición de productos alimentarios de una manera no selectiva, esto es, mediante la colección de un espectro, cromatograma o datos de composición multielemental. Los métodos que proporcionan un perfil mineral característico se pueden usar para la determinación del origen geográfico y, por lo tanto, generar una herramienta valiosa para la autenticación y trazabilidad de alimentos (Esslinger *et al.*, 2014).

La espectroscopia de absorción atómica por llama (FAAS), es una técnica sensible y especifica debido a que las líneas de absorción atómica son considerablemente estrechas (de 0,002 a 0,005 nm) y las energías de transición electrónica son únicas para

cada elemento. La sensibilidad de la técnica permite identificar pequeñas concentraciones de cada metal, desde mg/L a µg/L. Es una técnica ampliamente difundida, accesible y de relativamente bajo costo, no obstante, su sensibilidad solamente permite determinar elementos presentes a nivel de macro u oligoelementos (Skoog *et al.*, 2008).

Los espectrómetros de emisión atómica de plasma por microondas (MP-AES) ofrecen alta sensibilidad, límites de detección inferiores a las partes por billón (ppb), velocidad superior a la de la absorción atómica de llama y no necesita gases combustibles. La nueva generación de espectrómetros MP-AES funciona con aire, lo que reduce enormemente los costos operativos.

El objetivo de este capítulo fue la utilización de información química de contenido de elementos minerales en muestras de jugo de naranja de las variedades 'Salustiana' y 'Valencia late', para obtener modelos de clasificación adecuados a efectos de analizar orígenes de jugo de naranja de diferentes zonas productoras del NEA, mediante información proporcionada por las dos técnicas de análisis químico (FAAS y MP-AES).

6.2. Materiales y Métodos

6.2.1. Obtención y procesamiento de muestras

Mediante un muestreo en etapas sobre un padrón de productores con características tecnológicas medias, en cuatro zonas productivas del NEA (centro sur de Misiones: CSMN, centro oeste de Corrientes: COCR, sudeste de Corrientes: SECR y noreste de Entre Ríos: NEER), por cada variedad estudiada se seleccionaron 5 lotes al azar simple, en cada lote 5 plantas por azar sistemático ubicadas a lo largo de una línea que recorría el lote de NO a SE, en las que se extrajeron 10 frutos por planta de los que se obtuvo el jugo (unidad muestral), se realizó la digestión por vía seca y la determinación multielemental. Para la determinación del contenido multielemental

mediante espectroscopía de absorción atómica por llama, se trabajó con 120 muestras de naranjo dulce (60 muestras de cada variedad), obtenidas durante las dos primeras campañas siguiendo el mismo procedimiento.

En este Capítulo se evalúa y compara la eficacia de la información obtenida mediante espectroscopía de absorción atómica por llama (FAAS) y por espectroscopía de emisión atómica de plasma de microondas (MP-AES), para detectar la presencia de elementos que permitan conocer el origen de las muestras de jugos de naranja de diferentes zonas de la región NEA.

6.2.2. Datos

El conjunto de datos con el que se trabajó estuvo formado por una variable que identificaba los sitios de origen de las muestras (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) y las variables clasificadoras que corresponden a los contenidos de elementos minerales en jugos de naranja, que se encontraron por encima de los límites de detección con diferentes instrumentos de análisis químico: 7 variables clasificadoras en el análisis mediante FAAS y 16 en el análisis mediante MP-AES.

La técnica FAAS se utilizó debido a su amplia difusión, su accesibilidad en diferentes regiones del país y su relativo bajo costo, que permitió obtener información de contenidos de Ca, K, Fe, Mg, Mn, Na y Zn. La técnica de MP-AES se empleó por ser más sensible y específica para elementos traza, pero de menor difusión en nuestro país, más difícil acceso y mayores costos, se pudieron determinar en los jugos de naranja los siguientes elementos: Al, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Mo, Ni, Sb, Sr y Zn.

6.2.3. Análisis de datos

En primer lugar, se realizó un Análisis Exploratorio de Datos empleando herramienta gráficas y técnicas de Estadística Descriptiva, Análisis de Varianza (ANOVA) para cada uno de los elementos, con las correspondientes pruebas F combinadas con la prueba de Duncan para separación de promedios (Perelman *et al*., 2019).

Ante la necesidad de estudio simultáneo de varias variables, se utilizaron Métodos de Análisis Estadísticos Multivariantes y de Aprendizaje Automático. Dentro de los Métodos Multivariantes se aplicaron técnicas de Análisis de Componentes Principales (PCA) y Análisis Discriminante Lineal (LDA) (Peña, 2002). Entre las técnicas de Aprendizaje Automático se utilizaron Árboles de Decisión (DT), empleando los algoritmos C5.0, Rpart y Rpart2; K-vecino más cercano (KNN), buscando la mejor configuración; Redes Neuronales Artificiales (ANN), con datos normalizados, utilizando el algoritmo MLP, probando diferente cantidad de capas y número de neuronas por capa, seleccionando la cantidad de iteraciones en que se estabiliza la precisión; y Máquinas de Vectores Soporte (SVM) (James *et al.*, 2017).

6.3. Resultados y Discusión

Con la información de contenido de los diferentes elementos minerales en los jugos de naranja se realizó, en primer término, una caracterización de las variedades estudiadas y luego un análisis de los contenidos minerales de los jugos en las diferentes zonas a fin de establecer una diferenciación geográfica.

6.3.1. Caracterización multielemental

Los resultados de las concentraciones totales de los diferentes elementos se presentan en la Tabla 6.1. Las concentraciones de dos elementos (Pb y Se) no se presentan porque se encontraban por debajo de los límites de detección (LOD) en todas las muestras.

Tabla 6.1. Concentraciones de elementos en muestras de jugo de naranja (μg/g). Promedios (Me), desviaciones estándares (DE), mínimos (Mín) y máximos (Max)

	Variedad							
	SAL				VAL			
Elemento	Me	DE	Mín	Máx	Me	DE	Mín	Máx
Al	9,5	8,31	1,35	12,52	7,6	5,76	1,33	22,83
Ba	1,88	4,83	0	18,76	5,66	7,56	0	20,32
Ca	17,8	10,88	8,66	58,76	20,39	17,19	8,86	84,94
Cd	0,6	0,72	0	0,8	0,5	0,54	0	0,8
Co	0,01	0,01	0	0,03	0,01	0,02	0	0,04
Cr	0,32	0,47	0	1,78	0,75	1,16	0	4,27
Cu	5,32	3,06	1,7	16,6	7,88	5,55	1,6	17,56
Fe	0,46	0,32	0,11	0,98	0,28	0,18	0,08	0,78
K	1365,94	534,58	366,58	2478,56	1202,37	576,96	192,76	2288,72
Mg	125,37	59,54	61,5	281,77	121,93	54,05	63,12	283,1
Mn	1,33	1,54	0,15	5,92	2,55	2,4	0,35	8,73
Mo	0,0046	0,01	0	0,03	0,02	0,05	0	0,19
Ni	0,03	0,13	0	0,7	0,1	0,27	0	0,98
Sb	0,25	0,78	0	3,29	1,54	2,24	0	4,97
Sr	2,18	4,16	0	15,84	3,41	3,45	0,15	10,97
Zn	1,83	2,23	0,15	8,5	1,68	1,8	0,13	5,9

Teniendo en cuenta la concentración media de todas las muestras, el elemento más abundante fue K con contenidos medios superiores a los 1000 μg/g, seguido de Mg con contenidos medios superiores a 100 μg/g. En orden de importancia por sus contenidos medios le siguen Al y Ca, con valores medios ubicados entre 10 y 20 μg/g; Ba, Cu, Mn, Sr y Zn, entre 1 y 10 μg/g; y Cd, Co, Cr, Fe, Ni, Mo y Sb, con contenidos medios inferiores a 1 μg/g. Estos valores concuerdan con los encontrados por Hong *et al.* (2019), quienes establecieron que los cítricos eran fuentes adecuadas de K (95,13 – 270,4 μg/g), Ca (10,57 – 75,29 μg/g), Zn (0,47 – 1,61 μg/g) y Mn (0,035 – 1,90 μg/g).

Los niveles de concentración encontrados para K, Fe, Mg, Sb, Co, Ni y Mo son los similares a los obtenidos en otros estudios en naranja (USDA, 2020; Tufour *et al.*, 2011). Los valores de K, Ca y Mg son coincidentes con los hallados por Chuku y Chinaka (2014) en jugos de naranja. En relación con los valores reportados Simpkins *et al*. (2000) y Niu *et al.* (2008), en esta tesis se encontraron, en los jugos de naranja, niveles similares de K, Fe, Mg, Ni y Mo, niveles inferiores de Ca y niveles superiores de Zn, Cu, Al, Mn, Ba y Co. En comparación con los resultados encontrados por Harmankaya *et al.* (2011) y Farid y Enani (2010), en los jugos de naranja analizados en esta tesis se encontraron concentraciones similares de Fe, Co, Ni y Mo, superiores de Zn, Cd, Cu, Mn y Mg e inferiores de Ca. De acuerdo con los valores reportados por Cautela *et al.* (2009), los niveles encontrados en esta tesis para K son coincidentes, los de Zn, Cu, Al, Mn, Sr, Cr, Fe, Mg, Ba, Sb, Co, Ni y Mo son mayores y los de Ca son menores. En relación con los valores obtenidos por Turra *et al*. (2017), los contenidos de Ba, Mn y Sr de las muestras estudiadas presentaron valores similares, los de Al y Cu fueron algo superiores y los de Fe, Mg y Zn inferiores. La concentración de Cu encontrada en esta tesis fue superior al rango de concentración notificado por Raja *et al*. (2016) para cítricos de Irán.

La Junta de Alimentación y Nutrición de Estados Unidos, ha especificado algunos niveles críticos para los denominados macroelementos, llamados Ingestas de Referencia Dietética (DRIs) e Ingestas Superiores Tolerables (TUIs). Los valores DRI (mg/kg) para los macroelementos estudiados son: 1000-1200 (Ca), 200-2300 (K), 350 (Mg), y 6-8 (Fe), mientras que los TUI (mg/kg) son: 2500 (Ca), 4700 (K), 350 (Mg), y 45 (Fe) (FNB, 2005). Observando estos valores, los resultados obtenidos, en jugos de naranja en la presente tesis, se encuentran dentro de los límites recomendados (Khan *et al*., 2014, Belitz *et al*., 2009).

Los microelementos son aquellos que se presentan con una concentración media de más de 10 µg/g en los alimentos. Entre los elementos estudiados en esta tesis se incluyen seis de los denominados microelementos Zn, Cu, Mn, Ni, Cr y Ba, aunque algunos en concentraciones mucho menores. Entre estos elementos, las

concentraciones de Cu (5-18 μg/g), Zn (2-8 μg/g), Ba (2-20 μg/g) y Mn (1-9 μg/g) fueron las más altas, seguido de Cr (0,3-4 μg/g) y Ni (0,1-1 μg/g).

Diversos son los factores que pueden contribuir a variaciones en los niveles de elementos minerales en los jugos de naranja según su origen geográfico. La disponibilidad del elemento en el suelo para su extracción por la planta es el factor de mayor incidencia. No obstante, otros factores como la madurez de los frutos en la cosecha y diversas prácticas agrícolas como las aplicaciones de agroquímicos o el riego pueden influir en las concentraciones de elementos minerales en los jugos de frutas (Szymczycha Madeja *et al.*, 2014).

Si bien se observan perfiles de concentraciones de elementos minerales semejantes entre ambas variedades, el MANOVA ha detectado diferencias significativas entre ellas ($F = 67,19$, p-valor < 0,0001 para 'Salustiana' y $F = 23,96$, p-valor = 0,0001 para 'Valencia late') y la prueba de Hotelling corregida por Bonferroni separa, en ambas variedades, los jugos procedentes de SECR de las demás regiones. A fin de detectar los elementos que permiten hacer esta diferenciación, en la Tabla 6.2 se presentan los valores de F posteriores al análisis de variancia para cada elemento.

Tabla 6.2. Valores de F para las hipótesis de igualdad de concentraciones por elemento mineral en jugos de naranja de las variedades 'Salustiana' (SAL) y 'Valencia late' (VAL) y sus correspondientes p-valores

Elemento	F	p-valor
Al	3,41	0,0709
Ba	4,54	0,0384*
Ca	0,41	0,5226
Cd	0,15	0,7024
Co	1,44	0,2364
Cr	3,05	0,0871
Cu	4,25	0,0447*
Fe	5,27	0,0262*
K	1,05	0,3108
Mg	0,04	0,8361
Mn	4,61	0,0369*
Mo	2,5	0,1209
Ni	1,38	0,2457
Sb	8,01	0,0068*
Sr	1,22	0,2754
Zn	0,07	0,7996

(*) indica significancia estadística

Estos resultados permiten afirmar que solamente son significativamente diferentes entre ambas variedades, los contenidos de Cu, Fe, Ba y Sb por lo que, en un sistema de trazabilidad en el que se requiera diferenciar entre estas dos variedades de naranja empleando análisis de la variancia, deberían incluirse estos elementos en los análisis.

6.3.2. Elementos de interés ambiental, toxicológico y nutricional

Los metales pesados son elementos químicos con alta densidad, masa y peso atómico, algunos de ellos son: Al, Ba, Be, Co, Cu, Sn, Fe, Mn, Cd, Hg, Pb, As, Cr, Mo, Ni, Ag, Se, Tl, V, Au y Zn (Shimada, 2005; Concon, 2009). Los seres vivos requieren pequeñas cantidades de algunos de estos elementos para varias funciones biológicas, por ejemplo, Co, Cu, Fe, Mn, Se y Zn, pero una excesiva concentración puede alterar procesos bioquímicos y/o fisiológicos en el organismo, por lo que resultan tóxicos. (OMS, 1980; Molina *et al*., 2013; Hernández y Hansen, 2012).

Estos elementos son absorbidos por las plantas y se encuentran presentes en los alimentos a nivel de vestigios o ultra traza, dependiendo de su disponibilidad en el suelo y de los mecanismos de selectividad propios de cada especie o variedad. A fin de detectar posibles problemas de interés ambiental, toxicológico o nutricional, se estudiaron los niveles de dichos elementos, comparativamente con valores de referencia.

Teniendo en cuenta los valores máximos establecidos en el Código Alimentario Argentino, en las muestras analizadas en esta tesis y en relación con los niveles permitidos para jugos cítricos se observa que los contenidos promedios de Cu de la variedad 'Salustiana' se encontraron dentro de los estándares, pero los de la variedad 'Valencia late' superan esos máximos permitidos. No obstante, en 25 de las muestras analizadas se encontraron valores mayores a los permitidos de Cu, 13 de la variedad 'Salustiana' (4 de sudeste de Corrientes, 3 de centro oeste de Corrientes, 3 de noreste de Entre Ríos y 3 de centro sur de Misiones) y 12 de la variedad 'Valencia late' (8 de sudeste de Corrientes, 1 de noreste de Entre Ríos y 3 de centro sur de Misiones).

La Association of American Feed Control Oficials (AAFCO, 2020) clasifica los metales en altamente tóxicos, tóxicos, moderadamente tóxicos y ligeramente tóxicos y propone valores máximos recomendables, que se presentan en la Tabla 6.3.

Tabla 6.3. Categorización de metales pesados y valores máximos permitidos en alimentos (Association of American Food Control Oficials, AAFCO)

Categoría	Nivel máximo	Metal
Altamente tóxico	10	Cd, Hg, Se
Tóxico	40	Ba, Co, Cu, Pb, Mo, V, W
Moderadamente tóxico	400	As, I, Ni, Sb
Levemente tóxico	1000	Al, Bi, Bo, Br, Cr, Mn, Zn

Los niveles de todos los elementos determinados en las muestras de jugo de naranja analizadas en esta tesis se encuentran por debajo de los valores máximos establecidos por la AAFCO, si bien se considera que esos valores máximos son bastante elevados en comparación con otros estándares.

El Al es reportado en la literatura como neurotóxico y se supone que participa en la fisiopatología de trastornos neurodegenerativos como la enfermedad de Alzheimer

y el Parkinson (Mutsuko *et al*., 2011). El Comité de Expertos en Aditivos Alimentarios de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) y la Organización Mundial de la Salud (OMS) (JECFA, 2006) ha definido una ingesta semanal tolerable provisional (PTWI) de 1 mg/kg de peso corporal, para Al de todas las fuentes. En los jugos de naranja analizados en esta tesis, los valores de las concentraciones de Al en los jugos de naranja se encontraron a niveles inferiores a los 15 µg/g en todas las muestras, por lo que se puede afirmar que el contenido de Al de los jugos de naranja está por debajo del PTWI especificado.

Entre los denominados microelementos (por encontrarse en pequeñas concentraciones), la Organización Mundial de la Salud (OMS) ha categorizado a Cu, Cr y Zn, como esenciales, mientras que Ni y Mn son considerados elementos esenciales probables (OMS, 1996). El Comité de Expertos en Aditivos Alimentarios de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) y la Organización Mundial de la Salud (OMS) (JECFA, 2006) ha reportado algunos niveles críticos importantes, como nivel de ingesta diaria máxima tolerable provisional (PMTDI), nivel PTWI, referencias de ingesta dietética (DRI), ingesta dietética total (TDI) y un nivel superior de ingesta regional (UL). Según estos valores, los DRIs para Zn, Cu, y Cr son 8-11, 0,7-0,9, y 0,015-0,030 mg/día respectivamente y el UL especificado para Zn, Mn, Ni y Cu es de 40, 11, 1 y 10 mg/día respectivamente, para adultos de entre 30 y 70 años. En los jugos estudiados en esta tesis se encontró el Zn en valores aceptables, mientras que Cu y Cr presentan valores algo elevados.

Entre los elementos traza se incluyen aquellos que tienen valores de concentración promedio inferior a 10 g/kg. Sobre la base de su naturaleza, estos se agrupan en no tóxicos y tóxicos, como se explica para los productos lácteos por Llorent Martínez *et al.* (2012) y Khan *et al.* (2014b) y para las especias de Khan *et al.* (2014c).

Entre los elementos traza estudiados en esta tesis, el Co se considera como oligoelemento no tóxico porque no se encontró ningún efecto tóxico en la literatura y su rango de concentración fue entre 0,01 y 0,04. El Cd, por el contrario, presenta toxicidad para los seres humanos y otros mamíferos y los PTWI que informan la Organización Mundial de la Salud y la Junta de Alimentación y Nutrición de EE. UU.

(JECFA, 2006), para 15 g/kg de peso corporal para Cd son de 7 g/kg. La FAO y la OMS publican el Codex Alimentarius (FAO-OMS, 2020) que establece, para los frutos o jugos de frutos, los contenidos máximos de As en 0,2 mg/kg, Cd en 0,05 mg/kg y Pb en 0,1 mg/kg. Entre estos elementos, en esta tesis solamente se determinó el contenido de Cd, cuyo valor promedio (0,60 mg/kg) se encuentra muy por encima del máximo permitido si se tienen en cuenta estos estándares.

Hong *et al*. (2019) establecieron que las concentraciones de elementos tóxicos (Al, As, Cd, Hg, Pb) en jugos cítricos eran muy bajos, estos resultados coinciden con las concentraciones halladas en esta tesis en jugos de naranja para el Al, aunque los citados autores reportaron valores superiores de concentraciones de Cd.

6.3.3. Análisis Exploratorio

Con el objeto de explorar la presencia posibles diferencias o asociaciones en la composición mineral de los jugos de naranja entre las diferentes zonas de producción, se realizó en primer término un PCA con las muestras agrupadas de acuerdo a la procedencia geográfica de las mismas. En base a los resultados, se pudo reducir la dimensionalidad de la matriz de datos a las dos primeras componentes principales, resumiendo un 83,6% de la variabilidad total (49,2% en la primera componente y 34,4% en la segunda).

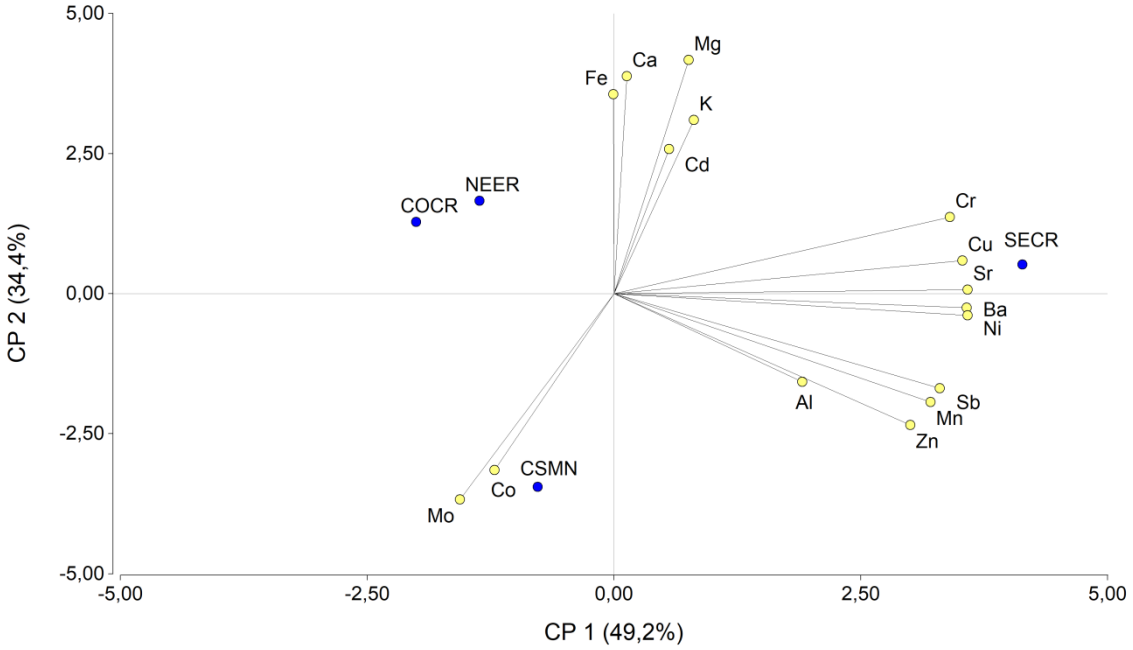

Figura 6.1. Biplot que representa los contenidos de elementos minerales en jugo de naranja y las zonas productoras (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) sobre las dos primeras coordenadas resultantes del Análisis de Componentes Principales

Según se observa en el gráfico biplot, las dos primeras Componentes Principales permiten separar a las muestras de jugo de naranja provenientes del sudeste de Corrientes, por un lado, que se encuentran asociados a mayores contenidos de Ba, Cu, Cr, Mn, Ni, Sb, Sr y Zn, respecto de las otras zonas productivas. Por otro lado, a su vez los jugos provenientes del centro sur de Misiones se pueden diferenciar de los del centro-oeste de Corrientes y noreste de Entre Ríos, por sus asociaciones con mayores contenidos de Co y Mo. Mientras que los jugos del centro-oeste de Corrientes y noreste de Entre Ríos presentan características muy similares, asociados a mayores contenidos de Ca, Cd, Fe, K y Mg.

Paralelamente, las muestras se agruparon de acuerdo a su variedad botánica en dos grupos, Salustiana (SAL) y Valencia Late (VAL). Los resultados fueron analizados mediante análisis de la varianza univariante (ANOVA) y multivariante (MANOVA).

En la Tabla 6.4 se presentan los valores de los estadísticos F obtenidos con posterioridad a los Análisis de la Variancia y sus correspondientes p-valores.

Tabla 6.4. Estadísticos F y sus p-valores, correspondientes a los ANOVA por elemento y variedad entre zonas productoras

	SAL		VAL	
Elemento	F	p-valor	F	p-valor
Al	5,58	0,0047*	2,25	0,1191
K	2,04	0,1346	5	0,0115*
Mg	2,02	0,1382	1,93	0,1629
Mo	1,89	0,1575	1,69	0,2067
Ba	3,65	0,0268*	3,82	0,0294*
Ca	0,96	0,4286	1,94	0,1625
Cd	2,75	0,0645	2,17	0,1291
Co	0,78	0,5141	4,1	0,0232*
Cr	3,84	0,0224*	1,92	0,1653
Cu	0,56	0,6477	4,9	0,0124*
Fe	1,24	0,3159	1,75	0,194
Mn	25,67	<0,0001*	2,79	0,0721
Ni	0,55	0,6521	0,58	0,6364
Sb	2,33	0,0994	1,8	0,1857
Sr	7,63	0,0009*	10,93	0,0003*
Zn	0,69	0,5644	0,35	0,792

(*) indica significancia estadística

Se han detectado diferencias significativas entre zonas, en ambas variedades, en los contenidos de Sr y Ba; en 'Salustiana' en los contenidos de Al, Cr y Mn y en 'Valencia late' en los contenidos de Cu y K.

En la Tabla 6.5 se presentan la caracterización de las concentraciones de los diferentes elementos minerales en jugos de naranja por variedad y por zona estudiada.

Tabla 6.5. Concentraciones de elementos en muestras de jugo de naranja por variedad y zona productora (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR), Promedios (Me), desviaciones estándares (DE), mínimos (Mín) y máximos (Máx)

		Zona															
		COCR															
	Ele	Zn	Cd	Cu	Al	Mn	Sr	Cr	K	Fe	Mg	Ca	Ba	Sb	Co	Ni	Mo
SAL	Me	1,63	0,04	6,28	3,94	0,48	0,20	0,12	1150,10	0,66	149,57	21,28	0,18	0,00	0,017	0,00	0,01
	DE	2,12	0,07	2,93	2,12	0,19	0,13	0,26	308,63	0,31	71,83	18,44	0,45	0,00	0,004	0,00	0,02
	Mín	0,18	0,00	4,10	1,35	0,25	0,00	0,00	883,66	0,11	77,89	11,29	0,00	0,00	0,00	0,00	0,00
	Máx	5,44	0,13	11,80	7,41	0,71	0,39	0,64	1556,89	0,93	281,77	58,76	1,10	0,00	0,01	0,00	0,03
VAL	Me	0,43	0,74	3,40	6,45	0,38	0,18	0,36	988,10	0,34	117,27	16,62	0,00	0,00	0,00	0,00	0,00
	DE	0,08	0,07	0,01	0,11	0,02	0,04	0,19	11,46	0,03	0,28	0,29	0,00	0,00	0,00	0,00	0,00
	Mín	0,37	0,69	3,39	6,37	0,36	0,15	0,22	980,00	0,32	117,07	16,41	0,00	0,00	0,00	0,00	0,00
	Máx	0,48	0,79	3,41	6,53	0,39	0,20	0,49	996,20	0,36	117,46	16,82	0,00	0,00	0,00	0,00	0,00
		CSMN															
SAL	Me	1,90	0,44	4,19	8,42	0,42	0,22	0,07	1103,07	0,37	83,99	12,50	0,00	0,00	0,01	0,02	0,00
	DE	1,44	0,67	2,11	4,84	0,37	0,11	0,07	533,18	0,32	11,20	4,53	0,00	0,00	0,01	0,02	0,00
	Mín	0,50	0,00	1,70	3,12	0,15	0,08	0,00	366,58	0,11	61,50	8,66	0,00	0,00	0,00	0,00	0,00
	Máx	4,00	1,83	8,00	15,05	1,22	0,37	0,17	1805,71	0,98	96,23	22,14	0,00	0,00	0,03	0,03	0,00
VAL	Me	1,83	0,41	5,43	11,75	3,82	1,96	0,19	686,88	0,15	86,98	11,03	3,55	1,50	0,03	0,07	0,06
	DE	1,67	0,35	3,91	3,59	1,91	1,16	0,25	268,25	0,05	10,27	1,63	5,86	2,33	0,02	0,16	0,09
	Mín	0,25	0,00	1,60	7,38	2,2	0,90	0,02	192,76	0,08	76,93	8,86	0,00	0,00	0,00	0,00	0,00
	Máx	4,01	0,89	12,19	16,83	7,2	4,14	0,64	899,66	0,22	102,30	13,61	13,84	4,84	0,04	0,40	0,19
		NEER															
SAL	Me	0,87	0,78	0,51	24,33	0,36	0,21	0,27	1697,88	0,36	119,53	16,54	0,00	0,00	0,01	0,02	0,01
	DE	0,72	0,41	0,20	12,13	0,14	0,06	0,33	594,23	0,23	62,04	7,31	0,00	0,00	0,01	0,03	0,01
	Mín	0,28	0,27	0,25	11,64	0,19	0,15	0,00	1100,66	0,12	63,55	9,84	0,00	0,00	0,00	0,00	0,00
	Máx	1,81	1,26	0,75	40,73	0,54	0,32	0,90	2478,56	0,74	236,93	27,62	0,00	0,00	0,03	0,07	0,01
VAL	Me	1,60	1,05	0,43	31,29	0,53	0,32	0,30	1761,66	0,37	164,39	35,79	0,00	0,00	0,02	0,01	0,02
	DE	1,72	0,19	0,13	11,23	0,21	0,08	0,39	319,93	0,27	81,75	33,10	0,00	0,00	0,02	0,00	0,01
	Mín	0,28	0,79	0,33	22,75	0,35	0,21	0,05	1340,29	0,21	107,52	14,85	0,00	0,00	0,00	0,01	0,01
	Máx	4,00	1,20	0,60	46,49	0,83	0,39	0,88	2118,52	0,78	283,10	84,94	0,00	0,00	0,03	0,01	0,02
		SECR															
SAL	Me	2,56	6,29	14,54	10,64	3,26	6,33	0,70	1493,00	0,47	145,32	20,44	5,73	0,78	0,01	0,08	0,003
	DE	3,28	9,33	18,37	96,32	1,33	5,94	0,63	524,28	0,36	61,89	9,79	7,34	1,26	0,01	0,23	0,01
	Mín	0,15	0,00	0,25	1,30	1,72	1,50	0,04	779,45	0,16	63,74	9,61	0,00	0,00	0,00	0,00	0,00
	Máx	8,50	25,00	44,24	117,61	5,92	15,84	1,78	2050,52	0,97	214,33	34,41	18,76	3,29	0,03	0,70	0,01
VAL	Me	1,90	0,33	12,41	150,65	3,08	6,48	1,40	1345,08	0,32	127,40	20,62	10,83	2,59	0,002	0,19	0,002
	DE	2,19	0,66	8,53	161,30	2,71	3,05	1,55	598,95	0,18	53,62	11,76	7,99	2,48	0,004	0,39	0,004
	Mín	0,13	0,00	0,28	20,09	1,03	1,16	0,00	552,02	0,12	63,12	9,24	0,00	0,00	0,00	0,00	0,00
	Máx	5,90	1,50	27,08	199,19	8,73	10,97	4,27	2288,72	0,65	209,12	42,72	20,32	4,97	0,01	0,98	0,01

El MANOVA ha permitido determinar que existe un efecto significativo de la zona en la composición mineral de los jugos de naranja ($F = 3,68$, p-valor < 0,0001), además de una interacción significativa entre la variedad y la zona ($F = 2,55$, p-valor = 0,0001).

En la Tabla 6.6 se presentan los resultados de la prueba de Hotelling corregida por Bonferroni, donde se pueden observar las zonas entre las que se encuentran las diferencias.

Tabla 6.6. Resultados de la prueba de Hottelling-Bonferroni por Zona productora (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona		
SECR	A	
NEER		B
CSMN		B
COCR		B

Promedios seguidos de la misma letra no presentan diferencias significativas

La inclusión conjunta de todos los elementos ha permitido diferenciar los jugos provenientes del sureste de Corrientes, con mayores contenidos de Al, Ba, Cd, Cu, Cr, Mn, Ni, Sb, Sr y Zn, respecto de las demás zonas productoras. El MANOVA no ha permitido diferenciar entre los jugos de naranja provenientes de noreste de Entre Ríos, centro sur de misiones y centro oeste de Corrientes.

6.3.4. Propuesta de modelos clasificatorios

A continuación, se presentan diferentes análisis que generan modelos con el fin de clasificar el origen de los jugos con errores aceptables, a partir de contenidos de elementos minerales determinados mediante espectrometría de absorción atómica de llama (FAAS) y espectroscopía de emisión atómica de plasma de microondas (MP-AES).

Una vez seleccionados los valores óptimos para cada modelo, para la evaluación de la clasificación alcanzada con los diferentes métodos en el conjunto de pruebas, se tuvo en cuenta la sensibilidad (muestras pertenecientes a la clase y clasificadas correctamente en esta clase), la especificidad (muestras que no pertenecen a la clase modelada y clasificadas correctamente como no pertenecientes) y el porcentaje de acierto general (1 - porcentaje de error) (Marcelo *et al*., 2014).

Se utilizó también el índice Kappa (κ), un coeficiente propuesto originalmente por Cohen en 1960 que permite medir la concordancia entre los resultados de dos o más variables cualitativas. El índice κ, aplicado a la matriz de confusión permite evaluar si la clasificación observada es similar (concordante) con la clasificación predicha por el clasificador (Warrens, 2020).

6.3.4.1. Selección de modelos con información de FAAS

En este ítem se describen los procedimientos y resultados obtenidos al aplicar los diferentes métodos de clasificación con información procedente del análisis de espectroscopía de absorción atómica de llama (FAAS).

6.3.4.1.1. Análisis Discriminante Lineal

En una primera etapa de clasificación se realizó un Análisis Discriminante, con un modelo lineal y se evaluaron los resultados de interés que arroja la matriz de confusión.

En la Figura 6.2 se observa la distribución de las nubes de puntos correspondientes a cada una de las zonas productoras. Las observaciones del sudeste de Corrientes se ubican entre el 2do y 3er cuadrante, perfectamente separadas de las otras tres zonas. Las observaciones del nordeste de Entre Ríos, ubicadas en el cuarto cuadrante. El grupo de observaciones de las otras dos zonas no se separan de la misma manera, los jugos del centro oeste de Corrientes y del centro sur de Misiones se localizan en el cuarto cuadrante.

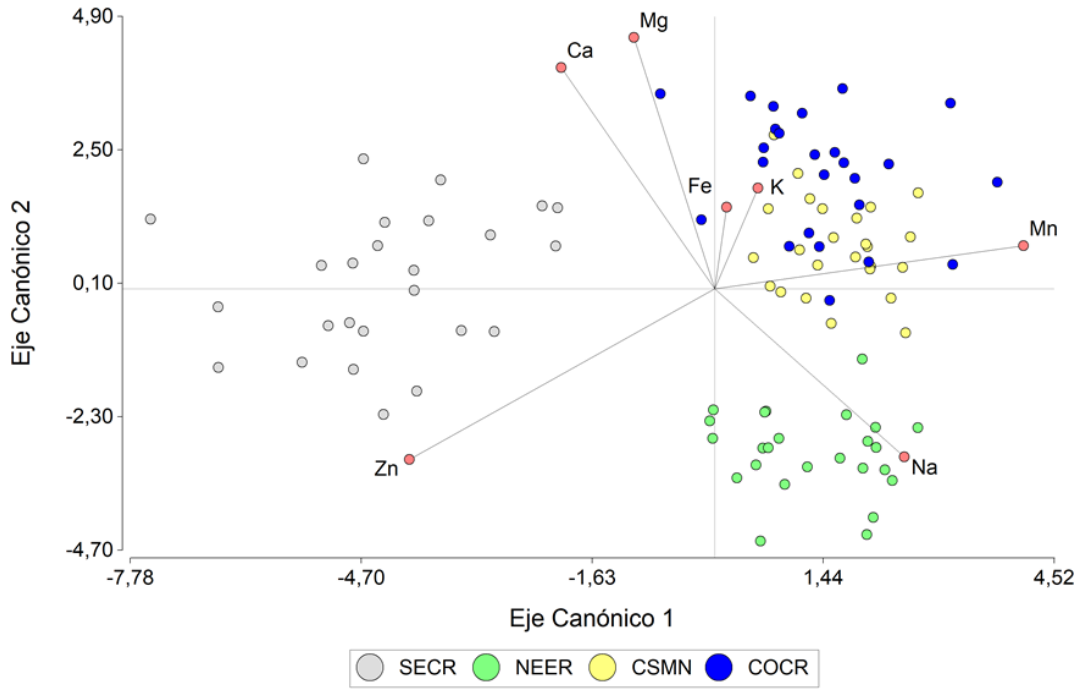

Figura 6.2. Biplot de los contenidos de elementos minerales en jugo de naranja y las zonas productoras (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) sobre los dos primeros ejes canónicos del Análisis Discriminante Lineal

Las funciones discriminantes canónicas son presentadas en la Tabla 6.7.

Tabla 6.7. Funciones Discriminantes Canónicas de contenidos minerales de jugos de naranja del nordeste argentino

	Función discriminante canónica	
Elemento	1	2
Ca	-0,30	**0,59**
Fe	0,02	0,22
K	0,09	**0,74**
Mg	-0,16	**0,67**
Mn	**0,61**	0,11
Na	0,37	-0,45
Zn	**-0,60**	-0,45

La primera función discriminante, que permite separar los jugos provenientes del sudeste de Corrientes, se construye, principalmente, con los contenidos de Mn y Zn. La segunda función discriminante, sobre la que se diferencian las nubes de puntos correspondientes al noreste de Entre Ríos del conjunto formado por las observaciones de centro sur de Misiones y centro oeste de Corrientes, se construye fundamentalmente con los contenidos de K, Mg y Ca en orden de importancia.

A continuación, se presentan los resultados de interés de la clasificación con Análisis Discriminante Lineal para evaluar el desempeño del método en la matriz de confusión (Tabla 6.8), donde se observa un 88% de acierto en la clasificación de las observaciones de centro oeste de Corrientes, un 96% en las de centro sur de Misiones y un 100% de acierto, con todas las observaciones bien clasificadas, en el caso de noreste de Entre Ríos y sudeste de Corrientes.

Tabla 6.8. Criterios de Selección de Modelos para Análisis Discriminante Lineal por zona de producción (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona	Acierto (%)	Sensibilidad	Especificidad	κ
COCR	88	0,88	0,98	0,89
CSMN	96	0,96	0,96	0,89
NEER	100	1,00	1,00	1,00
SECR	100	1,00	1,00	1,00

Clasifica correctamente, con un 100% de acierto, los jugos de naranja del noreste de Entre Ríos y sudeste de Corrientes. En el caso de los jugos provenientes del centro sur de Misiones existe un 12% de jugos de esa procedencia mal clasificados y un 4% proveniente del centro oeste de Corrientes, clasificado como centro sur de Misiones. El comportamiento general del método fue muy bueno, con un acierto de 94%, sensibilidad de 95%, especificidad de 97%, exactitud de 96% e índice κ = 0,94.

6.3.4.1.2. Árboles de Decisión

Se han generado Árboles de Decisión utilizando diferentes algoritmos, los que han presentado comportamientos similares en relación con los criterios de decisión empleados y para las diferentes zonas productoras. En la Tabla 6.9 se presentan los valores de los criterios de selección de modelos para los casos de mejor comportamiento.

Tabla 6.9. Criterios de selección de modelos por algoritmo usado en la generación de Árboles de Decisión por zona productora (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Algoritmo	Porcentaje de acierto	Índice κ
C5.0	**0,90**	**0,86**
CART (Rpart)	0,85	0,80
CART (Rpart 2)	0,85	0,80

Dentro de los Árboles de Decisión, el algoritmo C5.0 presentó mejor comportamiento que los CART (Rpart y Rpart2), en relación con el porcentaje de acierto y el índice κ, el clasificador obtenido es de tipo regla, sin filtrado de variables y 10 iteraciones. El algoritmo C5.0 arrojó un acierto general de 90% y un índice κ global de 0,86, resultados similares fueron informados por Lubinska-Szczygieł *et al.* (2018), al clasificar jugos de limas. En la tabla 6.10, se presentan los valores de sensibilidad y especificidad, logradas por zona productiva con la aplicación de este algoritmo.

Tabla 6.10. Sensibilidad y especificidad logradas con Árboles de Decisión generados con el algoritmo C5.0 por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona productiva	Sensibilidad	Especificidad
COCR	0,90	0,93
CSMN	0,80	0,97
NEER	1,00	0,97
SECR	0,90	1,00

Al analizar por zona de producción utilizando este algoritmo, para el noreste de Entre Ríos, sudeste y centro oeste de Corrientes se consigue alta sensibilidad y especificidad, mientras que en el centro sur de Misiones los valores de sensibilidad son levemente inferiores, en todas las zonas se observaron altos valores de especificidad.

En la Tabla 6.11 se presentan las reglas de clasificación obtenidas mediante el algoritmo C5.0 considerando cada zona de producción.

Tabla 6.11. Reglas de decisión para la clasificación de los jugos de naranja procedentes de las diferentes zonas productivas (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) obtenidas mediante Árboles de Decisión mediante algoritmo C5.0

Clasificadores	Zona productiva	Acierto (%)
Na ≤ 38,52, Zn ≤ 0,54	COCR	94,0
K > 819,09, Zn ≤ 0.54	COCR	88,4
Mg > 243,68	COCR	77,0
K ≤ 819,90, Mg ≤ 243,68, Na > 38,52, Zn ≤ 0,54	CSMN	90,3
Mn > 0,05, Zn > 0,54	NEER	93,3
Mn ≤ 0,05	SECR	83,9

La siguiente figura refleja la importancia de las distintas variables para determinar la trazabilidad de los jugos de naranja.

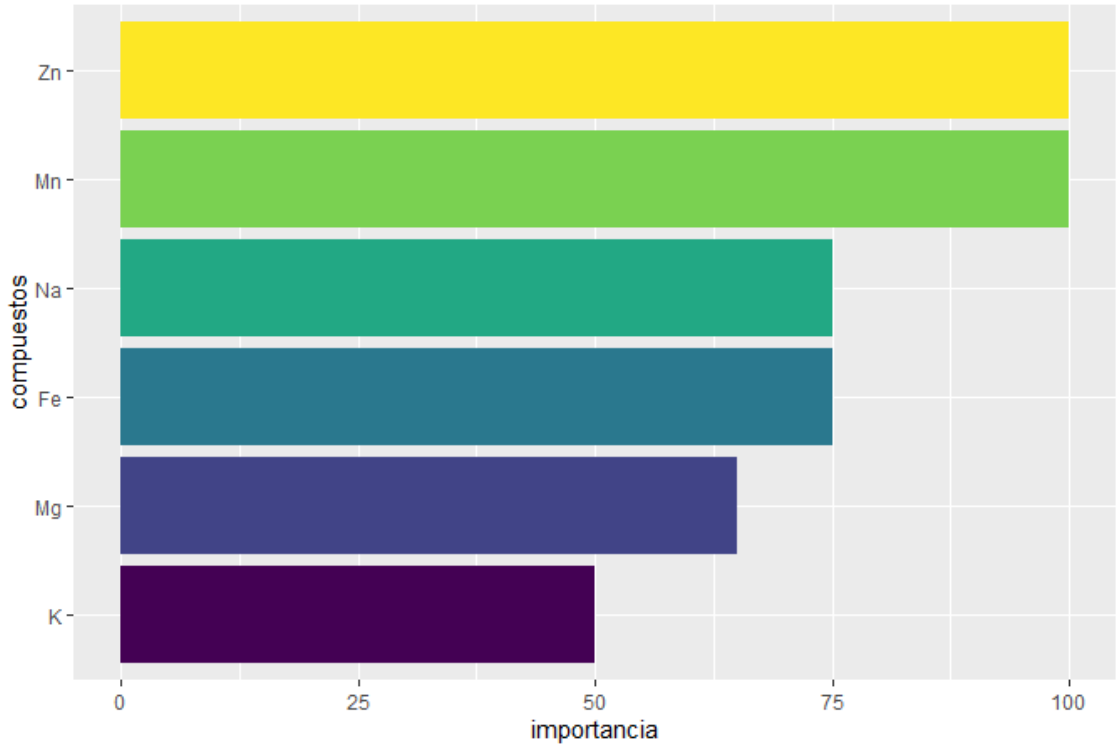

Figura 6.3. Importancia de las distintas variables para determinar la trazabilidad de los jugos de naranja.

Los contenidos de Ca no revistieron importancia para la clasificación por zona de los jugos de naranja. K y Mg participan en la clasificación con menos del 65% y los demás elementos presentan valores del 75% o más, se destacan Zn y Mn con 100% de importancia en la determinación del origen de los jugos.

6.3.4.1.3. K-Vecino más Cercano

Con el método KNN se han probado diferentes opciones del parámetro *k* (número de vecinos que se tiene en cuenta para predecir). En la Tabla 6.12 se presentan los valores de acierto e índice κ.

Tabla 6.12. Criterios de selección de modelos para diferentes valores del parámetro k usando métodos va2gos de K-Vecinos más Cercanos

Valor de k	Porcentaje de acierto	κ
5	**0,67**	**0,55**
9	0,66	0,54
17	0,64	0,52

Los mejores resultados se obtuvieron con 5 vecinos más cercanos, no obstante, los valores obtenidos de porcentaje de acierto e índice κ, son relativamente bajos, lo que indica que la técnica de K-Vecinos más Cercanos no sería la mejor opción para clasificar jugos de naranja en función de su contenido mineral.

Los porcentajes de acierto obtenidos con el método KNN en este Capítulo son muy inferiores a los obtenidos por Li *et al.* (2014), quienes encontraron que el KNN fue la técnica de mejor comportamiento, para clasificar arándanos en diferentes etapas de crecimiento, con un acierto entre 85 y 98%.

6.3.4.1.4. Redes Neuronales Artificiales

Variando el número de capas ocultas, la cantidad de neuronas por capa y de iteraciones se han podido generar diferentes modelos. En la Tabla 6.13 se presentan los criterios de selección para los diferentes modelos definidos mediante la técnica de Redes Neuronales Artificiales (ANN).

Tabla 6.13. Criterios de selección de modelos por algoritmo usado en la generación de Redes Neuronales Artificiales

Nro de capas ocultas	Neuronas por capa oculta	Porcentaje de acierto	Índice κ
1	**10**	**0,92**	**0,89**
1	25	0,88	0,84
1	35	0,90	0,86
2	16-13	0,88	0,84
3	14-10-5	0,81	0,74

Los mejores resultados se obtuvieron con el algoritmo MLP con 1 capa oculta de 10 neuronas, lográndose un acierto del 92% y un índice $\kappa = 0,89$, resultados superiores a los obtenidos por Turra *et al*. (2017) al clasificar naranjas en orgánicas y no orgánicas con ANN logrando un 83% de acierto.

En la Tabla 6.14 se presentan la sensibilidad y especificidad por zona productiva para el modelo escogido.

Tabla 6.14. Sensibilidad y especificidad logradas con Redes Neuronales con 1 capa de 35 neuronas por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona de producción	Sensibilidad	Especificidad
COCR	0,80	0,96
CSMN	0,90	0,93
NEER	1,00	1,00
SECR	1,00	1,00

Con este algoritmo se podrían clasificar con muy buena sensibilidad y especificidad los jugos provenientes de dos de las cuatro zonas estudiadas (noreste de Entre Ríos y sureste de Corrientes) mientras que para la zona del centro sur de Misiones y el centro oeste de Corrientes se observan valores de sensibilidad y especificidad más bajos.

Los gráficos que se presentan a continuación indican la importancia de los diferentes elementos en la construcción del modelo seleccionado por zona productora, en ellos se observa que, para las distintas zonas, la incidencia de los elementos en la clasificación de las muestras es diferente.

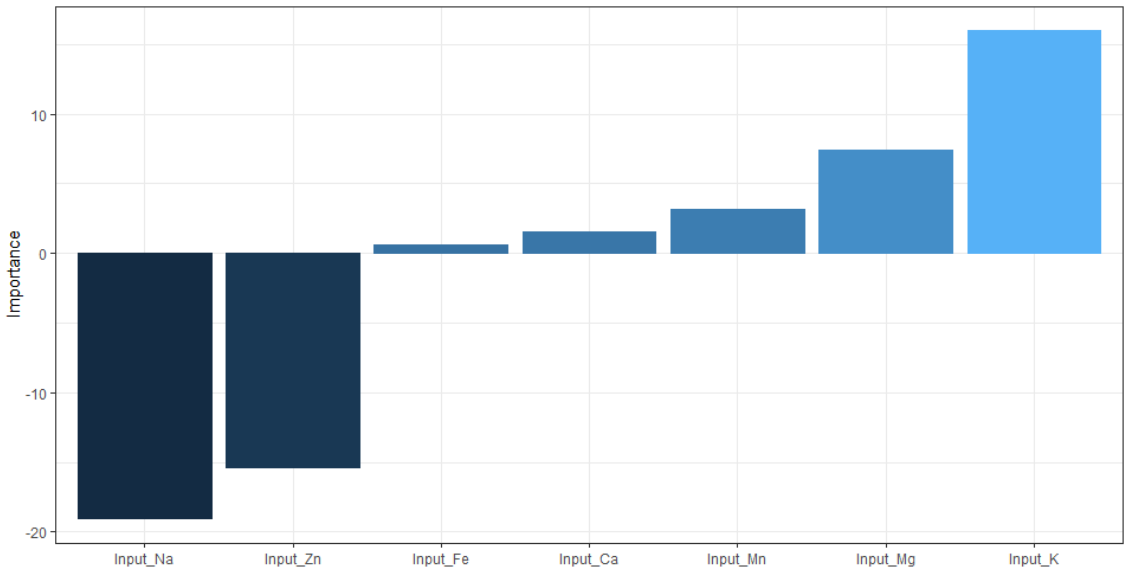

Figura 6.4. Importancia de los elementos en la construcción del modelo ANN para la zona productora del centro oeste de Corrientes

Para la clasificación de los jugos en la zona del centro oeste de Corrientes el método considera principalmente los contenidos de Na y Zn en sentido negativo y K y Mg en sentido positivo.

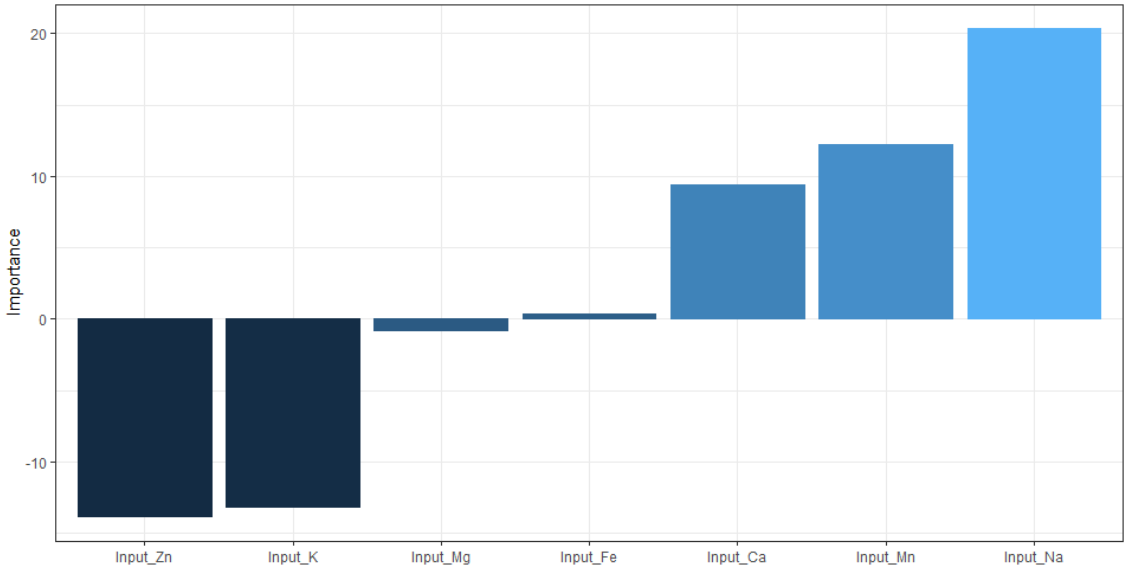

Figura 6.5. Importancia de los elementos en la construcción del modelo ANN para la zona productora del centro sur de Misiones

En la clasificación de los jugos en la zona del centro sur de Misiones se consideran principalmente los contenidos de Zn y K en sentido negativo y Na, Mn y Ca en sentido positivo.

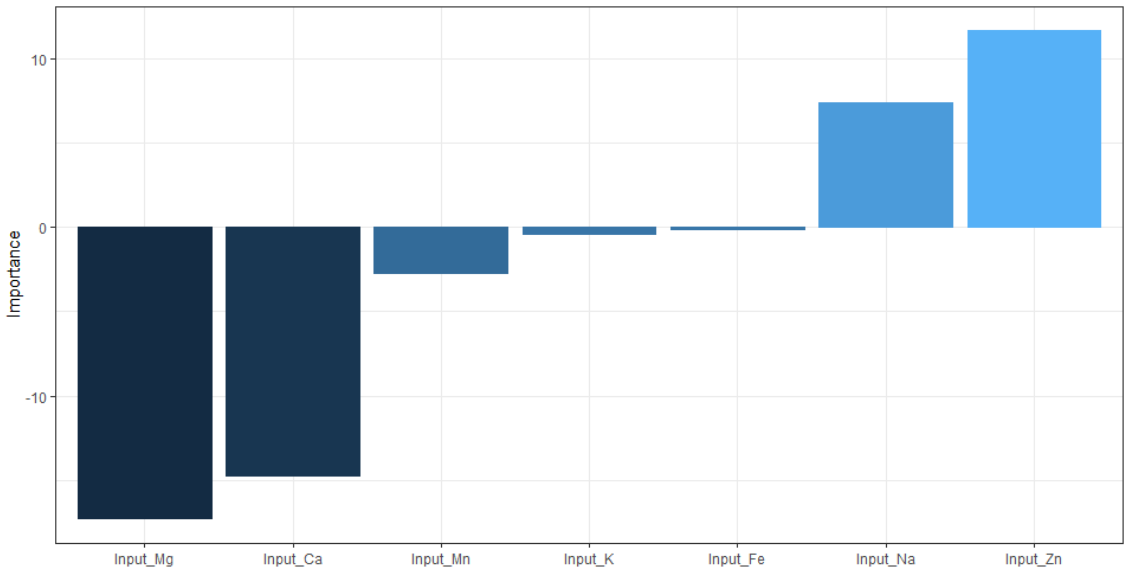

Figura 6.6. Importancia de los elementos en la construcción del modelo ANN para la zona productora del noreste de Entre Ríos

En la zona del noroeste de Entre Ríos los elementos que participan en la clasificación son de Mg y Ca en sentido negativo, Zn y Na en sentido positivo.

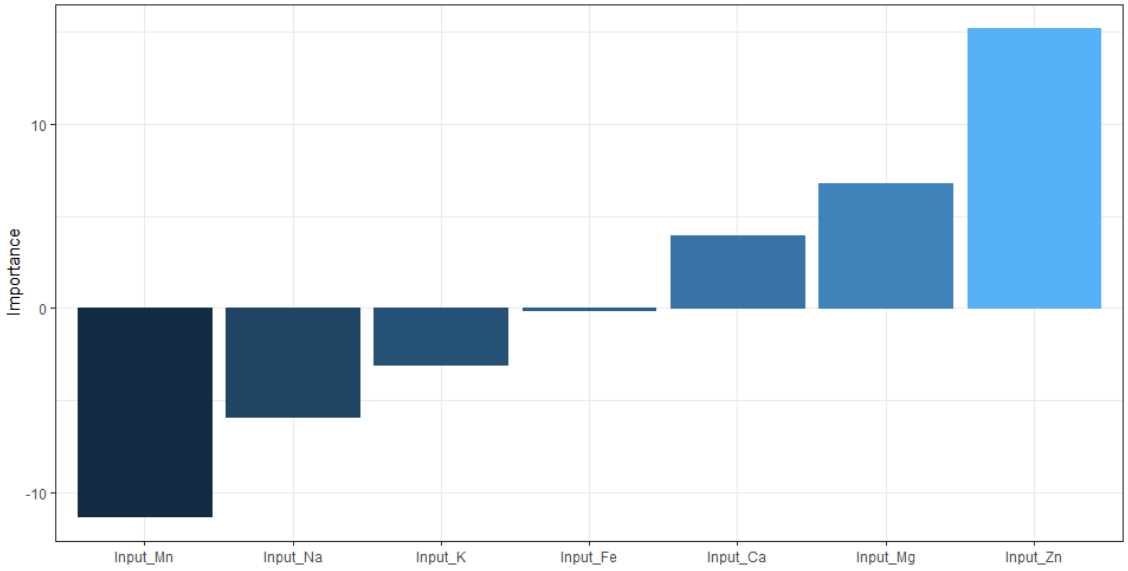

Figura 6.7. Importancia de los elementos en la construcción del modelo ANN para la zona productora del sudeste de Corrientes

En la zona del sudeste de Corrientes Mn y Na en sentido negativo y Zn y Mg en sentido positivo, son los elementos con mayor importancia en la clasificación.

Analizando las Figuras 6.4 a 6.7 se puede establecer que cuando se utilizaron ANN, el elemento Fe, fue el único que no tuvo importancia en la clasificación por origen geográfico de los jugos estudiados.

6.3.4.1.5. Máquinas de Vectores Soporte

Las SVM calculan un hiperplano de separación óptimo mediante un algoritmo iterativo que aprende la distribución de la muestra en los límites de cada clase considerada. Para evitar el sobreajuste se utilizaron las funciones *kernel* radial y polinomial para la clasificación. En la Tabla 6.15 se presentan los valores de acierto e índice κ para cada modelo.

Tabla 6.15. Criterios de selección de modelos para cada función *kernel* definida en Máquinas de Vectores Soporte

Kernel	Parámetros	Porcentaje de acierto	Índice κ
Radial	σ = 1 y C = 5	0,87	0,83
Polinomial	**Grado = 2 y c= 1**	**0,90**	**0,87**

El modelo con un *kernel* polinomial de grado 2 y c = 1 fue seleccionado por ser el de mejor comportamiento, con un acierto general de 90% y un índice κ de 0,87. Estos resultados son similares a los obtenidos por Turra *et al.* (2017) con SVM para clasificar naranjas en orgánicas y no orgánicas, alcanzando un 93% de acierto.

Tabla 6.16. Sensibilidad y especificidad logradas con Máquinas de Vectores Soporte con un *kernel* grado 2 y *c* = 1 por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona de producción	Sensibilidad	Especificidad
COCR	0,70	0,97
CSMN	0,90	0,90
NEER	1,00	1,00
SECR	1,00	1,00

Como se puede ver en la tabla 6.16 con el algoritmo seleccionado se logra la mayor sensibilidad y especificidad para la zona productora del nordeste de Entre Ríos y el sureste de Corrientes (100% para ambos valores); seguidos del centro sur de Misiones, y el peor desempeño del método puede observarse para la zona del centro oeste de Corrientes con valores muy bajos de sensibilidad (70%), por lo que no sería un buen método para clasificar los jugos de naranjas.

6.3.4.1.6. Comparación de modelos

A los efectos de comparar los diferentes métodos de clasificación, en la Tabla 6.17 se presentan los diferentes criterios de porcentaje de acierto e índice κ.

Tabla 6.17. Criterios de selección de modelos de clasificación

Método	Porcentaje de acierto	κ
LDA	**96**	**0,95**
DT	90	0,86
KNN	67	0,55
ANN	92	0,89
SVM	90	0,87

Se puede establecer el siguiente orden para las técnicas, en la clasificación de jugos de naranja con información proveniente de FAAS: LDA 94% > ANN 92% > SVM = DT 90% > KNN 67%.

Lubinska-Szczygieł *et al.* (2018) informaron resultados similares al clasificar jugos de limas utilizando DT. Los aciertos logrados con ANN son algo inferiores a los de Sabanci *et al.* (2016), quienes utilizaron KNN y ANN para clasificar diferentes variedades de manzana y encontraron que ANN presentó mejor comportamiento con un acierto de 98,89%. Astuti *et al.* (2018), comparan en uso de ANN y SVM para definir algoritmos a fin de realizar una clasificación automática de frutas, la exactitud alcanzada con SVM (100%) fue mejor que con ANN (50%). Si bien el orden jerárquico de los métodos según porcentajes de acierto coincide, el acierto obtenido por estos autores para SVM fue superior y para ANN fue muy inferior.

Los resultados obtenidos señalan para SVM porcentajes de acierto inferiores a los encontrados para limones en el Capítulo IV (publicado en Gaiad *et al.*, 2016) no obstante, los rendimientos de LDA son similares y los de KNN muy inferiores. En relación con los obtenidos en el Capítulo V para jugos de mandarina, los porcentajes de acierto fueron similares para SVM, ANN y DT, algo superiores para LDA e inferiores para KNN.

6.3.4.2. Selección de modelos con información de MP-AES

En este ítem se describen los procedimientos y resultados obtenidos al emplear los diferentes métodos de clasificación con información procedente del MP-AES.

6.3.4.2.1. Análisis Discriminante Lineal

En una primera etapa de clasificación se realizó un Análisis Discriminante, con un modelo lineal y se evaluaron los resultados de interés que arroja la matriz de confusión.

En la Figura 6.8 se observa la distribución de las nubes de puntos correspondientes a cada una de las zonas productoras. Las observaciones del sudeste de Corrientes se ubican entre el 2do y 3er cuadrante, perfectamente separadas de las otras tres zonas. Las observaciones del nordeste de Entre Ríos, ubicadas en el primer y segundo cuadrantes. El grupo de observaciones de las otras dos zonas no se separan de la misma manera, los jugos del centro oeste de Corrientes y del centro sur de Misiones se localizan en el cuarto cuadrante.

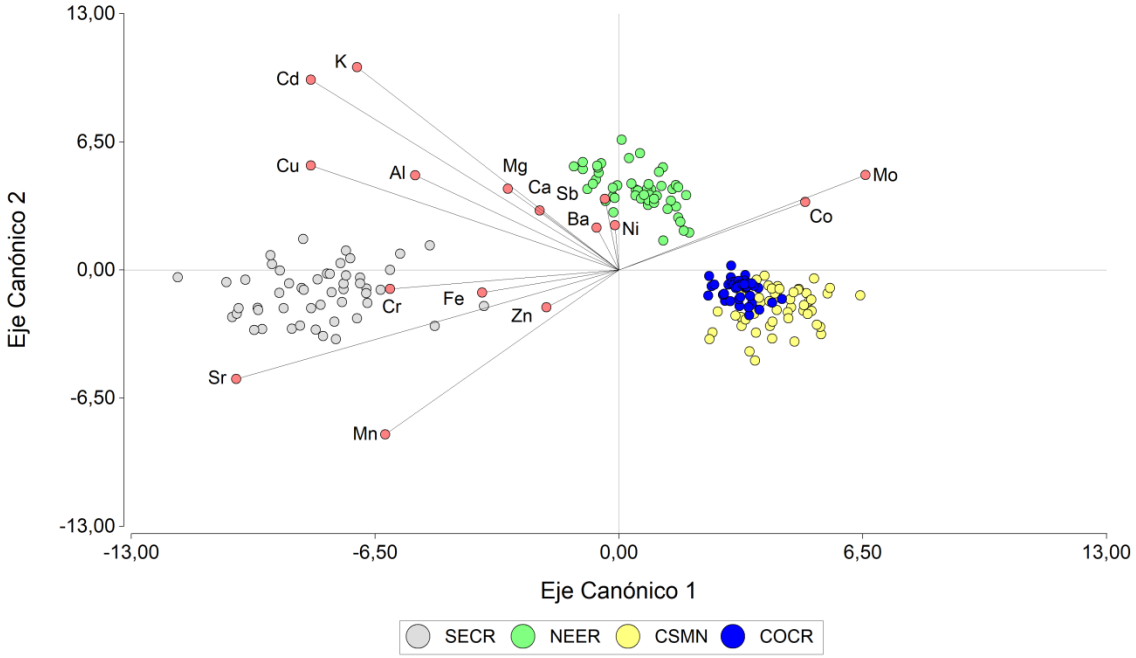

Figura 6.8. Biplot que representa los contenidos de elementos minerales en jugo de naranja y las zonas productoras (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) sobre los dos primeros ejes canónicos del Análisis Discriminante Lineal

Las funciones discriminantes canónicas son presentadas en la Tabla 6.18.

Tabla 6.18. Funciones Discriminantes Canónicas de contenidos minerales de jugos de naranja del nordeste argentino

	Función discriminante canónica	
Elemento	1	2
Al	-0,39	0,35
Ba	-0,04	0,16
Ca	-0,15	0,22
Cd	**-0,59**	**0,70**
Co	0,36	0,25
Cr	-0,44	-0,07
Cu	**-0,59**	0,38
Fe	-0,26	-0,08
K	**-0,51**	**0,74**
Mg	-0,21	0,30
Mn	-0,45	**-0,60**
Mo	0,48	0,35
Ni	-0,01	0,16
Sb	-0,03	0,26
Sr	**-0,74**	-0,40
Zn	-0,14	-0,14

La primera función discriminante, permite separar los jugos provenientes del sudeste de Corrientes y se construye, principalmente, con los contenidos de Sr, Cu, Cd y K (en orden de importancia). La segunda función discriminante, sobre la que se diferencian las nubes de puntos correspondientes al noreste de Entre Ríos de las de centro sur de Misiones y centro oeste de Corrientes, se construye fundamentalmente con los contenidos de K, Cd y Mn (en orden de importancia).

A continuación, se presentan los resultados de interés de la clasificación con Análisis Discriminante Lineal para evaluar el desempeño del método en la matriz de confusión (Tabla 6.19), donde se observa un 94% de acierto en las de centro sur de Misiones y 100% de acierto en la clasificación de las observaciones de centro oeste de Corrientes, sudeste de Corrientes y noreste de Entre Ríos.

Tabla 6.19. Criterios de Selección de Modelos para Análisis Discriminante Lineal por zona de producción (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona	Acierto (%)	Sensibilidad	Especificidad	κ
COCR	100	1,00	0,98	0,96
CSMN	94	0,94	1,00	0,96
NEER	100	1,00	1,00	1,00
SECR	100	1,00	1,00	1,00

El comportamiento general del LDA resulta adecuado, con un acierto global de 99%, sensibilidad de 99%, especificidad de 99%, exactitud de 99% e índice κ = 0,98. Altos valores de sensibilidad y especificidad simultáneos son atributos de un buen método de clasificación (Takaya & Rehmsmeier, 2015).

Estos resultados indican mejor comportamiento del método que los obtenidos en los Capítulos IV y V de esta tesis, en los que, mediante la aplicación de LDA para clasificar limones y mandarinas según su origen geográfico, se obtuvieron valores menores de sensibilidad, especificidad e índice Kappa y también que los de Hong *et al.* (2019), quienes analizaron las concentraciones de elementos macro, micro y traza en diferentes variedades de frutos cítricos (obtenidas por ICP-MS y ICP-OES) y lograron separar las muestras en grupos mediante LDA con 94% de clasificaciones correctas.

6.3.4.2.2. Árboles de Decisión

Los algoritmos probados han demostrado diferente comportamiento en relación con los criterios de decisión empleados y para las diferentes zonas productoras. En las Tablas 6.20 y 6.21 se presentan los valores de los criterios de selección de modelos para los casos de mejor comportamiento.

Tabla 6.20. Criterios de selección de modelos para diferentes algoritmos usados en la generación de Árboles de Decisión

Algoritmo	Porcentaje de acierto	Índice κ
C5.0	**0,99**	**0,98**
CART (Rpart)	0,99	0,98
CART (Rpart 2)	0,99	0,98

Con el algoritmo C5.0 y con los CART (Rpart y Rpart2) se obtuvieron idénticos resultados en la fase de prueba, no obstante, en la fase de entrenamiento el algoritmo C5.0 tuvo mejor comportamiento, a lo que se suma su mayor sencillez, motivos por los que fue seleccionado. El clasificador obtenido es de tipo regla, sin filtrado de variables y con una iteración.

Tabla 6.21. Sensibilidad y especificidad logradas con Árboles de Decisión generados con el algoritmo C5.0 por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona productiva	Sensibilidad	Especificidad
COCR	1,00	1,00
CSMN	0,95	1,00
NEER	1,00	1,00
SECR	1,00	0,98

Con el algoritmo C5.0, se logra la máxima sensibilidad y especificidad para las zonas productoras del centro oeste de Corrientes y nordeste de Entre Ríos (NEER); en la de centro sur de Misiones, la especificidad resultó del 100% pero la sensibilidad fue menor; y en el sudeste de Corrientes se obtuvo la máxima sensibilidad, pero con menor especificidad.

En esta tesis se han obtenido valores superiores con el uso de DT que los que encontraron Buratti *et al*. (2004), quienes clasificaron vinos italianos empleando PCA, LDA y DT, obteniendo, con DT, aciertos de 87%.

En la Tabla 6.22 se presentan las reglas de decisión para la clasificación, con el modelo seleccionado, de los jugos procedentes de las diferentes zonas productivas.

Tabla 6.22. Reglas de decisión para la clasificación de los jugos de naranja procedentes de las diferentes zonas productivas (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR) obtenidas mediante Árboles de Decisión mediante algoritmo C5.0

Clasificadores	Zona productiva	Acierto (%)
Cd ≤ 0,35, Mg > 94,13, Sr ≤ 1,70	COCR	95,2
Mg > 94,13, Mo ≤ 0, Sr ≤ 1,70	COCR	82,6
Ca ≤ 17,43, Mg ≤ 94,13	CSMN	97,7
Cd > 0,35, Mg > 94,13, Mo > 0, Sr ≤ 1,70	NEER	97,8
Ca > 17,43, Mg ≤ 94,13	NEER	83,0
Mg > 94,13, Sr > 1,70	SECR	90,9

La siguiente figura refleja la importancia de las distintas variables para determinar la trazabilidad de los jugos de naranja.

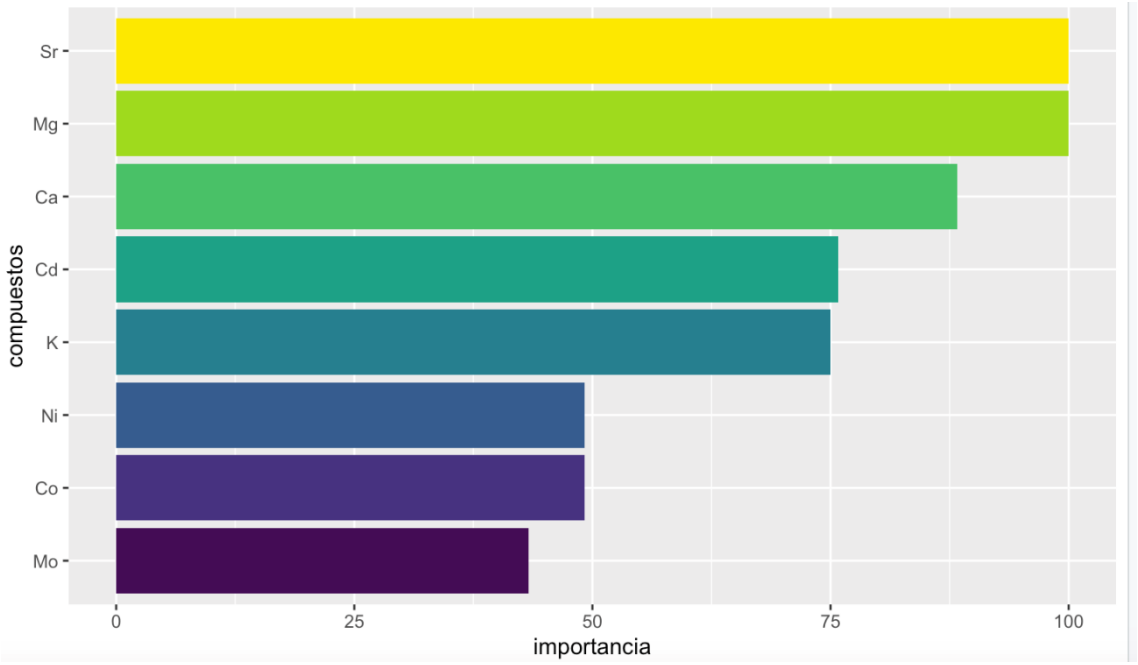

Figura 6.9. Importancia de las distintas variables para determinar la trazabilidad de los jugos de naranja

Al, Ba, Cr, Cu, Fe, Mn, Pb, Sb y Zn no resultaron de importancia para la clasificación por zona de los jugos de naranja. Mo, Co y Ni tienen baja participación en la clasificación (menos del 50%), Ca, Cd, K, Mg y Sr son los más importantes.

6.3.4.2.3. K-Vecino más Cercano

El análisis de KNN se realizó variando los valores el parámetro k (cantidad de vecinos más cercanos), los resultados obtenidos se presentan en la Tabla 6.23.

Tabla 6.23. Criterios de selección de modelos para diferentes valores del parámetro k usando métodos vagos de K-Vecinos más Cercanos

Valor de k	Porcentaje de acierto	κ
5	0,73	0,63
7	0,68	0,58
9	0,73	0,63
11	0,70	0,60
13	0,71	0,61
15	**0,74**	**0,66**
17	0,73	0,63
19	0,72	0,62
21	0,73	0,65
23	0,72	0,62

Los mejores resultados se obtuvieron con 15 vecinos más cercanos, no obstante, los bajos porcentajes de acierto e índice κ indican que K-Vecinos más Cercanos no sería la técnica de mejor comportamiento para clasificar jugos de naranja en función de su contenido mineral.

6.3.4.2.4. Redes Neuronales Artificiales

Variando el número de capas ocultas, la cantidad de neuronas por capa y de iteraciones se han podido generar diferentes modelos con el algoritmo MLP (perceptrón multicapa). En la Tabla 6.24 se presentan los criterios de selección para los diferentes algoritmos usados en la técnica de Redes Neuronales Artificiales (ANN).

Tabla 6.24. Criterios de selección de modelos por algoritmo usado en la generación de Redes Neuronales Artificiales

Nro de capas ocultas	Neuronas por capa oculta	Porcentaje de acierto	Índice κ
1	2	0,73	0,63
1	10	0,95	0,93
3	**15-11-5**	**0,99**	**0,98**

Los mejores resultados se obtuvieron con el algoritmo MLP con 3 capas ocultas de 15, 11 y 5 neuronas, lográndose un acierto del 99% y un índice κ = 0,98. En la Tabla 6.25 se presentan la sensibilidad y especificidad por zona productiva para el modelo escogido.

Tabla 6.25. Sensibilidad y especificidad logradas con Redes Neuronales con 3 capas de 11-15-5 neuronas por zona productiva (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona de producción	Sensibilidad	Especificidad
COCR	1,00	1,00
CSMN	1,00	1,00
NEER	1,00	0,98
SECR	0,95	1,00

Con este algoritmo se logra máxima sensibilidad y especificidad para las zonas productoras de centro oeste de Corrientes y centro sur de Misiones; para el nordeste de Entre Ríos la sensibilidad resultó del 100% con una pequeña reducción de la especificidad; y para el sudeste de Corrientes, se obtuvo la máxima especificidad, con una pequeña reducción de la sensibilidad.

Los porcentajes de acierto obtenidos son superiores a los logrados por Turra *et al.* (2017) al clasificar naranjas en orgánicas y no orgánicas por su contenido en elementos químicos, quienes lograron un 83% de acierto con ANN.

También son mayores a los de Vijayarekha & Govindaraj (2006), quienes utilizaron ANN para identificar mandarinas con y sin defectos y lograron un 84% de frutas bien clasificadas para frutas con picaduras, 50% con pudrición y 100% con rajaduras.

Alonso Salces *et al*. (2005) estudiaron los perfiles poli fenólicos de manzanas para sidra de acuerdo con su estado de maduración empleando DA, KNN y ANN; las Redes

Neuronales Artificiales presentaron un excelente comportamiento con éxitos en la predicción de 97% en la categoría de frutos inmaduros y el 99% para la de maduros, valores similares a los obtenidos en este Capítulo.

A continuación, se presentan gráficos que indican la importancia de los diferentes elementos en la construcción del modelo seleccionado por zona productora, en los que se observa, que la incidencia de los elementos en la clasificación de las muestras es diferente para las distintas zonas.

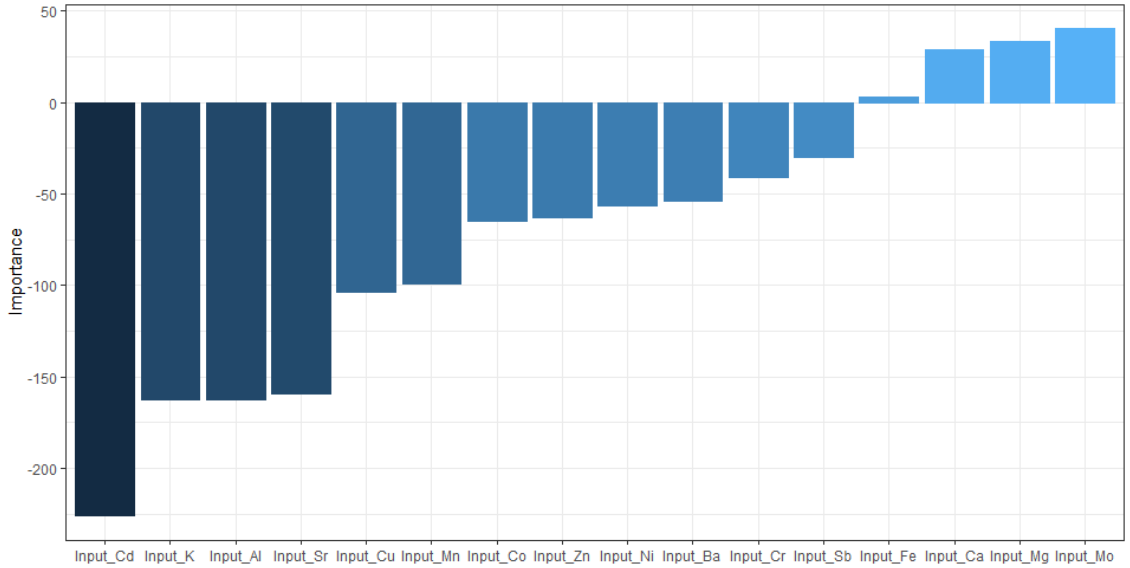

Figura 6.10. Importancia de los elementos en la construcción del modelo ANN para la zona productora del centro oeste de Corrientes

En la zona del centro oeste de Corrientes la mayoría de los elementos tienen sentido negativo, siendo los más importantes el Cd, K, Al y Sr. Solamente Mo, Mg, Ca y Fe se presentan en sentido positivo, pero con baja importancia para la clasificación.

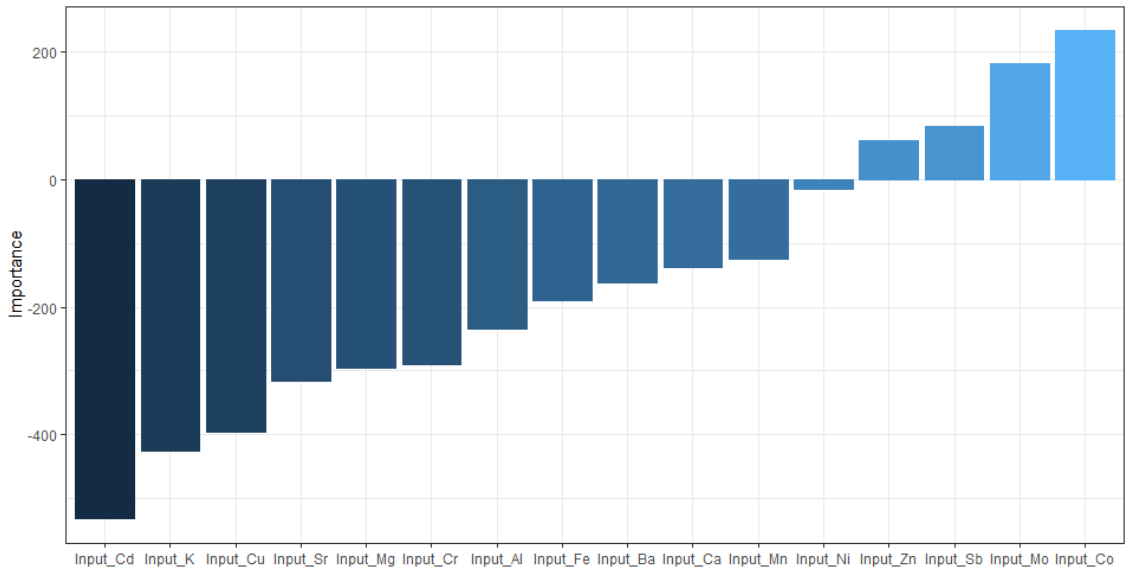

Figura 6.11. Importancia de los elementos en la construcción del modelo ANN para la zona productora del centro sur de Misiones

Al igual que en el caso anterior, en la clasificación de los jugos del centro sur de Misiones, la mayoría de los elementos tienen sentido negativo, siendo los más importantes el Cd, K, Cu y Sr. Solamente Co, Mo, Sb, Zn y Ni se presentan en sentido positivo, pero con menor importancia para la clasificación.

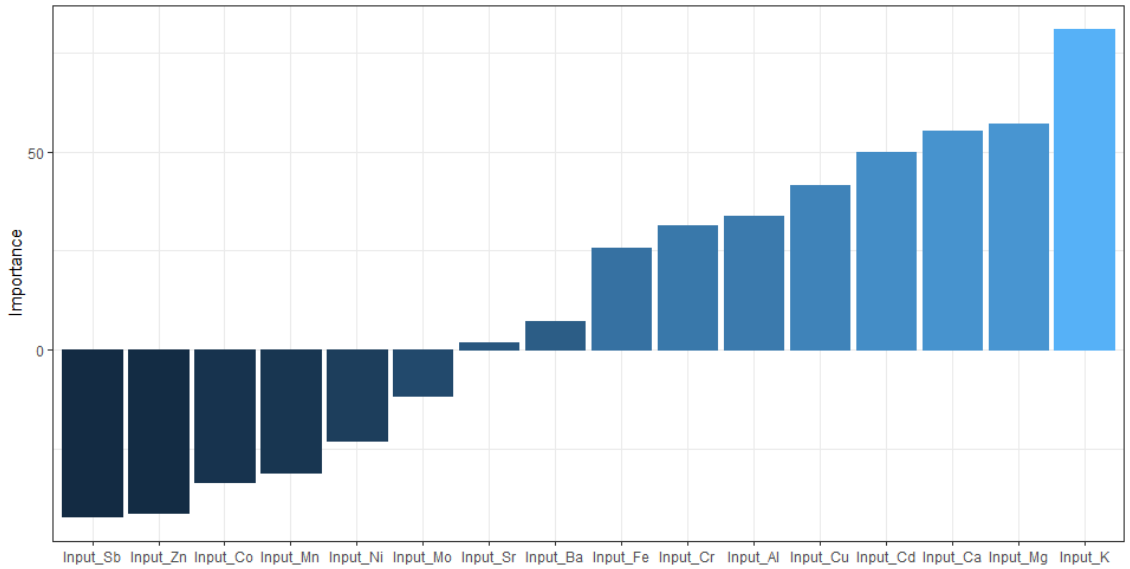

Figura 6.12. Importancia de los elementos en la construcción del modelo ANN para la zona productora del noreste de Entre Ríos

En la clasificación de los jugos del noreste de Entre Ríos, los elementos más importantes son el K, Mg, Ca y Cd en sentido positivo y con menor importancia Sb y Zn en sentido negativo.

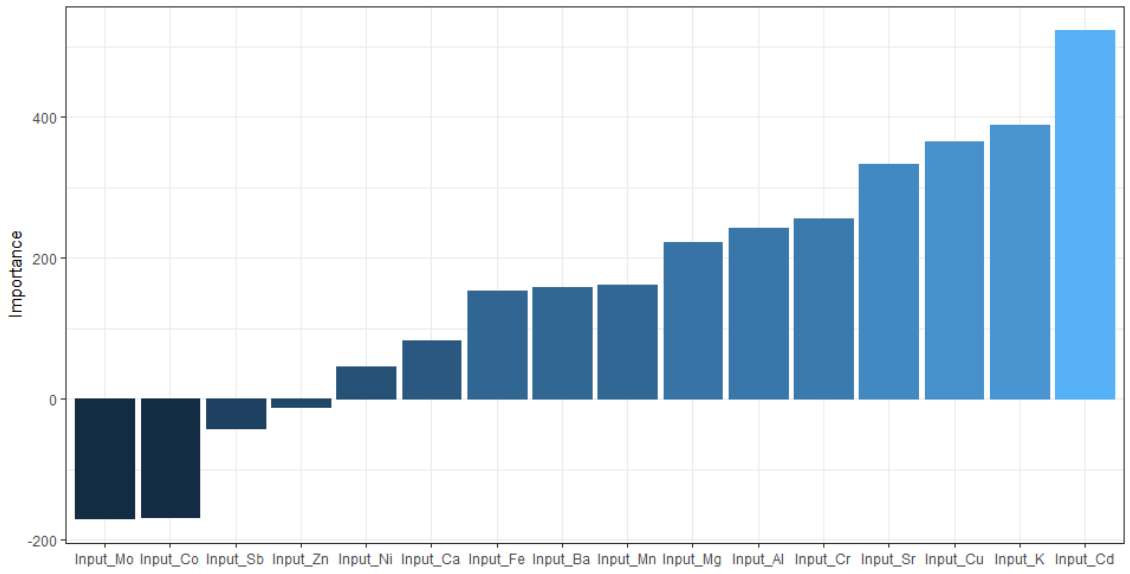

Figura 6.13. Importancia de los elementos en la construcción del modelo ANN para la zona productora del sudeste de Corrientes

A diferencia de las zonas centro oeste de Corrientes y centro sur de Misiones, en la clasificación de los jugos del sudeste de Corrientes, la mayoría de los elementos se presentan en sentido positivo, los elementos más importantes son Cd, K, Cu y Sr. Mo, Co, Sb y Zn tienen sentido negativo pero menor importancia en la clasificación.

Analizando las Figuras 6.10 a 6.13 se puede establecer que, dentro de los elementos con importancia en la discriminación, los que se encuentran presentes en todas las zonas productivas son Cd y K, a los que se suma el Sr en 3 y Cu en 2 de las 4 zonas analizadas.

6.3.4.2.5. Máquinas de Vectores Soporte

Debido a que las SVM calculan un hiperplano de separación óptimo mediante un algoritmo iterativo que aprende la distribución de la muestra en los límites de cada clase considerada, se suele dar un sobreajuste por lo que para evitarlo se utilizaron las funciones *kernel* radial y polinomial para la clasificación. En la Tabla 6.26 se presentan los valores de acierto en índice κ para cada modelo.

Tabla 6.26. Criterios de selección de modelos para cada función kernel definida en Máquinas de Vectores Soporte

Kernel	Parámetros	Porcentaje de acierto	Índice κ
Radial	σ = 1 y C = 1	0,74	0,66
Radial	σ = 1 y C = 5	0,76	0,68
Radial	σ = 3 y C = 1	0,47	0,30
Radial	σ = 3 y C = 5	0,49	0,32
Polinomial	Grado = 2 y c= 5	0,98	0,97
Polinomial	**Grado = 3 y c= 1**	**0,99**	**0,99**

El modelo con un *kernel* polinomial de grado 3 y c = 1 fue seleccionado por ser el de mejor comportamiento, con un acierto general de 99% y un índice κ de 0,99.

Tabla 6.27. Sensibilidad y especificidad logradas con Máquinas de Vectores Soporte con un *kernel* grado 3 y *c* = 1 por zona productora (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Zona de producción	Sensibilidad	Especificidad
COCR	1,00	1,00
CSMN	1,00	1,00
NEER	1,00	0,98
SECR	0,95	1,00

Con este algoritmo se logra la mayor sensibilidad y especificidad para las zonas productoras del centro oeste de Corrientes y el centro sur de Misiones (100% para ambos valores); en el noreste de Entre Ríos se obtuvo la mayor sensibilidad, pero menor especificidad; y en el sudeste de Corrientes se consiguió logró el valor máximo de especificidad, pero con menor sensibilidad.

Los resultados obtenidos con SVM en este Capítulo son similares a los de Astuti *et al.* (2018), quienes desarrollaron un sistema para clasificación automática de frutos basado aplicando ANN y SVM, logrando una tasa de acierto del 100% con SVM. A la vez que son algo superiores a los logrados por Turra *et al.* (2017) al clasificar naranjas en orgánicas y no orgánicas por su contenido en elementos químicos, quienes lograron un 93% de acierto con SVM.

6.3.4.2.6. Comparación de modelos

A los efectos de comparar los diferentes métodos de clasificación, en la Tabla 6.28 se presentan los diferentes criterios de porcentaje de acierto e índice κ.

Tabla 6.28. Criterios de selección de modelos para diferentes métodos

Método	Porcentaje de acierto	κ
LDA	99	0,98
DT	99	0,98
KNN	74	0,66
ANN	99	0,98
SVM	**99**	**0,99**

El método con peor comportamiento para la clasificación por origen de jugos de naranjas del NEA fue el de K-Vecinos más Cercanos, que en su mejor configuración solamente alcanzó un 74% de acierto, con un índice κ de 0,66. Los demás métodos probados tuvieron comportamientos muy buenos, con valores acierto e índice κ cercanos al máximo.

Se han obtenido valores superiores con el uso de DT a los informados por Buratti *et al*. (2004), quienes clasificaron vinos italianos empleando PCA, LDA y DT, obteniendo, con DT, aciertos de 87%. Lubinska-Szczygieł *et al.* (2018), al clasificar jugos de limas, encontraron valores inferiores al utilizar DT (87,5%) en comparación con SVM (100%).

Estos resultados coinciden con los de Sabanci *et al.* (2016), quienes utilizaron KNN y ANN para clasificar diferentes variedades de manzana y encontraron que ANN presentó mejor comportamiento con un acierto de 98,89%.

Astuti *et al.* (2018), comparan en uso de ANN y SVM para definir algoritmos a fin de realizar una clasificación automática de frutas, la exactitud alcanzada con SVM (100%) fue mejor que con ANN (50%) y con menor tiempo de entrenamiento. Si bien el orden jerárquico de los métodos coincide con los resultados de este Capítulo, el acierto obtenido por estos autores para el método ANN fue muy inferior. De manera similar, comparando los resultados obtenidos con los de Turra *et al*. (2017) se observa

mejor comportamiento de SVM con 93% de acierto y luego ANN con 83%, este último porcentaje de acierto inferior al presentado en este apartado.

Pérez *et al.* (2006) estudiaron el origen de frutillas, arándano y pera a partir de los perfiles multielementales, aplicaron Análisis de la Variancia (ANOVA), DA con funciones lineal y cuadrática, ANN y redes neuronales genéticas (GNN). Todos los modelos estudiados presentaron el 100% de exactitud en la clasificación de frutillas y arándanos, pero en el caso de las peras, con el LDA solamente obtuvieron entre 60-80% de acierto, con el DA cuadrático entre 85-100% y con ANN entre 80-90%, la técnica de mejor comportamiento fue GNN con 100% de acierto. Estos resultados coinciden con los obtenidos en este Capítulo en cuanto a la jerarquía de los métodos, pero los porcentajes de acierto para el LDA y ANN fueron inferiores.

Los resultados obtenidos en este Capítulo para jugos de naranjas señalan porcentajes de acierto similares para SVM a los encontrados para limones en el Capítulo IV (publicado en Gaiad *et al.*, 2016) no obstante, los rendimientos de LDA son algo superiores y los de KNN muy inferiores. En relación con los obtenidos en el Capítulo V para jugos de mandarina, los porcentajes de acierto fueron superiores para SVM, ANN, DT y LDA e inferiores para KNN.

Se puede establecer para las configuraciones de mejor comportamiento de la técnicas utilizadas en la clasificación de jugos de naranja el siguiente orden: SVM = ANN = DT = LDA 99% > KNN 74%.

Debido a los altos porcentajes de acierto obtenidos en todos los métodos, y aquellos que tienen niveles de concordancia muy buenos (SVM, ANN, LDA y DT), se debería considerar también la regla de Occam, que propone como más plausible al modelo más sencillo que se ajusta a los datos (Abu Mostafa *et al.*, 2012).

6.3.5. Marcadores químicos de identidad

Algunos de los métodos estudiados permiten detectar las variables de mayor peso en la discriminación. En la Tabla 6.29 se presentan esos clasificadores por método.

Tabla 6.29. Clasificadores con mayor poder discriminante por método usado en la clasificación por origen de jugos de naranja del noreste argentino, por zona productora (centro oeste de Corrientes COCR, centro sur de Misiones CSMN, noreste de Entre Ríos NEER y sudeste de Corrientes SECR)

Método	Clasificadores	Zona de producción
ANOVA/MANOVA	Sr y Ba	SECR
PCA	Ba, Cu, Cr, Mn, Ni, Sb, Sr y Zn	SECR
	Co y Mo	CSMN
LDA	Sr, Cu, Cd, K, Mo y Mn	SECR
	K, Cd, Mn y Sr	NEER
DT	Cd, Mg, Mo y Sr	COCR y NEER
	Ca y Mg	CSMN
	Mg y Sr	SECR
ANN	Cd, K, Al y Sr	COCR
	Cd, K, Cu, Sr, Mg y Cr	CSMN
	K, Mg, Ca y Cd	NEER
	Cd, K, Cu y Sr	SECR

Dentro de los elementos definidos por alguno de los métodos estudiados como marcadores químicos de identidad presentes en los jugos de naranja, Al, Ba, Ca, Mg, Mn, Mo, y Sr concuerdan con los hallados por García Ruiz *et al.* (2007) quienes lograron clasificar cidras de acuerdo con su origen, mediante el uso de LDA. Zhang *et al.* (2018) definió algunos elementos como buenos marcadores para identificar el origen geográfico de muestras de té, utilizando PCA y LDA, entre ellos, Cr y K, coinciden con los definidos en este Capítulo para clasificar jugos de naranjas.

Veljkovic *et al.* (2016) emplearon las concentraciones de 14 elementos para separar cuatro grupos de té, mediante el uso de PCA y Análisis Cluster, entre ellos Al, Ba, Ca, Cr, Cu, Mg, Mn, Mo, Ni, Sr y Zn resultados que concuerdan con los encontrados para jugos de naranjas.

Abdrabo *et al.* (2015) establecieron que Cd, Co, Cr y Ni pueden ser exitosamente aplicados para discriminar orígenes de dátiles, todos ellos también se han demostrado útiles en la clasificación de origen de jugos de naranjas.

Van der Linde (2008) clasificó vinos por su origen utilizando los contenidos de 4 elementos, Co, Cu, Mn y Ni todos ellos han sido detectados como buenos clasificadores por alguno de los métodos estudiados en este Capítulo.

Entre todos los elementos que han demostrado su capacidad para diferenciar jugos de naranja por su origen geográfico, se debe destacar al Sr que fue el elemento presente en las clasificaciones definidas por todos los métodos estudiados.

6.4. Resumen de resultados

La técnica FAAS permitió obtener información de contenidos de Ca, K, Fe, Mg, Mn, Na y Zn. Mediante la técnica de MP-AES se pudieron determinar en los jugos de naranja las concentraciones de Al, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Mo, Ni, Sb, Sr y Zn, las concentraciones de Pb y Se, se encontraban por debajo de los límites de detección (LOD) en todas las muestras.

Los jugos de naranja del NEA han podido ser caracterizados en función de los contenidos de 16 elementos minerales, según su origen y variedad. El elemento más abundante en jugos de naranja fue K con contenidos medios superiores a los 1000 μg/g, seguido de Mg con contenidos medios superiores a 100 μg/g. En orden de importancia le siguen Al y Ca, con contenidos medios entre 10 y 20 μg/g, Ba, Cu, Mn, Sr y Zn, con contenidos medios entre 1 y 10 μg/g. Mientras que Cd, Co, Cr, Fe, Ni, Mo y Sb, presentaron contenidos medios inferiores a 1 μg/g.

En relación con los elementos considerados nocivos para la salud y teniendo en cuenta los valores máximos establecidos en el Código Alimentario Argentino, las concentraciones de Cu en naranja 'Valencia late' superan los máximos permitidos.

Según los estándares del Codex Alimentarius (FAO-OMS), los valores promedio de las muestras analizadas en esta tesis se encontraron por debajo de los máximos permitidos, con excepción del Cd.

El PCA permitió detectar que los jugos de naranja provenientes del sudeste de Corrientes se encuentran asociados a mayores contenidos de Ba, Cu, Cr, Mn, Ni, Sb, Sr y Zn; los del centro sur de Misiones se pueden diferenciar de los del centro-oeste de Corrientes y noreste de Entre Ríos, por sus asociaciones con mayores contenidos de Co y Mo. Mientras que los jugos del centro-oeste de Corrientes y noreste de Entre Ríos presentan características muy similares, con mayores contenidos de Ca, Cd, Fe, K y Mg.

Según la técnica utilizada para la determinación de los contenidos de minerales en jugos de naranjas, el LDA permitió clasificar los jugos de naranja según su origen geográfico con un acierto del 94% y un índice κ = 0,94 (FAAS) y un acierto del 99% y un índice κ = 0,98 (MP-AES). Las funciones discriminantes se construyen principalmente con los contenidos de Mn y Zn (primera función) y Ca, K y Mg (segunda función) (FAAS) y con los contenidos de Cd, Cu, K y Sr (primera función) y de Cd, K y Mn (segunda función) (MP-AES).

Los DT permitieron generar reglas de decisión para diferenciar los jugos de naranja provenientes de las diferentes zonas, logran un 90% de acierto con un κ = 0,86 (información de FAAS) y un acierto de 99% con un κ = 0,98 (información de MP-AES). Con información de FAAS: K y Mg participan en la clasificación con menos del 65% y Na y Fe con valores alrededor del 75% y se destacan Zn y Mn con 100% de importancia en la determinación del origen de los jugos. Con información de MP-AES: Ca, Cd, K, Mg y Sr son los elementos más importantes en la clasificación.

La técnica de K-Vecinos más Cercanos no resultó ser la mejor opción para clasificar jugos de naranja en función de su contenido mineral, independientemente de la técnica utilizada para determinar el contenido elemental de los jugos estudiados, con un acierto de 67%, e índice κ = 0,55 (FAAS) y un acierto de 74%, e índice κ = 0,66 (MP-AES).

El uso de las Redes Neuronales permitió diferenciar los jugos de naranja provenientes de las diferentes zonas productoras, lográndose un acierto del 92% y un índice κ = 0,89 (FAAS) y un acierto del 99% y un índice κ = 0,98 (MP-AES). Dentro de los elementos con importancia en la discriminación por ANN, los que se encuentran presentes en mayor cantidad de zonas productivas son K, Mg, Mn, Na y Zn, (FAAS) y Cd, Cu, K y Sr (MP-AES).

Con las SVM se consiguió un acierto general de 90% y un índice κ de 0,87 (FAAS), y un acierto general de 99% y un índice κ de 0,99 (MP-AES).

El orden de las 5 técnicas de clasificación probadas según sus porcentajes de acierto se puede establecer, cuando se trabajó con datos obtenidos por medio de espectroscopía de absorción atómica de llama (FAAS): LDA 94% > ANN 92% > SVM = DT con 90% > KNN 67%. Con datos obtenidos por medio de espectroscopía de emisión atómica de plasma de microondas (MP-AES): SVM = 99% = ANN = DT = LDA 99% > KNN 74%.

Los elementos considerados marcadores químicos de identidad por alguna de las técnicas estudiadas fueron K, Mg, Mn, Na y Zn (FAAS) y Ca, Cd, Cu, K, Mg, Mn y Sr (MP-AES). El elemento Sr, es utilizado como clasificador por todas las técnicas utilizadas, cuando se trabajó con datos obtenidos por espectroscopía de emisión atómica de plasma de microondas (MP-AES).

Considerando todas las técnicas utilizadas se puede determinar que el acierto que se logra es algo mayor cuando se trabaja con concentraciones de elementos obtenidas mediante espectroscopía de emisión atómica por plasma de microondas (MP-AES) que cuando se utiliza información derivada de espectroscopía de absorción atómica de llama (FAAS), no obstante, el porcentaje de acierto logrado con información obtenida por ambas técnicas resulta adecuado para la propuesta de modelos clasificatorios que consideren un mayor número de muestras.

6.5. Referencias

AAFCO. 2020. Association of American Feed Control Officials 2020. Disponible: https://www.aafco.org/. Visita 12/02/2020.

Abdrabo, SS; Grindlay, G; Gras, L; Mora, J. 2015. Multi-element analysis of Spanish date palm (*Phoenix dactylifera* L.) by inductively couple plasma-based techniques. Discrimination using multivariate statistical analysis. Food Analytical Methods. 8: 1268-1278.

Abu Mostafa, YS; Magdon Ismail, M; Lin HT. 2012. Learning from Data: A Short Course: AMLBook.com.

Agustí, M. 2010. Citricultura. Ediciones Mundi-Prensa, Madrid, España. 507 pp.

Alonso Salces, RM; Herrero, C; Barranco, A; Berrueta, L; Gallo, B; Vicente, F. 2005. Classification of apple fruits according to their maturity stage by the pattern recognition analysis of their polyphenolic compositions. Food Chemistry. 93: 113-123.

Astuti, W; Dewanto, S; Soebandrija, KEN; Tan, S. 2018. Automatic fruit classification using support vector machines: a comparison with artificial neural networks. The 2nd International Conference on Eco Engineering Development (ICEED 2018). IOP Conference Series: Earth and Environmental Science 195 012047.

Belitz, HD; Grosch, W; Schieberle, P. 2009. Food chemistry, 4th revised and extended Ed. Springer Berlin Heidelberg. 421-425.

Cautela, D; Santelli, F; Boscaino, F; Laratta, B; Servillo, L; Castaldo; D. 2009. Elemental content and nutritional study of blood orange juice. Journal of the Science of Food and Agriculture. 89 (13): 2283-2291.

Chuku, LC; Chinaka, NC. 2014. Protein and mineral element levels of some fruit juices (Citrus spp.) in some Niger Delta areas of Nigeria. International Journal of Nutrition and Food Sciences. 3(6-1): 58-60.

Código Alimentario Argentino. 2020. Ley 18284. Disponible en línea: https://www.argentina.gob.ar/anmat/codigoalimentario. Visita 12/02/2020.

Concon, JM. 2009. Heavy metals in food. In: Food Toxicology, Part B: Contaminents and Additives. New York, Dekker. 3 (4): 1043-1045.

Drivelos, SA; Georgiou, CA. 2012. Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union. Trends in Analytical Chemistry. 40: 38-51.

Esslinger, S; Riedl, J; Fauhl-Hassek, C. 2014. Potencial y limitaciones de la toma de huellas dactilares no dirigida para la autenticación de alimentos en control oficial. International Food Research. 60: 189-204.

FAO. 2017. Citrus fruits statistics 2017. Disponible en línea: http://www.fao.org/economic/est/est-commodities/citricos/es/. Visita: 04/09/2019.

FAO-OMS. 2020. Codex Alimentarius. Disponible en línea: http://www.fao.org/fao-who-codexalimentarius/es/. Visita: 12/02/2020.

Farid, SM; Enani, MA. 2010. Levels of trace elements in commercial fruits juices in Jeddah, Saudi Arabia. Medicine Journal Islamic World Academy Science. 18 (1): 31-38.

Federcitrus. 2018. La actividad citrícola Argentina. Disponible en línea: https://www.federcitrus.org/. Visita: 04/09/2019.

FNB. Food and Nutrition Board (FNB), Institute of Medicine, National Academies, Washington, DC, USA. 2005. Dietary Reference Intakes (DRIs), Recommended Intakes for individual elements 2005. Disponible en línea: http://iom.edu/Activities/Nutrition/SummaryDRIs/~/media/Files/Activity%20Files/Nut

rition/DRIs/New%20Material/5DRI%20Values%20SummaryTables%.pdf. Visita: 10/03/2020.

Fu, L; Shi, S. 2019. A novel strategy to determine the compositions of inorganic elements in fruit wines using ICP-MS/MS. Food Chemistry. 299: 1-8.

Gaiad, JE; Hidalgo, MJ; Villafañe, RN; Marchevsky, EJ; Pellerano, RG. 2016. Tracing the geographical origin of Argentinean lemon juices based on trace element profiles using advanced chemometric techniques. Microchemical Journal. 129: 243-248.

García Ruiz, S; Modovan, M; Fortunato, G; Wunderli, S; García Alonso, JI. 2007. Evaluation of strontium isotope abundance ratios in combination with multi-elemental analysis as a possible tool to study the geographical origin of ciders. Analytica Chimica Acta. 590: 55-66.

Harmankaya, M; Gezgin, S; Ozcan, MM. 2011. Comparative evaluation of some macro- and micro-element and heavy metals contents in commercial fruit juices. Environ Monitoring Asses. 184: 5415-5420.

Hernández, A; Hansen, A. 2012. Uso de plaguicidas en dos zonas agrícolas de México y evaluación de la contaminación de agua y sedimentos. Revista Internacional de Contaminación Ambiental. 27(1): 115-127.

Hong, YS; Choi, JY; Nho, EY; Hwang, IM; Khan, N; Jamila, N; Kim, KS. 2019. Determination of macro, micro and trace elements in citrus fruits by inductively coupled plasma-optical emission spectrometry (ICP-OES), ICP-mass spectrometry and direct mercury analyzer. Journal of the Science of Food and Agriculture. 99 (4): 1870-1879.

JECFA. Joint FAO/WHO Expert Committee on Food Additives. 2006. Summary and conclusions of the sixty-seventh meeting of the Joint FAO/WHO Expert Committee on Food Additives p-1-11. Disponible en línea: ftp://ftp.fao.org/ag/agn/jecfa/jecfa67_final.pdf. Visita: 10/03/2020.

Khan, N; Choi, JY; Nho, EY; Hwang, IM; Habte, G; Khan, MA; Park, K; Kim, K. 2014a. Determination of mineral elements in milk products by inductively coupled plasma-optical emission spectrometry. Analytical Letters. 47: 1606-1613.

Khan, N; Jeong, IS; Hwang, IM; Kim, JS; Choi, SH; Nho, EY; Choi, JY; Park, KS; Kim, KS. 2014b. Analysis of minor and trace elements in milk and yogurts by inductively coupled plasma-mass spectrometry (ICP-MS). Food Chemistry. 147: 220-224.

Khan, N; Choi, JY; Nho, EY; Jamila, N; Habte, G; Hong, JH; Hwang, IM; Kim, KS. 2014c. Determination of minor and toxic elements in aromatic spices by micro-wave assisted digestion and inductively coupled plasma-mass spectrometry. Food Chemistry. 158: 200-206.

Lantz, B. 2015. Machine Learning with R: Packt Publishing.

Llorent Martínez, EJ; De Córdova, MLF, Ruiz Medina, A; Ortega Barrales, P. 2012. Analysis of 20 trace and minor elements in soy and dairy yogurts by ICP-MS. Microchemical Journal. 102: 23-27.

Lubinska-Szczygieł, M; Różańska, A; Namieśnik, J; Dymerski, T; Shafreen, R; Weisz, M; Ezra, A; Gorinstein, S. (2018). Quality of limes juices based on the aroma and antioxidant properties. Food Control. 89: 270-279.

Luykx, D & Van Ruth S. 2008. An overview of analytical methods for determining the geographical origin of food products. Food chemistry. 107 (2): 897-911.

Marcelo, MCA; Martins, CA; Pozebon, D; Dressler, VL; Ferro, MF. 2014. Classification of yerba mate (*Ilex paraguariensis*) according to the country of origin based on element concentrations. Microchemical Journal. (117): 164-171.

Molina, C; Ibañez, C; Gibon, FM. 2013. Proceso de biomagnificación de metales pesados en un lago hiperhalino (Poopó, Oruro, Bolivia): posible riesgo en la salud de consumidores. Ecología. 47(2): 99-118.

Mutsuko, HK; Fujii, S; Ono, A; Hirose, A; Imai, T; Ogawa, K; Ema, M; Nishikawa, A. 2011. Two-generation reproductive toxicity study of aluminum sulfate in rats. Reproduction and Toxicology. 31: 219–230.

Niu, L; Wu, J; Liao, X; Chen, F; Wang, Z; Zhao, G; Hu, X. 2008. Physicochemical characteristics of orange juices samples from seven cultivars. Agricultural Sciences in China. 7 (1): 41-47.

OMS. Organización Mundial de la Salud. 1980. Inorganic Mercury. IPCS Environment Health Criteria, 118(1), Ginebra (Suiza). 92 pp.

Palacios, J. 2013. Citricultura. Talleres Gráficos Alfa Beta S.A. ISBN: 9789874383266. 518 pp.

Perelman, SB, Garibaldi, LA; Tognetti, PM. 2019. Experimentación y Modelos Estadísticos. Ed. Facultad de Agronomía. Universidad de Buenos Aires. 475 pp.

Pérez, AL; Smith, BW; Anderson, KA. 2006. Stable isotope and trace elements profiling combined with classification models to differentiate geographic growing origin for three fruits: effect of subregion and variety. Journal of Agriculture and Food Chemistry. 54: 4506-4516.

Raja, OR; Sobhanardakani, S; Cheraghi, M. 2016. Health risk assessment of citrus contaminated with heavy metals in Hamedan city, potential risk of Al and Cu. Environmental Health Engineering Management J. 3:131-135.

Sabanci, K; Ünlersen, MF. 2016. Different apple varieties classification using k algorithms. International Journal of Intelligent Systems and Applications in Engineering. Special Issue-146967: 166-169.

Shimada, AM. 2005. Nutrición animal. 3 Ed. Trillas. México. 388 pp.

Simpkins, WA; Louie, H; Wu, Michael; Harrison, M; Goldberg, D. 2000. Trace elements in Australian orange juice and other products. Food Chemistry. 71: 423-433.

Singh, A; Sharma, RK; Agrawal, M; Marshall, FM. 2012. Health risk assessment of heavy metals via dietary intake of foodstuffs from the wastewater irrigated site of a dry tropical area of India. Food and Chemical Toxicology. 48 (1): 611-619.

Skoog, D; Agacharse, S; Holler, F. 2008. Principles of Instrumental Analysis: Cengage Learning Latin America.

Szymczycha Madeja, A; Welna, M; Jedryczko, D; Pohl, P. 2014. Developments and strategies in the spectrochemical elemental analysis of fruit juices. TrAC Trends in Analytical Chemistry. (55): 68-80.

Takaya, S; Rehmsmeier, M. 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. Plos One. Disponible en línea: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432. https://doi.org/10.1371/journal.pone.0118432. Visita: 02/05/2020.

Tufour, J; Bentum, J; Essumang, D; Koranteng Addo, J. 2011. Analysis of heavy metals in citrus juice from the Abura-Asebu-Kwamankese District, Ghana. Journal of Chemical and Pharmaceutical Research. 3 (2): 397-402.

Turra, C; Días de Lima, M; De Nadai Fernandes, EA; Arruda Bacchi, M; Barbosa, F; Barbosa, R. 2017. Multielement determination in orange juice by ICP-MS associated with data mining for the classification of organic samples. Information Processing in Agriculture. Accepted Manuscript. http://dx.doi.org/10.16/j.impa.2017.05.004.

USDA. 2020. United States Departament of Agriculture, Citrus Fruits USDA Foreign Agricultural Service. http://www.fas.usda.gov/commodities/citrus-fruit, Consultado: 12/02/2020.Waisberg, M; Joseph, P; Hale, B; Beyersmann, D. 2013. Molecular and cellular mechanisms of cadmium carcinogenesis. Toxicology. 3(4): 95-117.

Van der Linde, G. 2008. Multi-element analysis of South African wines un their provenance soils by ICP-MS and their classification according to geographical origin

using multivariate Statistics. Tesis Magister Scientiae of Chemistry in the Faculty of Science at the University of Johannesburg.

Veljkovic, JN; Pavlovic, AN; Brcanovic, JM; Mitic, SS, Tosik, SB; Pecev-Marinkovic, ET; Mitic, MN. 2016. Differentiation of black, green, herbal and fruit bagged teas based on multi-element analysis using inductively coupled plasma atomic emission spectrometry. Chemical papers. 70 (4): 488-494.

Vijayarekha, K; Govindaraj, R. 2006. Citrus fruit external defect classification using Wavelet Packet transform features and ANN. 1-4244-0726-2006-IEEE. 2872-2877.

Warrens, MJ. 2020. Kappa coefficients for dichotomous-nominal classifications. Advances in Data Analysis and Classification. Disponible en línea: https://link.springer.com/content/pdf/10.1007/s11634-020-00394-8.pdf. Visita 20/05/2020.

Weil, RR; Brady, N. 2017. The nature and properties of soils. 15th ed. Pearson Education.

Zhang, J; Yang, R; Chen, R; Li, YC; Peng, Y; Liu, C. 2018. Multielemental analysis associated with chemometric techniques for geographical origin discrimination of tea lives (*Camelia sinensis*) in Ghizou Province, SW China. Molecules. 23: 1-16.

CAPÍTULO VII

CONCLUSIONES GENERALES

Confirmando la hipótesis planteada en esta tesis, se han podido establecer modelos estadísticos multivariados y de aprendizaje automático basados en la composición mineral determinada por técnicas analíticas espectrométricas, que contribuyan a establecer sistemas de trazabilidad química de cítricos.

Se caracterizaron diferentes cítricos producidos en el norte argentino en función de los contenidos de elementos minerales en sus jugos.

En los jugos de limón se encontraron concentraciones mayores a 10 μg/g para Fe, Zn y Rb, entre 1 y 10 μg/g para Al, BA, Cu, Mn y Ni y menos de 1 μg/g para La, Cr, Se, Li, Mo, Co, Sn, Sc, V y Bi.

En los jugos de mandarina y naranja se presentaron perfiles similares, donde el elemento más abundante fue K (cercano a 1000 μg/g); seguido de Al, Mg y Ca con concentraciones mayores a 10 μg/g; Mn, Cu, Zn y Sr entre 1 y 5 μg/g; y Cd, Cr y Fe con contenidos menores a 1 μg/g.

Para los jugos de naranja se han descripto, además, los contenidos de elementos considerados nocivos para la salud. En general las concentraciones de los elementos analizados se encontraban dentro de los límites establecidos. Los contenidos de Cu de la variedad 'Valencia late' superaron los máximos establecidos en el Código Alimentario Argentino para metales pesados, y Pb superó los máximos en algunas muestras, no en promedio. El contenido promedio de Cd de los jugos de naranja superó el máximo permitido por la OMS y la FAO.

Entre los elementos nutricionalmente importantes, los jugos de frutas cítricas demostraron ser fuentes adecuadas de K, Ca, Zn y Mn.

Todos los instrumentos analíticos de medida utilizados se han demostrado eficaces para detectar la mayoría de los elementos estudiados.

Para los jugos de limón, el instrumento analítico de medida fue un espectrómetro de masas acoplado a plasma inductivo y el porcentaje de acierto en la clasificación de 4 sitios osciló entre 76 y 67% y de 3 provincias entre 95 y 100%. Demostrando la factibilidad del uso de la información multielemental combinada con métodos quimiométricos resulta útil para la propuesta de modelos capaces de determinar la procedencia geográfica de estos frutos.

En los jugos de mandarina se empleó la técnica de espectrometría de emisión atómica de plasma de microondas y se obtuvieron precisiones entre 85 y 96%. Demostrando, al igual que el caso anterior, la aptitud del uso de esta información para la propuesta de modelos clasificatorios de mandarinas, de acuerdo a su procedencia geográfica.

En el caso de jugos de naranja, cuando se empleó la técnica de espectroscopía de absorción atómica de llama, el acierto osciló entre 67 y 94% y con espectrometría de emisión atómica de plasma de microondas entre 74 y 99% según el método de análisis de datos utilizado. Estos resultados indican que la calidad de datos obtenidos mediante espectrometría de emisión atómica de plasma de microondas permite porcentajes de acierto mayores al momento de la clasificación por origen de los jugos de naranja, no obstante, los porcentajes de acierto obtenidos con ambos instrumentos de medida pueden considerarse adecuados, teniendo en cuenta la disponibilidad de equipos en laboratorios de mediana complejidad.

Se detectó la presencia de marcadores químicos de trazabilidad en limones del NEA y NOA y de mandarinas y naranjas del NEA, a partir de su composición multielemental.

La diferenciación de los jugos de limón entre regiones se puede establecer por los contenidos de Fe, La, V, Cu y Zn determinados por espectrometría de masas acoplada a plasma inductivo.

En los jugos de mandarina se han encontrado patrones para separar las muestras según la zona productora basados en los contenidos de Al, Ca, Cd, Cr, Cu, K, Mg, Mn, Sr y Zn, determinados espectrometría de emisión atómica de plasma de microondas.

En el caso de los jugos de naranja, con información de espectroscopía de absorción atómica de llama los elementos que han demostrado capacidad como marcadores químicos de identidad fueron K, Mg, Mn, Na y Zn. Cuando los datos se obtuvieron mediante espectrometría de emisión atómica de plasma de microondas estos elementos fueron Ca, Cd, Cu, K, Mg, Mn y Sr.

Se demostró la eficacia de los métodos de análisis multivariados y de aprendizaje automático, para proponer modelos para autenticación o confirmación de identidad de los frutos.

Las técnicas probadas pueden ordenarse según los porcentajes de acierto:

Para los jugos de limón:

SVM 76% > RF 71% > LDA = PLS-DA = KNN 67% (4 sitios).

SVM 100% > RF = LDA = PLS-DA = KNN 95% (3 provincias).

En los jugos de mandarina:

ANN 96% > SVM 94% > DT 91% > LDA 86% > KNN 85%.

En el caso del jugo de naranja, según la metodología de análisis químico:

LDA 96 % > ANN 92% > SVM = DT 90% > KNN 67% (FAAS); y

SVM 99% > ANN = DT = LDA 99% > KNN 74% (MP-AES).