Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

Trabajo Final de Maestría en Tecnologías de la Información

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales

Autor: Hilda Rosana Alcantre

Director: Dra. María del Carmen Montserrat la Red Martínez

Co-Director: Dra. Sonia Itatí Mariño

Año 2024

Dedicatoria

A mi familia por su amor, paciencia y acompañamiento, mis amados hijos Anto, Mica, Gonza y mi esposo Gustavo.

A mi padre Victor que me guía desde el cielo, mi gran maestro, fuente de inspiración y sabiduría.

A mi madre Hilda por ser un ejemplo a seguir, por su amor incondicional y buenas enseñanzas.

A mis hermanos Pedrín, Fabian, Lily y Anto, que con su cariño de siempre me motivan a seguir adelante.

A mis hermosos sobrinos y sus padres, por tantas alegrías compartidas en familia.

A mis amigos y personas queridas que me brindaron su apoyo y motivación en este camino.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Resumen

En las últimas décadas, ante la necesidad de analizar grandes cantidades de datos digitalizados en bases de datos de las organizaciones, se produjo un gran avance en el campo de la explotación de datos, especialmente mediante técnicas de minería de datos. Estos adelantos fueron favorecidos por el desarrollo de las tecnologías de la información, las que permitieron su aplicación en diversas áreas de investigación.

Como resultado de este avance, han surgido numerosas herramientas que permiten el tratamiento y análisis de la información con métodos descriptivos y predictivos potentes, que facilitan la producción de conocimiento y que sirven de apoyo a la toma de decisiones.

En este Trabajo Final de Maestría se propone el diseño de un procedimiento, basado en una metodología, que permite sistematizar las tareas de explotación de información, para estudiar el comportamiento de una Población Económicamente Activa.

Se determinarán los procesos de explotación de información apropiados, que serán aplicados a un conjunto de datos estructurados de población. A través de enfoques supervisados y no supervisados se podrán identificar grupos dentro del conjunto de datos, crear modelos que representen a esos grupos y hallar patrones que ayuden a obtener conocimiento sobre los datos, mediante técnicas de minería de datos.

Palabras claves: *Explotación de la Información, Minería de Datos, ISO, Procedimientos*

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Abstract

In recent decades, given the need to analyze large amounts of digitized data in organizational databases, there has been a great advance in the field of data mining, especially through data mining techniques. These advances were favored by the development of information technologies, which allowed their application in various areas of research.

As a result of this progress, numerous tools have emerged that allow the treatment and analysis of information with powerful descriptive and predictive methods, which facilitate the production of knowledge and support decision making.

This Final Master's work proposes the design of a procedure, based on a methodology, that allows systematizing the tasks of information mining, to study the behavior of an Economically Active Population.

Appropriate information mining processes will be determined and applied to structured population data set. Through supervised and unsupervised approaches it will be possible to identify groups within the dataset, create models that represent those groups and find patterns that help to obtain knowledge about the data, using data mining techniques.

Keywords: *Information Mining, Data Mining, ISO, Procedure*

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Agradecimientos

Agradezco a Dios y la Virgen María por bendecir mi vida con mi amada familia y amigos, por guiar mis decisiones cada dia, por darme paciencia y fortaleza en los momentos difíciles.

A mi directora Dra. Montserrat la Red Martínez, por su ayuda y por haberme guiado en este trabajo.

A mi co-directora Dra. Sonia Mariño por su valiosa ayuda y colaboración en todo momento.

A mi estimada colega Lic. Gabriela González por su aporte a este Trabajo Final de Maestría.

A mis profesores de la maestría, por su acompañamiento y guía en todo el desarrollo de esta carrera.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Índice de contenidos

Índice de Tablas

Índice de figuras

Lista de abreviaturas

A.C.	Análisis de Correspondencia (Correspondence Analysis).
ASD-BI	Adaptive Software Development – Business Intelligence (Desarrollo de Software Adaptativo para Inteligencia de Negocios).
BPMN	Business Process Model and Notation (Modelo y Notación de procesos de Negocio).
CRISP-DM	CRoss-Industry Standard Process for Data Mining (Proceso Estándar para Minería de Datos Multi Industria). Es una metodología estudiada, probada y ampliamente utilizada para el desarrollo de proyectos de Explotación de Información.
E.I.	Explotación de la Información.
EPH	Encuesta Permanente de Hogares.
FCE	Facultad de Ciencias Económicas.
INDEC	Instituto Nacional de Estadísticas y Censos.
ISO/IEC	International Organization for Standardization / International Electrotechnical Commission (Organización Internacional de Normalización / Comisión Electrotécnica Internacional).
OIT	Organización Internacional del Trabajo.
OMG	Object Management Group (Grupo de Gestión de Objetos).
PEA	Población Económicamente Activa.
SEMMA	Sample, Explore, Modify, Model and Assess (Muestra, Exploración, Modificación, Modelado y Evaluación). Enfoque utilizado para el desarrollo de proyectos de Explotación de Información.
SOM	Self-Organizing Maps (Mapas Auto-Organizados).
TDIDT	Top-Down Induction Decision Trees (Inducción descendente de Árboles de Decisión).
TDSP	Team Data Science Process (Proceso de Ciencia de Datos en Equipo).
TFM	Trabajo Final de Maestría.
UNNE	Universidad Nacional del Nordeste.
UML	Unified Modeling Language (o Lenguaje de Modelado Unificado).

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Capítulo 1

Introducción

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

1. Introducción

En este capítulo, se realiza una introducción al Trabajo Final de Maestría (TFM). En la primer sección se realiza la contextualización del TFM (sección 1.1); en la segunda sección se expone la motivación que ocasionó la elección de la temática de este trabajo (sección 1.2); en la siguiente sección se presentan los problemas que originaron la realización de este TFM (sección 1.3); luego, se definen objetivos de este trabajo (sección 1.4) y finalmente, se explica como está organizado este TFM (sección 1.5).

1.1. Contextualización del Trabajo Final de Maestría

Las actividades que se realizan en el presente trabajo corresponden al área de conocimiento de las Ciencias de la Computación. En particular, la definición de un procedimiento [1], para sistematizar las actividades en un proyecto de explotación de información, mediante el uso de una metodología.

El tema propuesto en este TFM se centra en el diseño de un procedimiento de explotación de información, para detectar problemáticas laborales y sus factores de incidencia en una Población Económicamente Activa, mediante minería de datos.

La Población Económicamente Activa (PEA) o Fuerza de Trabajo [2], está compuesta por "todas las personas que aportan su trabajo (lo consigan o no) para producir bienes y servicios económicos, definidos según y como lo hacen los sistemas de cuentas nacionales durante un período de referencia determinado". En la EPH (Encuesta Permanente de Hogares), para medir este concepto, se considera como parte de la PEA a "todas las personas de 10 años y más que en un período de referencia corto tienen trabajo y aquellos que sin tenerlo están disponibles y buscan activamente un trabajo. Son parte de la PEA tanto los *ocupados* como los *desocupados*".

El tesauro de la Organización Internacional del Trabajo (OIT) [3] define al trabajo como el "conjunto de actividades humanas, remuneradas o no, que producen bienes o servicios en una economía, o que satisfacen las necesidades de una comunidad o proveen los medios de sustento necesarios para los individuos".

En este contexto, partiendo del concepto de trabajo como una actividad que genera bienes o servicios para el mercado, se consideran *ocupadas* [2] a "todas las personas que tengan cierta edad especificada (10 años o más) y que durante un período de referencia (una semana) hayan trabajado al menos una hora. Se incluye a: a) las personas que durante el período de referencia realizaron algún trabajo de al menos

una hora, hayan recibido pago (en dinero o en especie) o no por dicha actividad. b) las personas que tienen una ocupación pero que no estaban trabajando temporalmente durante el período de referencia y mantenían un vínculo formal con su empleo. El grupo de los *desocupados* está conformado por todas aquellas personas que sin tener trabajo se encuentren disponibles para trabajar y han buscado activamente una ocupación en un período de referencia determinado".

En este contexto, el objeto de estudio de este TFM se centra en el diseño de un procedimiento de explotación de la información que permita sistematizar el trabajo a realizar en un proyecto de explotación de la información mediante una metodología que ayude a obtener conocimiento en conjuntos de datos estructurados de población, y posibilite la detección de problemáticas laborales y sus factores de incidencia en una PEA, mediante la aplicación de distintas técnicas de minería de datos.

Se espera que el procedimiento diseñado facilite la ejecución de proyectos de explotación de la información y la evaluación de los resultados obtenidos, y se pueda aplicar a otras poblaciones similares, previa adaptación a los datos disponibles y en relación con cada problema tratado.

Para lograr esto, se estudiarán métodos existentes en la literatura para el desarrollo de proyectos de explotación de la información, que posibiliten el diseño del procedimiento, ayuden a ordenar el trabajo realizado y sirvan de guía en la ejecución de este tipo de proyectos.

1.2. Motivación

En la actualidad, el mercado laboral global enfrenta varios desafíos, entre los que destacan el aumento de la desocupación y la informalidad laboral. En Argentina, y específicamente en la ciudad de Corrientes, estas problemáticas son igualmente significativas (en el Anexo A se aporta información sobre el mercado de trabajo en Argentina. Tasas e indicadores socioeconómicos). Esto, sumado a la escasez de estudios en la región de referencia que utilicen herramientas de explotación de información para conocer la realidad de estas poblaciones y atender las problemáticas que se presentan, motiva el presente trabajo.

En este TFM se propone enfocar el estudio en el diseño de un procedimiento de explotación de la información que facilite y guie la ejecución de proyectos de explotación de la información y permita, mediante la aplicación de técnicas de minería de datos, obtener conocimiento de los datos o patrones de comportamiento que ayuden a detectar problemáticas laborales en estudios de población.

En un proyecto de explotación de la información los datos utilizados o datos de entrada generalmente provienen de diferentes bases de datos y su principal objetivo es encontrar información oculta o implícita en esos repositorios, lo que sería dificil obtener mediante métodos estadísticos convencionales [4]. Los proyectos de explotación de la información permiten lograr esto, debido a que proporcionan las herramientas necesarias para convertir los datos de una base de datos en conocimiento útil, para la toma de decisiones.

En este contexto, en este trabajo se estudian distintas metodologías que puedan ser aplicadas a un proyecto de explotación de la información y se analizan enfoques *supervisados y no supervisados* que puedan ser utilizados en estudios sobre PEA en zonas urbanas. Asimismo, se indaga sobre notaciones estandarizadas que permitan diseñar los procedimientos de explotación de la información.

1.3. Problema

En esta sección se describe la problemática que se aborda en este trabajo y la justificación de las decisiones tomadas para llevar a cabo su resolución. Se finaliza aportando los objetivos que guian el desarrollo de este TFM.

En los últimos años, cada vez más empresas desarrollan proyectos de explotación de la información, pero en muchos casos se identifican deficiencias en la ejecución exitosa de los mismos debido a la complejidad creciente en este tipo de proyectos. Es por eso que, la comunidad científica trabaja continuamente en el perfeccionamiento de diferentes metodologías y herramientas que permitan mejorar la calidad en el proceso de explotación de la información, a fin de obtener resultados exitosos [5].

Como se comentó anteriormente, en la actualidad son escasos los estudios en la región que sirvan como guía en la aplicación de procesos de explotación de la información que permitan obtener conocimiento a partir de un conjunto de datos estructurados de PEA. Los procesos de explotación de información permiten obtener patrones y tendencias en grandes cantidades de datos; estos patrones o modelos pueden aportar información útil o conocimiento sobre el conjunto de datos estudiado.

En este contexto y ante la necesidad de lograr resultados confiables en proyectos de explotación de la información aplicados a estudios de población, en este TFM interesa conocer y analizar metodologías que posibiliten la aplicación de procesos de

explotación de información sobre datos estructurados de manera ordenada, a fin de obtener resultados exitosos en el proyecto y el logro de los objetivos planteados.

Por otra parte, dada la importancia de documentar las tareas y actividades que se llevan a cabo en un proyecto de explotación de la información, se considera necesario indagar acerca de estándares para el diseño de procedimientos, que permitan documentar las actividades realizadas a fin de poder aplicar los conocimientos adquiridos, a casos de estudio similares, previa adaptación de datos y procesos de explotación de información a los problemas estudiados.

En este sentido, se espera que los procedimientos diseñados faciliten la tarea de selección de herramientas, algoritmos y métodos para la detección de patrones en conjuntos de datos de población.

El alcance del procedimiento es descubrir patrones de comportamiento o conocimiento en un conjunto de datos, mediante distintas técnicas de minería de datos, que serán seleccionadas acorde a la problemática planteada y al conjunto de datos disponible. Asimismo, se pretende que los procedimientos generados ayuden a la evaluación de los procesos y resultados obtenidos que serán documentados, visualizados y analizados con la opinión de expertos a fin de obtener las interpretaciones y conclusiones finales que contribuyan al estudio realizado y la mejora de los procesos, procedimientos y resultados obtenidos.

En relación al problema planteado, se establecen los siguientes objetivos para este TFM:

1.4. Objetivo general

Diseñar un procedimiento de explotación de la información para la detección de problemáticas laborales y sus factores de incidencia, basado en normas internacionales, que permitan establecer pautas para orientar estudios sobre la población económicamente activa en zonas urbanas.

1.5. Objetivos específicos

Se definen los siguientes objetivos específicos como una guía de pasos a seguir para el logro del objetivo general de este TFM:

1)- Indagar en la literatura sobre metodologías o modelos de procesos que se puedan aplicar a la explotación de datos estructurados.

2)- Analizar en la bibliografía que métodos y herramientas de explotación de información se utilizan en distintos estudios, para detectar problemáticas

laborales y sus factores de incidencia en conjuntos de datos estructurados de población.

3)- Examinar estándares para el diseño de procedimientos que permitan establecer el ordenamiento de las actividades requeridas en un proyecto de explotación de la información aplicado a datos estructurados.

4)- Averiguar sobre técnicas o métricas que permitan validar el procedimiento de explotación de la información con datos estructurados y evaluar los modelos obtenidos con algoritmos de minería de datos.

1.6. Organización del TFM

A fin de lograr los objetivos establecidos, el presente trabajo se ha estructurado de la siguiente manera:

En el **Capítulo 1** se realiza una *introducción* a este TFM, una descripción de la motivación que ocasiono la elección de la temática de este trabajo y los problemas que se abordarán en el mismo; además se exponen los objetivos del TFM y al final del capítulo se explica cómo está organizado este trabajo.

En el **Capítulo 2** se realiza una descripción que sintetiza el estado de la cuestión de los conocimientos disciplinares específicos sobre los que se trabaja en este TFM; se explica qué es el Aprendizaje Automático y se hace referencia a distintos tipos de aprendizaje. Luego, se define la Explotación de información y los procesos de explotación de información. Seguidamente, se introducen algunos conceptos fundamentales relacionados con la explotación de la información, como la definición de Ingeniería de Explotación de Información, los roles que se pueden establecer en un proyecto de explotación de la información; se analizan los términos Explotación de la Información, Minería de Datos y Ciencia de Datos; se describen los principales enfoques que se utilizan en explotación de información y se detalla la metodología CRISP-DM. Finalmente, se describen algunas técnicas de minería de datos, los software que se pueden usar en minería de datos y se analizan los estándares internacionales que serán utilizados en este TFM.

En el **Capítulo 3** de este TFM se expone el marco metodológico usado en el desarrollo de este trabajo, estructurado en tres etapas.

En el **Capítulo 4** se presenta la solución propuesta a la problemática planteada y se explican los aspectos generales de la misma. Se expone un modelo general a usar en el diseño del procedimiento de explotación de la información y una propuesta de aplicación del procedimiento diseñado, que consta de un procedimiento principal

(adaptación de CRISP-DM) y de procedimientos específicos planteados en relación con cada objetivo de minería de datos definido y proceso de explotación de información identificado. Finalmente, se valida el procedimiento diseñado mediante la experimentación con datos reales, en la que se explica el desarrollo del proyecto de explotación de la información mediante una adaptación de la metodología CRISP-DM, se especifican las tareas a realizar en cada fase de la metodología y se exponen los resultados obtenidos.

En el **Capítulo 5** se exponen las conclusiones en relación a los objetivos planteados y los resultados obtenidos a través de la experimentación, mediante la aplicación del procedimiento diseñado en este TFM.

Se finaliza este apartado con los aportes realizados en este TFM y las futuras líneas de investigación que se pueden desarrollar a partir del trabajo realizado.

Luego, se presentan las referencias o lista de todas las publicaciones consultadas para el desarrollo de este TFM.

Al final del trabajo se presenta un **Anexo** en el que se expone:

El **Anexo A** presenta información sobre el mercado de trabajo en Argentina. Tasas e indicadores socioeconómicos.

El **Anexo B** presenta información sobre el relevamiento realizado en el Barrio Industrial, de la ciudad de Corrientes. Incluye los formularios utilizados para realizar el relevamiento, mapas e información recolectada sobre el Barrio industrial de la ciudad de Corrientes (caso de estudio).

Anexo C: Presenta un glosario con terminología utilizada en este TFM.

Anexo D: Otros enfoques utilizados en explotación de la información.

Anexo E: Datos del Relevamiento del Barrio Industrial.

Capítulo 2

Estado de la Cuestión

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

2. Estado de la cuestión

En este capítulo se realiza una revisión del estado de la cuestión respecto a los contenidos disciplinares específicos que se trabajan en este TFM, a fin de aportar los conocimientos necesarios para comprender la problemática abordada y el porqué de la solución planteada. En la primera sección (sección 2.1), se explica qué es el Aprendizaje Automático y se hace referencia a distintos tipos de aprendizaje. Luego se detalla en qué consiste la Explotación de información y los procesos de explotación de información (sección 2.2). En la sección siguiente, se define la Ingeniería de Explotación de la Información y se comentan los roles que se pueden establecer en un proyecto de explotación de la información (sección 2.3). Posteriormente, se hace un análisis de los términos explotación de información, minería de datos y ciencia de datos en relación con estudios de distintos autores (sección 2.4); se describen los principales enfoques que se utilizan en explotación de la información (sección 2.5); seguidamente, se detalla la metodología CRISP-DM (sección 2.6) y finalmente, se describen algunas técnicas de minería de datos (sección 2.7), se hace una breve reseña sobre los software que se pueden usar en minería de datos (sección 2.8), y se analizan los Estándares Internacionales que serán utilizados en este TFM (sección 2.9).

2.1. Aprendizaje automático

En [6] se define el aprendizaje automático (*machine learning* o aprendizaje de máquina) como el "campo de estudio que brinda a las computadoras la capacidad de aprender sin estar programadas explícitamente". El aprendizaje automático impulsó los últimos avances en inteligencia artificial y proporciona tecnologías diversas como ser herramientas prácticas para analizar datos y hacer predicciones [7].

En [8], se define el aprendizaje de máquina como un área de la ciencia de la computación dedicada al desarrollo y aplicación de técnicas y algoritmos computacionales que pueden aprender y perfeccionar sus aprendizajes a través de la experiencia y la adaptación a condiciones cambiantes; de esta manera, aumentan el rendimiento de las máquinas y su capacidad de inferir y obtener nuevos conocimientos de grandes conjuntos de datos. El aprendizaje de máquina permite manejar e interpretar grandes cantidades de datos mediante el uso de diferentes métodos estadísticos, matemáticos y lógicos. En los últimos años, el aprendizaje de máquina ha tenido mucho protagonismo debido al aumento en la digitalización de datos y procesos [9].

Las técnicas de aprendizaje de máquina o algoritmos de aprendizaje automático se pueden clasificar en *supervisados*, *no supervisados y semi-supervisados* [10], cada una de las cuales tiene una función diferente según el tipo de datos con el que se trabaja y el objetivo o análisis que se realiza.

2.1.1. Aprendizaje supervisado: La experiencia de algoritmos de aprendizaje supervisado contiene un conjunto de características y el ejemplo está asociado con una etiqueta u objetivo. El término aprendizaje supervisado surge de la visión del objetivo o salida proporcionada por un instructor que muestra al sistema de aprendizaje lo que debe hacer. En el caso del aprendizaje no supervisado, no hay instructor que indique la salida, entonces el algoritmo debe aprender a comprender los datos sin esta guía. Ambos tipos de aprendizaje (supervisado y no supervisado), no se definen formalmente, esto lleva a que se puede resolver un problema aparentemente no supervisado, dividiéndolo en problemas de aprendizaje supervisados.

La mayoría de los algoritmos de aprendizaje automático, experimentan con un conjunto de datos (dataset) que consiste en una colección de ejemplos, que tienen determinadas características. Por ejemplo, una tabla donde cada fila representa un ejemplo diferente y cada columna de la matriz corresponde a una característica diferente. Es decir, se tiene un conjunto de *p* características X1, X2, ..., X*p*, medidas en *n* observaciones, y una respuesta "Y" también medida en esas mismas *n* observaciones. El objetivo en los problemas supervisados, es intentar predecir algún vector de resultado "Y", usando X1, X2, ..., X*p*. Si se ajusta un modelo predictivo usando una técnica de aprendizaje supervisado, se puede verificar qué tan bien predice el modelo la respuesta "Y" en observaciones no utilizadas para el ajuste del modelo. También se llaman técnicas predictivas y se basan en algoritmos que necesitan saber cuál es la salida correcta para un conjunto de datos. Es decir, se tiene idea de que existe una relación entre la entrada y la salida; se conoce previamente datos etiquetados para ser entrenados (conjunto de entrenamiento) y a partir de este entrenamiento se puede predecir la etiqueta de un nuevo dato de entrada sin etiqueta (conjunto de prueba) [8].

Dentro de estas técnicas se encuentran diferentes tipos de algoritmos como clasificación, regresión, máquinas de vectores de soporte (SVM) y árboles de decisión.

2.1.2. Aprendizaje *no supervisado:* Los algoritmos de aprendizaje no supervisado, experimentan con un conjunto de datos que contiene muchas características y aprende propiedades útiles de la estructura de este conjunto de datos [10]. Es decir, en este tipo de aprendizaje no se conoce la salida esperada, pero sí se conocen algunas características o propiedades de los datos examinados. Implica la observación de varios ejemplos de un vector aleatorio x, e intenta aprender en forma implícita o explícita, la distribución de probabilidad p(x) o algunas de sus propiedades interesantes. También se conoce como aprendizaje descriptivo, porque permite obtener patrones y conocimientos a partir de las características intrínsecas de los datos [8].

En [11] se compara a este tipo de aprendizaje con el método que utilizan los bebes para aprender a hablar. Es decir, en un principio escuchan hablar a los padres y no entienden nada, pero luego de escuchar miles de conversaciones, su cerebro formará un modelo sobre cómo funciona el lenguaje, comienzan a reconocer patrones y a esperar determinados sonidos.

En este tipo de aprendizaje, se tiene un conjunto de características X1, X2, ..., X*p* medidas en *n* observaciones. Como en el aprendizaje *no supervisado* no se tiene una variable de respuesta asociada Y, entonces, no interesa la predicción sino que el objetivo en este caso es descubrir cosas interesantes en ese conjunto de datos u observaciones [12]. Este tipo de aprendizaje suele ser más desafiante que el aprendizaje *supervisado*, porque no hay un objetivo o no se conoce la verdadera respuesta que permita verificar el trabajo realizado. Al no existir una forma estándar de realizar la validación, es más *subjetivo* y puede ser difícil evaluar los resultados obtenidos, por lo que suele realizarse como parte de un análisis de datos exploratorio.

En otras palabras, este tipo de algoritmo consiste en agrupar una serie de vectores según un criterio de similitud en distintos grupos o clusters, es decir, agrupa vectores similares entre si [13]. Cuando se agrupan las observaciones de un conjunto de datos en grupos distintos, se busca que las observaciones de cada grupo sean muy similares entre sí y las observaciones en grupos diferentes sean muy diferentes entre sí. Establecer que dos o más observaciones sean similares o diferentes, tiene que ver con el dominio, es decir, se debe realizar en base al conocimiento de los datos que se están estudiando [12]. Estas técnicas permiten, a través de la información que se tiene de los datos, identificar la estructura de los grupos, tendiendo a la mayor

similitud dentro de los grupos (intra-*clusters*) y la mayor diferencia entre los miembros que pertenecen a otros [10]. Existe una gran cantidad de métodos de agrupación. Los dos enfoques de agrupación en clúster más conocidos son: agrupación de K-medias y agrupación jerárquica.

2.1.3. Aprendizaje *semi-supervisado:* Estos algoritmos reciben un conjunto de datos, donde una parte está *etiquetada* y la otra parte está *sin etiquetar*, para construir modelos de descripción y predicción de los datos**.** Estas técnicas son un hibrido entre las técnicas supervisadas y no supervisadas, combinan diferentes algoritmos y técnicas independientes en un solo modelo, con la finalidad de mejorar la predicción de los algoritmos. Esto se debe a que las diferentes técnicas presentan algún tipo de debilidad o sesgo para identificar patrones específicos o trabajar con un tipo datos. Mediante la combinación de diferentes tipos de algoritmos se pueden obtener modelos predictivos más sólidos, en comparación con la utilización de un solo algoritmo [8].

2.1.4. Aprendizaje profundo: Este tipo de aprendizaje se puede aplicar a diferentes áreas de la ciencia, motivo por el cual se ha convertido en una de las áreas del aprendizaje de máquina que más ha llamado la atención de la comunidad científica en los últimos años.

El aprendizaje profundo se basa en la aplicación de redes neuronales artificiales de varias capas, que pueden encontrar e inferir patrones complejos dentro de los datos. Estos algoritmos intentan imitar el comportamiento del cerebro humano, tratando de reproducir los procesos de conexión de las neuronas. Son redes hábiles para aprender tareas complejas a partir de las entradas recibidas y se adaptan a los cambios de las entradas. El aumento de la capacidad computacional ocurrido en las últimas décadas ha favorecido el desarrollo de las redes neuronales artificiales y el aprendizaje profundo. Los algoritmos de aprendizaje profundo se pueden clasificar en *supervisados* y *no supervisados*, al igual que el aprendizaje de máquina, en relación a su capacidad de predecir o describir grandes conjuntos de datos [10].

2.1.5. Aprendizaje por refuerzo: Este algoritmo de aprendizaje automático no se ajusta a la experiencia de un conjunto de datos fijo, sino que interactúa con un entorno, aprende observando el mundo que le rodea. El sistema aprende a base de ensayo-error y su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones**.**

2.2. Explotación de información

Las terminologías usadas por distintos autores para referirse a la extracción de conocimiento de un set de datos mediante el uso de distintas tecnologías, ha cambiado en las ultimas décadas. En [14], se define la Explotación de Información como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información.

En un estudio publicado en [4] se define a la Explotación de Información como una sub-disciplina Informática que aporta a la Inteligencia de Negocio (Business Intelligence o BI), las herramientas que se necesitan para transformar la información en conocimiento. Además, se define la Inteligencia de Negocio como las estrategias y herramientas orientadas a la gestión y creación de conocimiento a partir del análisis de datos de una organización, que se refieren a software que permiten analizar y exhibir los datos. Asimismo, explica que la explotación de información basada en sistemas inteligentes se refiere a la aplicación de estas tecnologias (como ser los algoritmos TDIDT, SOM, etc.), para descubrir y enumerar patrones presentes en la información.

La Explotación de la Información tiene como principal objetivo encontrar información oculta en un conjunto de datos, que no se puede obtener con los métodos estadísticos convencionales [15]. En este estudio, el autor se refiere a los términos “Minería de datos” y “Explotación de la información” como procesos de extracción de conocimiento de bases de datos. Según el punto de partida del proceso, los proyectos de explotación de la información se puede clasificar en:

1)- Proyectos en los que se aplica la minería de datos a una situación organizacional (un problema o una oportunidad), en los que se buscan patrones y relaciones que puedan colaborar con la misma.

2)- Situaciones en las que el proyecto comienza con un conjunto de datos y el objetivo es explorarlos para encontrar relaciones interesantes que puedan ser útiles en el dominio de aplicación.

La Explotación de la Información es una disciplina que engloba un conjunto de técnicas que permiten la extracción de conocimiento de los repositorios de una organización y permite solucionar problemas de predicción, clasificación y segmentación [4].

2.2.1. Procesos de explotación de la información

Un proceso de explotación de información se define como un grupo de tareas relacionadas lógicamente, que ejecutadas sobre los datos o información de una organización, permite obtener información de mayor importancia para la misma. Para un analista de datos, en general no son los datos lo más importante, sino el conocimiento que se puede hallar en los mismos, siempre que los modelos encontrados reflejen la realidad, sean válidos y aporten información útil para la toma de decisiones de una organización [16], o el logro de objetivos planteados.

En cada proceso de explotación de información se especifica un conjunto de datos de *entrada*, las transformaciones que sufren esos datos y las *salidas* de información que resultan del *proceso*. Este proceso puede ser parte de un proceso mayor o bien incluir otros procesos [4]. En los procesos de explotación de información se aplican técnicas de minería de datos en su mayoría provenientes del campo del Aprendizaje Automático [17], [5]. Luego de identificar un problema de inteligencia de negocio y las técnicas de explotación de información que se van a usar, un proceso de explotación de información permite detallar las tareas a desarrollar, para que con la aplicación de las técnicas de explotación de información, se solucione el problema planteado. Se puede decir entonces, que la inteligencia de negocio aporta el *problema*, la explotación de información las *tecnologías* a utilizar y los procesos de explotación de información indican como usar esas tecnologías para abordar el problema identificado.

En [18] se realiza un estudio en el que se caracterizan los procesos de explotación de información, se identifican las tecnologías asociadas a tales procesos y se validan los mismos en distintos dominios. Se definen cinco procesos de explotación de información que se pueden considerar en un proyecto de explotación de información, los que a su vez especifican las técnicas o algoritmos que se pueden usar en base a las características y necesidades del problema de minería de datos. Estos procesos son:

1)- **Descubrimiento de reglas:** Este proceso se usa cuando se quiere identificar condiciones para obtener resultados del dominio del problema. Por ejemplo, se puede usar para establecer las características de los clientes con alto grado de fidelidad a la marca, o descubrir las características de las personas que pertenezcan a una determinada clase. Se propone el uso de algoritmos de

inducción TDIDT para descubrir las reglas de comportamiento de cada atributo clase.

2)- **Descubrimiento de grupos:** Este proceso es útil en los casos en que se necesita identificar una partición en el conjunto de información disponible sobre el dominio de un problema. Por ejemplo, puede ser utilizado para la identificación de grupos sociales con las mismas características, en un conjunto de datos de población. Los autores proponen utilizar en este proceso, los algoritmos de agrupamiento (*clustering*).

3)- **Ponderación de interdependencia de atributo (o descubrimiento de atributos significativos):** Este proceso es útil cuando se quiere identificar los factores que poseen mayor incidencia (o frecuencia de ocurrencia) sobre un determinado resultado de un problema. Es decir, permite ponderar en qué medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase. Por ejemplo, se puede aplicar para la determinación de factores que poseen incidencia sobre las ventas; o para determinar los factores (o variables) que tienen mayor incidencia sobre un atributo clase (grupo). En este proceso se propone la utilización de Redes Bayesianas.

4)- **Descubrimiento de reglas de pertenencia a grupos:** Se puede aplicar este proceso cuando se requiere identificar las condiciones de pertenencia a cada clase en una partición desconocida, pero que se encuentra presente en la masa de información disponible sobre el dominio del problema. Por ejemplo, se puede usar este tipo de proceso cuando se necesita realizar la agrupación de personas por edades y determinar el comportamiento de cada grupo, en un conjunto de datos de población. Los autores proponen el uso de algoritmos de agrupamiento (como SOM), para encontrar grupos y luego, se utilizan algoritmos de inducción (TDIDT) para establecer las reglas de pertenencia a cada uno.

5)- **Ponderación de reglas de comportamiento o de la pertenencia a grupos**: Se utiliza cuando se requiere identificar las condiciones con mayor incidencia sobre la obtención de un determinado resultado en el dominio del problema, ya sea por la mayor medida en la que inciden sobre su comportamiento o las que mejor definen la pertenencia a un grupo. Por ejemplo, se puede aplicar este tipo de proceso a la identificación del factor más predominante que incide en el alza o baja de las ventas de un producto dado.

Los autores proponen para este proceso, el uso de redes bayesianas. Pero, señalan dos casos posibles que se pueden dar, de acuerdo a la característica del problema a resolver:

- **Cuando hay grupos (clases) identificados**: Se puede usar los algoritmos de inducción (TDIDT) a fin de descubrir las reglas de comportamiento de cada atributo clase. Luego, se pueden aplicar redes bayesianas que permitirán descubrir qué atributo de los establecidos como antecedentes de las reglas, tiene mayor incidencia sobre el atributo establecido como consecuente.
- **En el caso en que no hayan grupos identificados:** Se puede aplicar algoritmos de agrupamiento (como SOM), que permita encontrar los grupos. Luego de identificar los grupos, se pueden usar las redes bayesianas para establecer la incidencia de cada atributo con respecto al atributo clase.

2.2.2. Ingeniería de explotación de información

Algunos estudios analizados sobre explotación de información explican la carencia de procesos o metodologías formales que den soporte a los proyectos de explotación de información y advierten sobre la necesidad de modificar la forma de trabajo artesanal a una forma más ingenieril al realizar este tipo de proyectos [19], debido a las características dinámicas y complejas de los mismos y al incremento de fuentes de información accesibles en la actualidad [5].

La Ingeniería de Explotación de Información se define como la "aplicación de un enfoque sistemático, disciplinado y cuantificable al desarrollo de proyectos de explotación de información" [20], es decir, se refiere a los procesos y metodologías que se usan para ordenar y gestionar la tarea de detección de patrones de conocimiento en grandes volúmenes de datos [5].

En el mismo sentido, en [21] se expone las ventajas de usar un proceso estándar de explotación de información y se propone crear un Modelo de Procesos para Proyectos de Explotación de Información orientado a Pymes. Este modelo se basa en CRISP-DM por ser considerada la más completa de todas las metodologías evaluadas y brinda solución a varias de sus limitaciones y puntos débiles.

En [15], se analizan varios enfoques de explotación de información y se considera a CRISP-DM como una metodología. Se explica que una metodología es la instancia de un proceso que además de definir las tareas, entradas y salidas, especifica cómo hacer las tareas. La diferencia que existe entre un modelo de proceso y una

metodología es que el *modelo de proceso* establece qué hacer y la *metodología* especifica cómo hacerlo. Es muy importante el uso de una metodología en el desarrollo de proyectos de explotación de información, debido a que ayuda a realizar el trabajo de una manera sistemática y no trivial; proporciona una guía para planificar y ejecutar un proyecto de explotación de información, ya que permite ordenar el proceso de descubrimiento de patrones en repositorios de una organización.

En [5] se considera que una metodología es un proceso en el que se ha identificado "con que técnica se desarrolla cada tarea de cada fase del proceso", es decir, supone que una metodología es igual a un proceso más las técnicas usadas.

2.2.3. Roles en un proyecto de explotación de información

En [21] se especifican los *roles* para los distintos actores en un proyecto de explotación de información, estos son:

1)- **Líder de proyecto**: Es aquél que tiene la responsabilidad de planear, coordinar, ejecutar e implementar el proyecto.

2)- **Cliente del proyecto de explotación de información**: Es el experto en el dominio que requiere el proyecto y que utilizará los resultados pero que no posee generalmente los conocimientos requeridos para participar en la ejecución de las fases más técnicas del proyecto como la preparación de los datos o el modelado.

3)- **Analista de explotación de información**: Es quien posee una gran comprensión, desde la perspectiva de negocio, de lo que el cliente desea lograr y asiste en la traducción de estos objetivos en requerimientos técnicos a ser utilizados para la construcción de modelos.

4)- **Ingeniero en explotación de información:** Es aquel que desarrolla, interpreta y evalúa los modelos de explotación de información en base a los objetivos de negocio y criterios de éxito, realiza las tareas en constante consulta con el cliente y el analista para ser asistido en el logro del fin de negocio.

5)- **Analista IT:** Su responsabilidad es proveer el acceso al hardware, software y datos necesarios para completar el proyecto exitosamente.

Dependiendo de la envergadura del proyecto estos roles pueden ser asumidos por varios individuos o un mismo individuo asumir varios de ellos.

2.2.4. Explotación de información, minería de datos y ciencia de datos

En [15] se refiere a la explotación de información y la minería de datos como el proceso de extraer conocimiento útil de los datos disponibles en distintas fuentes de información, con el objetivo de obtener conocimiento de los mismos.

En un estudio realizado en [5] se explica que el término explotación de información se usa como referencia genérica a distintos tipos de minería de datos, como la de texto, de imágenes, en la web, entre otras. Además, se hace énfasis en que los resultados obtenidos de la explotación de información deben ser comprensibles, validables y que provean valor al proceso de toma de decisiones.

Por otra parte, en [22] se expone que la minería de datos se corresponde con una de las fases dentro de un proceso más amplio del "Descubrimiento de Conocimiento en Bases de Datos" (KDD) o del proceso de explotación de información. Es decir, no alcanza con obtener patrones a partir de la ejecución de algoritmos de minería de datos, sino que se deben seguir los pasos interactivos e iterativos del KDD, esenciales para garantizar la extracción de conocimiento útil a partir de los datos.

En el mismo sentido, en [16] se considera que los términos minería de datos y explotación de información no se deben utilizar para referirse al mismo cuerpo de conocimientos, porque la minería de datos se relaciona con la tecnología (algoritmos) necesaria para transformar los datos en conocimiento, por lo que está más vinculada a las tareas propias de la programación, en cambio la explotación de información se refiere a los procesos y metodologías necesarias para lograr el objetivo, es decir, esta mas relacionada con los procesos de la Ingeniería de Software.

En el contexto del descubrimiento de conocimiento los términos usados para hacer referencia a la minería de datos han cambiado en los últimos años. Actualmente, "Ciencia de Datos" es una expresión muy usada en el campo de la explotación de datos y se describe, en un estudio realizado en [23], como un término derivado de Minería de Datos o de la Ingeniería de Explotación de Información (*Information Mining*).

Muchos autores se refieren a la ciencia de datos como una herramienta fundamental para la explotación de datos y la generación de conocimiento; uno de sus objetivos es buscar modelos que describan patrones y comportamientos en un conjunto de datos, para la toma de decisiones o para hacer predicciones [24]; usa principios,

procesos y técnicas, para tratar de entender fenómenos mediante el análisis de datos [25].

La diferencia entre minería de datos y ciencia de datos es que la primera se orienta a objetivos y se concentra en el proceso, en cambio, la ciencia de datos se orienta a los datos, es exploratoria e incluye otros aspectos relacionados con la cantidad de datos (big data) y la visualización de los mismos [26].

2.3. Principales enfoques que se utilizan en explotación de información

KDD: El KDD (Knowledge Discovery in Databases o Descubrimiento de Conocimiento en Bases de Datos) se define como el "proceso no trivial de identificar patrones novedosos, válidos, útiles y entendibles en los datos" [14]. Es decir, los patrones o modelos obtenidos a partir de los datos deberían aportar nuevo conocimiento sobre el dominio que se está trabajando, ser válidos para nuevos datos y finalmente, ser útiles y comprensibles para el usuario final, de modo que puedan utilizarse en la toma de decisiones [5]. KDD se considera como una primera aproximación a un *modelo de procesos*; se constituyó en el año 1996 como el primer modelo aceptado por la comunidad científica que describe el descubrimiento de conocimiento en bases de datos como un proceso que consta de distintas fases, entre ellas la preparación de los datos, la interpretación y la difusión de los resultados [15]. Este proceso se describe como *iterativo* porque puede ser necesario realizar varias iteraciones hasta extraer conocimiento útil de los datos, además, en la salida de alguna de las fases se puede retroceder a pasos anteriores. Por otra parte, se dice *interactivo* porque el usuario o el experto en el dominio del problema, colabora en la preparación de los datos y validación del conocimiento que se extrae de los mismos.

En este estudio, se resume el modelo de proceso KDD en las siguientes etapas:

1)- **Selección** de los datos sobre los que se trabajará.
2)- **Pre-procesamiento** de los datos, donde se realiza un tratamiento de los datos incorrectos y ausentes.
3)- **Transformación** de los datos y reducción de la dimensionalidad.
4)- **Minería de datos**, donde se obtienen los patrones de interés según la tarea de minería que llevemos a cabo (descriptiva o predictiva).
5)- **Interpretación y evaluación** del nuevo conocimiento en el dominio de aplicación.

Si bien el KDD se puede abreviar en las cinco etapas señaladas, el proceso en su versión completa está formado por nueve etapas. En [5] se realiza una adaptación de

los pasos KDD definidos en [14], donde se describen los pasos asociados al proceso y las salidas de los mismos. Se suele hacer referencia al proceso completo de descubrimiento de conocimiento como KDD o minería de datos, sin embargo, según el modelo KDD la minería de datos es una de las etapas en la que se extraen los patrones a partir del conjunto de datos.

Con el gran desarrollo que se ha producido en el área de explotación de información y minería de datos en las últimas décadas, a partir del año 2000 surgieron nuevos modelos que permiten realizar el proceso de explotación de información en forma sistemática, en varios trabajos realizados coinciden que entre los más probados y difundidos se encuentran CRISP-DM y SEMMA [15], [22], [21]:

SEMMA: SEMMA fue creada por el SAS Institute [27], organización relacionada con el desarrollo de software de inteligencia de negocios. Es un acrónimo que se usa para describir las fases que componen el proceso de explotación de información [22], [28], que son: Muestreo (Sample), Exploración (Explore), Modificación (Modify), Modelado (Model), Evaluación (Assess).

SEMMA provee una guía general del trabajo que debe realizarse en cada una de estas etapas y fue desarrollada para ser aplicada por quienes utilizan el software de minería de datos "SAS Enterprise Miner", por lo que su aplicación se considera limitada.

CRISP–DM: CRISP–DM (acrónimo en inglés de CRoss Industry Standard Process for Data Mining o Proceso Estándar para Minería de Datos Independiente de la Industria). Esta es una metodología de libre distribución creada por un grupo de empresas europeas [15]. CRISP-DM es un método que permite orientar trabajos de minería de datos.

Como modelo de proceso, ofrece un resumen de seis fases que puede aplicarse a un proyecto de explotación de información. Tiene como principal objetivo desarrollar proyectos a partir de un proceso estandarizado (independiente de la industria y la herramienta), minimizando los costos que implica un proyecto de este tipo en una organización.

El modelo se puede ver en la **Fig. 1:** Ciclo de vida del modelo CRISP-DM [29]. En la imagen, las flechas indican las dependencias más frecuentes e importantes entre las fases. La secuencia entre las distintas etapas no es estricta, es decir, se puede avanzar y retroceder entre fases si es necesario. CRISP-DM provee como herramienta de ayuda en el desarrollo de un proyecto de explotación de información, un *modelo de referencia* y una *guía del usuario*. El documento del *modelo de referencia* describe de forma general las fases, tareas generales y salidas de un proyecto de explotación de información; la *guía del usuario* aporta información sobre la aplicación práctica del *modelo de referencia* a proyectos específicos, como consejos y listas de comprobación respecto a las tareas de cada fase [4].

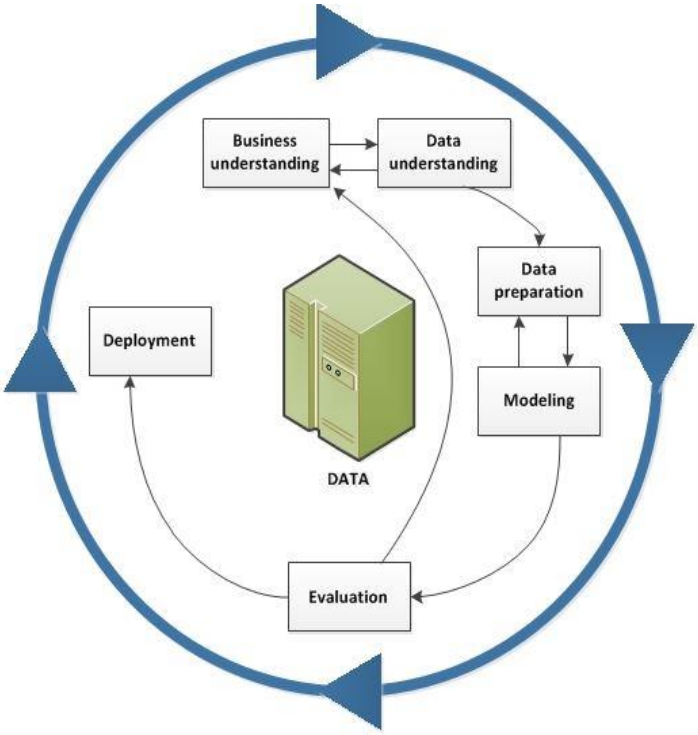


Fig. 1: Ciclo de vida del modelo CRISP-DM [29].

Como metodología, CRISP-DM incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. Algunas de las ventajas de esta metodología son que se la considera confiable y robusta, además de efectiva, ya que permite tener en cuenta las complejidades del proyecto a través de tareas fáciles de aplicar. Sin embargo, no establece las técnicas a implementar en cada actividad. La metodología CRISP-DM consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general al nivel más específico [19]. De acuerdo al problema que se quiera trabajar en una organización, puede suceder que en lugar de realizar el modelado, el trabajo se puede centrar en explorar y visualizar datos para descubrir patrones sospechosos en los datos. CRISP-DM permite crear un modelo de minería de datos que se adapte a las necesidades de una organización. En este

sentido, puede suceder que las fases de modelado, evaluación y despliegue sean menos importantes que las de preparación y comprensión de datos.

Otros enfoques: A partir del año 2000 surgieron otros enfoques con el avance de la disciplina [5], que se exponen en el Anexo D.

2.4. Metodología CRISP-DM

A partir del análisis de los diferentes estudios y de la opinión de distintos autores [16], [19], se considera que CRISP-DM es la más adecuada para utilizar en este TFM, motivo por el cual se describe esta metodología en detalle en esta sección.

En el nivel general el proceso se organiza en seis fases, que se estructuran en tareas generales de segundo nivel [19], [21], las que a su vez se dividen en tareas específicas donde se describen las acciones a desarrollar en determinadas situaciones. Las fases que establece la metodología CRISP-DM son las siguientes:

Fase 1: Comprensión del negocio

Esta fase requiere entender los objetivos del proyecto desde el punto de vista de la organización (perspectiva *no técnica*). Para lograr el éxito de un proyecto de explotación de información se necesita conocer los problemas del negocio (lo que el cliente desea lograr), luego convertir este conocimiento en una definición del problema de minería de datos y diseñar un plan para lograr dichos objetivos. Esta fase incluye las tareas que se detallan en la siguiente **Tabla 1:** Tareas de la fase 'Comprensión del negocio" de CRISP-DM [8].

Tabla 1: Tareas de la fase 'Comprensión del negocio" de CRISP-DM [8].

Fase	Tareas generales	Tareas específicas asociadas (salidas)
Comprensión del negocio	**1.Determinar los objetivos de negocio** (en términos organizacionales y desde una perspectiva no técnica): Entender y establecer desde el punto de vista del negocio los objetivos que el cliente desea lograr.	**1.1. Antecedentes**. Se puede incluir información sobre la situación actual de la organización, una descripción del problema y la solución para el mismo (si existe). **1.2. Objetivos de negocio**. Identificar los objetivos principales del cliente. **1.3. Criterios de éxito del negocio**. Describir los resultados esperados desde el punto de vista del negocio.
	2.Evaluar la situación: Evaluar la situación actual del negocio, analizar restricciones y factores que se deben tener en cuenta para el proyecto.	**2.1.** Inventario de recursos. **2.2.** Requisitos, supuestos y limitaciones. **2.3.** Riesgos y contingencias **2.4.** Terminología. **2.5.** Costos y beneficios.
	3.Determinar los objetivos de explotación de información. Los objetivos de minería de datos describen los objetivos del proyecto en "términos técnicos".	**3.1.**Objetivos de explotación de información **3.2.**Criterios de éxito de explotación de información
	4.Producir el plan del proyecto	**4.1.**Plan del proyecto. **4.2.**Evaluación inicial de herramientas y técnicas

Esta fase es muy importante porque de ella dependen muchas decisiones que se tomarán en las demás etapas, se deben tomar medidas adecuadas para reducir la probabilidad de fracasos en el desarrollo de proyectos de explotación de información [19]. A partir del análisis de los objetivos del proyecto y los repositorios de datos, es posible delimitar el alcance del proyecto en un conjunto de objetivos de minería de datos que se resolverán con la aplicación de procesos de explotación de información y mediante algoritmos de minería de datos correspondientes.

Los problemas de minería de datos representan los objetivos del proyecto en términos *técnicos* (objetivos de minería de datos) y la respuesta que se obtiene del problema de explotación de información, permite lograr los objetivos del negocio [5]. Además, comprender el dominio del proyecto incluye establecer un lenguaje común entre las personas involucradas [19]. Es importante manejar correctamente el vocabulario del negocio, usado en la investigación.

Fase2: Comprensión de los datos

Esta etapa comienza con la recolección inicial de datos e incluye acciones para familiarizarse con los mismos teniendo presente los objetivos del negocio (para cumplir con los objetivos definidos en la fase anterior). Esta fase incluye las tareas que se detallan en la siguiente **Tabla 2:** Tareas de la fase 'Comprensión de los datos' de CRISP-DM [8].

Tabla 2: Tareas de la fase 'Comprensión de los datos' de CRISP-DM [8].

Fase	Tareas generales	Tareas específicas asociadas (salidas)
Comprensión de los datos	**2.1. Recolectar los datos iniciales:** Recolectar los datos especificados en la lista de recursos del proyecto.	**2.1.1. Informe inicial de recopilación de los datos:** detalla la forma en que se obtuvieron los datos y los problemas que surgieron en el proceso
	2.2. Describir los datos: Describir en líneas generales los datos recolectados.	**2.2.1. Informe de descripción de los datos:** en este informe se describen los datos, el formato de los mismos y su tamaño (como cantidad de registros y variables). Es decir, se realiza el Diccionario de datos.
	2.3. Explorar los datos: Observar la distribución y el comportamiento de las variables mas relevantes.	**2.3.1. Informe de exploración de los datos:** en este informe se exponen los resultados del análisis (distribución de datos y comportamiento de las variables) . Es conveniente el uso de técnicas simples de análisis estadístico.
	2.4. Verificar la calidad de los datos: Examinar la calidad de los datos (completitud, errores y datos ausentes).	**2.4.1. Informe de calidad de los datos:** este informe se explica la calidad de los datos, como ser la completitud, errores en los datos y datos ausentes. Se documenta el análisis de calidad realizado y las posibles soluciones a los problemas encontrados.

En esta fase se identifican los problemas de calidad, ya que es común que en los repositorios de información existan errores de digitación, datos inconsistentes,

valores ausentes o duplicados. Estos problemas deterioran la calidad de los datos y por lo tanto, la calidad de las decisiones que se toman en relación a los mismos.

Fase3: Preparación de los datos

En esta fase el objetivo es obtener la vista minable o set de datos (dataset), es decir, se deben realizar las tareas de limpieza, formateo e integración de los datos recolectados para poder aplicar sobre esta vista minable, las técnicas de modelado.

En esta etapa, las tareas se pueden realizar muchas veces y sin un orden predeterminado, incluye la elección de registros y atributos, que se preparan para su posterior uso con las herramientas de modelado [21]. Las tareas a realizar en esta fase se detallan en la siguiente **Tabla 3**: Tareas de la fase 'Preparación de los datos' de CRISP-DM [8].

Tabla 3: Tareas de la fase 'Preparación de los datos' de CRISP-DM [8].

Fase	Tareas generales	Tareas específicas asociadas (salidas)
Preparación de los datos	**3.1.Especificar el conjunto de datos**: Descripción del Conjunto de datos	**3.1.1.Conjunto de datos** **3.1.2.Descripción del conjunto de datos**
	3.2. Selección de datos: Elegir los datos que se usarán en el análisis.Seleccionar los atributos (columnas) y las observaciones (filas o registros) con que se va a trabajar. Esta selección debe estar justificada..	**3.2.1. Justificación de la inclusión /exclusión de los datos**. Documento donde se justifiquen las causas por las cuales se incluyeron y excluyeron los datos.
	3.3. Limpieza de datos: tiene por objetivo mejorar la calidad de los datos. Se toman decisiones sobre los problemas de calidad, como datos ausentes o datos anómalos.	**3.3.1.Informe de limpieza de los datos**: en este reporte se incluyan las decisiones tomadas sobre los problemas de calidad de los datos.
	3.4. Construir los datos: En esta fase se construyen los nuevos datos, derivados de los disponibles, que son importantes para el análisis. Por ejemplo, atributos calculados o atributos transformados.	**3.4.1. Atributos derivados**. Estos atributos se calculan a partir de otros atributos del mismo registro. **3.4.2. Registros creados.** Estos registros se crean cuando son necesarios en la fase posterior de modelado
	3.5. Integrar los datos: Consiste en la integración de datos que provienen de diferentes tablas o registros.	**3.5.1.Datos combinados**: estos datos resultan de integrar la información de dos o más tablas que tienen diferente información de las mismas observaciones. En esta fase se incluye el cálculo de agregaciones, donde se calculan nuevos datos resumiendo información de diferentes tablas y registros.
	3.6.Formatear los datos: Se refiere al cambio que debe realizarse en el formato de los datos (pero no en su significado) por los requisitos de las técnicas de modelado elegidas.	**3.6.1.**Datos formateados. Conjunto de datos reformateados.

Fase4: Modelado

Esta etapa tiene como objetivo la aplicación de los algoritmos de explotación de información más adecuados sobre la *vista minable* o conjunto de datos trabajado en

la fase anterior. Se eligen y configuran las técnicas de modelado a utilizar (se determinan los parámetros óptimos). Pueden existir distintas técnicas para un mismo problema de minería de datos y a su vez, cada una de ellas tiene ciertos requisitos sobre los datos, por este motivo, suele ser necesario volver a la fase de preparación de los datos [21].

Esta fase incluye las tareas que se detallan en la siguiente **Tabla 4:** Tareas de la fase 'Modelado' de CRISP-DM [8].

Tabla 4: Tareas de la fase 'Modelado' de CRISP-DM [8].

Fase	Tareas generales	Tareas específicas asociadas (salidas)
Modelado	**4.1.Seleccionar la técnica de modelado**: Seleccionar qué técnica de minería de datos se usará.	**4.1.1.Técnica de modelado**: Documentar la técnica de modelado específica con la que se va a trabajar. **4.1.2.Supuestos del modelo**. Algunas técnicas asumen supuestos sobre el conjunto de datos, como por ejemplo distribución normal de una variable. Documentar todos los supuestos realizados.
	4.2. Generar el plan de prueba: Al construir los modelos, se necesita un mecanismo para determinar la calidad y validez de los mismos.	**4.2.1.Plan de pruebas:** Determinar y documentar de qué forma se entrenarán y evaluarán los modelos generados. Incluir las decisiones tomadas sobre los datos que se utilizarán para entrenamiento y prueba.
	4.3.Construir el modelo: se refiere a la aplicación de la técnica elegida sobre el conjunto de datos para generar uno o más modelos.	**4.3.1.Configuración de parámetros**: listar los parámetros proporcionados al modelo, justificando la elección de los mismos. **4.3.2.Modelo**: producido por las herramienta de minería. **4.3.3.Descripción del modelo**.
	4.4.Evaluar el modelo: se interpreta y evalúa el modelo en función del conocimiento del dominio, los criterios de éxito definidos para el proyecto y las pruebas diseñadas para el modelo. Los modelos pueden ser valorados y rankeados.	**4.4.1.Evaluación del modelo**. Generar un reporte de evaluación de los modelos obtenidos, describiendo sus características y un ranking para los mismos. **4.4.2.Revision de parámetros de configuración**. En función de la *evaluación* anterior, revisar los parámetros y ajustar los mismos para volver a la fase de construcción del modelo. Repetir las etapas 4.3 y 4.4 hasta asegurarse de que se han encontrado los "mejores" modelos.

Fase5: Evaluación

En la etapa anterior se pueden construir varios modelos y en esta fase se evalúan los modelos obtenidos, a fin de seleccionar los que contribuyan a una mayor calidad de análisis y que sean útiles a las necesidades del negocio [21]; es decir, se analiza si el modelo cumple con los criterios de éxito del proyecto identificados en la primera fase. Se requiere evaluar cada modelo, revisar los pasos ejecutados para la construcción del mismo, determinar la precisión del modelo e interpretarlo en el dominio del problema. Luego de este análisis, se determina si es necesario aplicar

nuevamente alguna de las fases anteriores por haber cometido algún error, o si se puede pasar a la siguiente fase de *implementación*.

Las tareas de esta fase son las que se especifican en la siguiente **Tabla 5:** Tareas de la fase 'Evaluación' de CRISP-DM [8].

Tabla 5: Tareas de la fase 'Evaluación' de CRISP-DM [8].

Fase	Tareas generales	Tareas específicas asociadas (salidas)
Evaluación	**5.1.Evaluar los resultados**: Evaluar el modelo en función de los objetivos del negocio.	**5.1.1.**Evaluar los resultados de la minería de datos con respecto a los criterios de éxito de negocio. **5.1.2.**Modelos evaluados y aprobados.
	5.2.Revisar el proceso: Realizar una revisión completa del proceso en búsqueda de posibles errores u omisiones.	**5.2.1.Revisión del proceso:** Revisar el proceso y documentar un resumen del mismo. Incluir las actividades omitidas o las que deberían ser repetidas.
	5.3.Determinar los próximos pasos: En relación a la evaluación de resultados y la revisión del proceso, decidir cómo continua el proyecto, si se pasa a la próxima fase (implementación) o si se retorna a una fase anterior.	**5.3.1.**Lista de posibles acciones **5.3.2.**Decisiones: Descripción de la decisión tomada.

Fase6: Implementación

Esta fase se suele llamar *despliegue* y tiene como objetivo la realización del plan de desarrollo, incluyendo la documentación y presentación de los resultados al cliente o partes interesadas. Es decir, se refiere a explotar los beneficios de los modelos, integrando el nuevo conocimiento en las tareas de toma de decisiones de la organización. Esta etapa puede ser tan simple como generar un reporte o tan compleja como implementar un proceso de explotación de información repetible en toda la empresa [22].

Las tareas de esta fase se detallan en la **Tabla 6:** Tareas de la fase 'Implementación' de CRISP-DM [8].

Tabla 6: Tareas de la fase 'Implementación' de CRISP-DM [8].

Fase	Tareas generales	Tareas específicas asociadas (salidas)
Implementación	**6.1.** Crear un plan de implementación para los resultados obtenidos con la minería de datos	**6.1.2. Ejecución del plan de implementación**: Incluir etapas y cómo llevarlas a cabo.
	6.2. Realizar el plan de monitoreo y mantenimiento: El monitoreo y mantenimiento es importante si los resultados de la minería formarán parte del trabajo diario del negocio y su entorno.	**6.2.1.** Ejecución del plan de monitoreo y mantenimiento.
	6.3. Realizar el informe final: Crear un informe final con el resumen del desarrollo del proyecto o mostrar un análisis comprensivo de los resultados obtenidos con la minería de datos.	**6.3.1.** Informe final del proyecto. **6.3.2.** Presentación final al cliente, con resultados y conclusiones.
	6.4. Revisar el proyecto: Identificar y analizar los puntos que fueron bien realizados, los que fueron mal realizados, y los que podrían mejorarse.	**6.4.1.** Documentación de la experiencia adquirida durante el desarrollo del proyecto.

2.5. Técnicas de minería de datos

En la minería de datos se pueden aplicar distintas técnicas provenientes del análisis estadístico y de sistemas inteligentes. Los métodos de análisis de datos considerados tradicionales trabajan con variables estadísticas, están orientados numéricamente y son esencialmente cuantitativos [4].

Luego de realizar el análisis de los datos se extrae conocimiento útil de los mismos; también se llaman modelos que consisten en relaciones, reglas, patrones y resúmenes obtenidos con el análisis de los datos [24]. En [22] se explica que los modelos de minería de datos pueden ser:

1)- **Predictivos**: Son los que permiten responder preguntas sobre datos futuros, a partir de los datos disponibles. Permiten estimar el valor de variables de interés (variables dependientes u objetivo), a partir de otras llamadas variables independientes. Además de los datos, se conoce la salida (clase, categoría o valor numérico) y a partir de los datos, se puede construir un modelo predictivo a través de distintas técnicas. Ejemplo: las técnicas de clasificación y la regresión son comúnmente usadas en este tipo de modelos.

2)- **Descriptivos**: Son los que permiten explorar las propiedades de los datos y aportan información sobre las relaciones existentes en los mismos [15], es decir,

consiste en hallar patrones o resumir datos, no pretenden predecir nuevos datos a partir de la información recabada. Los datos se presentan como un conjunto, sin estar ordenados ni etiquetados. Por ejemplo, el agrupamiento (*clustering*), las reglas de asociación.

En paralelo a los tipos de modelos, las tareas que se pueden abordar en un problema de minería de datos son [24]:

2.5.1. Técnicas predictivas

Clasificación: Es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo más cerca posible de otros y grupos diferentes estén lo más lejos posible de otros, donde la distancia se mide con respecto a las variables especificadas que se quieren predecir**.**

En un problema de clasificación se busca predecir los resultados en una salida discreta. Es decir, mapear variables de entrada en categorías discretas (etiquetas de clase) o encontrar un modelo que describa la distribución de las clases, que es una opción de una lista de posibilidades.

Ejemplo: árboles de decisión, aprendizaje bayesiano, regresión logística, métodos de vecinos más próximos [15].

Árbol de decisión: El aprendizaje de árboles de decisión se engloba dentro del aprendizaje supervisado. Los árboles de decisión son modelos predictivos formados por reglas binarias del tipo si/no, con las que se consigue repartir las observaciones en funcion de sus atributos y predecir asi el valor de la variable respuesta. Un árbol de decisión se puede interpretar como un conjunto de reglas representadas en forma de árbol, compuesto por ramas y nodos [24]. Las ramas, indican los posibles caminos generados automaticamente de acuerdo a la decisión tomada. El nodo interno contiene una evaluación en referencia a algún valor de una de las propiedades. El nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema. El nodo hoja representa el valor que devolverá el árbol de decisión.

Redes neuronales BP: Son redes formadas por múltiples capas que permiten resolver problemas que no son linealmente separables [4]. Utilizan un algoritmo de aprendizaje llamado regla delta generalizada, que consiste en minimizar el error, por medio del método del gradiente descendente en los parámetros de entrenamiento de la red neuronal. Estas redes se conocen como redes de retropropagación (Redes BP).

Redes bayesianas: Las redes bayesianas o probabilísticas se fundamentan en la teoría de la probabilidad. Combinan la potencia del teorema de Bayes con grafos dirigidos. Permiten modelar de forma cualitativa el conocimiento y expresar en forma numérica la intensidad de la relación entre las variables. Una red bayesiana se puede representar de dos formas, como una base de reglas o como una distribución de la probabilidad conjunta de las variables representadas en la red bayesiana [22]. Se definen como un grafo acíclico dirigido, en el que cada nodo representa a las variables aleatorias y las flechas muestran las influencias causales entre las variables. Las variables usadas pueden ser continuas o discretas. Si un nodo es padre de otro significa que es causa directa del mismo.

Las redes de creencias o bayesianas, son usadas para calcular la ponderación de interdependencia entre atributos. Obtener una red bayesiana a partir de datos es un proceso de aprendizaje que se divide en dos etapas: el aprendizaje estructural y el aprendizaje paramétrico. Una red bayesiana representa relaciones causales en el dominio del conocimiento a través de una estructura gráfica y las tablas de probabilidad condicional entre los nodos.

A pesar de su nombre, las redes bayesianas o de creencias no implican necesariamente el uso de estadística bayesiana; sin embargo, se denominan así por la utilización de la regla de Bayes para el cálculo de inferencia probabilística. Llamarlas gráficos acíclicos dirigidos podría ser más apropiado dado que un clasificador de bayes ingenuo (Naive Bayes) resulta útil para la representación de modelos jerárquicos condicionados a algún atributo.

2.5.2. Técnicas descriptivas

Algunas de las técnicas descriptivas son el agrupamiento y las reglas de asociación [24]:

Clustering: También llamadas técnicas de agrupamiento o segmentación, permite analizar datos para generar etiquetas y utiliza el aprendizaje no supervisado. Algunas técnicas de agrupación son: *Clustering* jerárquico, métodos basados en particiones (*K-Means*) y métodos basados en la densidad (*DBSCAN*).

Análisis de correspondencia (A.C.): El análisis de correspondencia (*Correspondence Analysis*) calcula la transformación lineal de los datos de entrada. Si bien es similar al análisis de componentes principales (*Principal Component Analysis* o PCA), A.C. calcula la transformación lineal en datos discretos en lugar de continuos. El análisis de correspondencia puede representar varias variables en un

gráfico bidimensional, lo que facilita ver las relaciones entre los valores de las variables [30].

El A.C. es una técnica de reducción de dimensión y visualización de la información que permite reducir el número de variables y visualizar la nube de puntos N-Dimensionales para poder detectar estructuras o características de forma visual [31]. Es una técnica descriptiva o exploratoria que permite resumir una gran cantidad de datos en pocas dimensiones, con la menor pérdida de información posible.

Una característica que diferencia al A.C. de otros métodos es el uso de variables cualitativas; es decir, se usan como datos las frecuencias de una tabla. También se puede definir al A.C. como una técnica estadística que permite analizar gráficamente las relaciones entre variables categóricas a partir de los datos de una tabla de contingencia. Una tabla de contingencia se puede visualizar mediante un gráfico de mosaico en el software Orange [32].

Reglas de asociación: Son similares a la correlación y tienen como objetivo encontrar relaciones no explícitas entre atributos. Tienen aplicación práctica en muchos campos; por ejemplo, se utilizan para comprender los hábitos de compra [24] o para analizar el contenido de un carrito de compra. Uno de los algoritmos más utilizados es el algoritmo Apriori.

Algoritmo Apriori: La generación de reglas de asociación se logra basándose en un procedimiento de *covering*. Las reglas de asociación son parecidas, en su forma, a las reglas de clasificación; si bien en su lado derecho puede aparecer cualquier par o pares atributo-valor. De manera que para encontrar ese tipo de reglas, es preciso considerar cada posible combinación de pares atributo-valor del lado derecho.

Redes autoorganizadas de Kohonen o redes SOM (Self-Organizing Map)

Las redes neuronales SOM, también conocidas como mapas auto-organizados de Kohonen, son un tipo de red no supervisada y competitiva que crea un mapa topológico con regiones que tienen datos similares. En el aprendizaje competitivo, las neuronas compiten entre sí con el objetivo de activarse, los objetos similares del conjunto de datos se clasifican dentro de una misma categoría, activando la misma neurona de salida.

Estas redes aplican vectores con valores de los datos que modifican los pesos sinápticos de diversas regiones de la red, de manera que se van identificando y diferenciando los grupos a los que pertenecen las instancias clasificadas. Esto permite descubrir la estructura subyacente de los datos ingresados, a partir de

establecer características comunes entre los vectores de información de entrada a la red [4]. A lo largo del entrenamiento de la red, los vectores de datos son introducidos en cada neurona y se comparan con el vector de peso característico de la misma. La neurona que presenta menor diferencia entre su vector de peso y el vector de datos es la neurona ganadora y ella y sus vecinas verán modificados sus vectores de pesos.

Estos mapas SOM auto-organizados son muy útiles cuando se trabaja con grandes cantidades de datos, son tolerantes al ruido y tienen la capacidad de extender la generalización al momento de necesitar manipular datos nuevos.

2.6. Software de minería de datos

Muchas de las herramientas utilizadas en minería de datos son software de código abierto (*open source*), lo que facilita la aplicación de minería de datos y disminuye los costos de implementación de proyectos de explotación de información en las organizaciones.

En [22] se describen distintos software que permiten realizar el proceso de minería de datos, entre ellos:

- Clementine de la empresa SPSS/Integral Solutions Limited (ISL).
- Enterprise Miner de la empresa SAS.
- RapidMiner, gratuito y basado en la filosofía *open source*.
- WEKA, gratuito y basado en la filosofía *open source*.
- Orange, gratuito y basado en la filosofía *open source*.

Se describen brevemente los software Weka y Orange:

Weka (Waikato Environment for Knowledge Analysis o Entorno para Análisis del Conocimiento de la Universidad de Waikato): Es una plataforma de software para el aprendizaje automático y la minería de datos, escrito en Java y desarrollada en la Universidad de Waikato [7]. Es de libre distribución (licencia GPL) y se destaca por la cantidad de algoritmos que presenta, como así también por la eficiencia de los mismos.

El software provee una gran cantidad de herramientas para la realización de tareas de minería de datos, visualización y además permite programar en Java algoritmos más sofisticados para el análisis de datos y el modelado predictivo. Cuenta con una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades, en la que

se pueden implementar técnicas de clasificación, asociación, agrupamiento y predicción.

Weka soporta varias tareas estándar de minería de datos, especialmente preprocesamiento de datos, clustering, clasificación, regresión, visualización y selección. Todas las técnicas de Weka trabajan con datos disponibles en un fichero plano, en el que cada registro de datos está descrito por un número fijo de atributos (normalmente numéricos o nominales, aunque también se soportan otros tipos).

Orange: Es un paquete de software para aprendizaje automático y minería de datos desarrollado en el Laboratorio de Bioinformática de Ia Facultad de Ciencias de la Computación e Informática de Ia Universidad de Liubliana, Eslovenia, junto con la comunidad de código abierto. Es un software libre que incluye un amplio rango de técnicas de preproceso, modelado y exploración de datos [24].

Una de sus principales ventajas es que permite realizar el análisis de datos y obtener visualizaciones mediante el modelo de arrastrar y soltar (*drag and drop*), a través de objetos GUI llamados Orange Widgets [33], que representan diferentes tareas. Además, puede ser instalado como una librería de Python.

Orange permite realizar visualizaciones interactivas de los datos, que ayudan a descubrir patrones de datos ocultos, brindan intuición detrás de los procedimientos de análisis de datos o respaldan la comunicación entre científicos de datos y expertos en el dominio. Los widgets de visualización [34], incluyen distintos tipos de diagramas, como ser, de dispersión, diagramas de caja e histogramas, y visualizaciones específicas del modelo como dendrogramas, diagramas de siluetas y visualizaciones de árboles, entre otros.

2.7. Estándares Internacionales

La ISO (International Organization for Standardization - Organización Internacional de Normalización), es una organización internacional no gubernamental, que tiene como miembros distintos organismos nacionales de normalización. Esta organización ha publicado miles de Estándares internacionales que abarcan casi todas las industrias, desde tecnología, seguridad alimentaria, agricultura y atención médica [35]. Estos estándares pueden ser empleados en organizaciones como guía en la adopción de mejores prácticas, para garantizar que sus productos y/o servicios sean seguros, confiables, de alta calidad.

Los estándares ISO son acordados internacionalmente por expertos y describen o especifican la mejor manera de llevar a cabo las tareas en las organizaciones.

En [36] se define como activo a "cualquier cosa que tenga valor para la organización". Los activos de procesos abarcan planes, políticas, procedimientos y lineamientos, formales o informales; también abarcan las bases de conocimiento de la organización, como las lecciones aprendidas e información histórica.

Dentro de los estándares que proporcionan modelos para la evaluación de la calidad del software, se encuentran las normas ISO 9001, esta norma pertenece a la familia ISO 9000 [1] de Gestión de la Calidad, que ayuda a las organizaciones a mejorar la calidad de sus productos y servicios.

En la Norma ISO 9001 se establecen los criterios para un sistema de gestión de la calidad y es el único estándar de la familia que puede certificarse (aunque esto no es un requisito). Puede ser utilizado por cualquier organización, grande o pequeña, independientemente de su campo de actividad.

Al realizar una comparación de la Norma ISO 9001:2008 y la Norma ISO 9001:2015 [37], se observa que existe una mayor flexibilidad en esta última en relación a los requisitos para los procesos, la información documentada y las responsabilidades de la organización.

Donde la Norma ISO 9001:2008 se refería a "documento" o "procedimientos documentados", "manual de la calidad", en la presente edición de esta Norma define requisitos para "mantener la información documentada". La Información documentada se refiere a la información generada para que la organización opere (documentación). La organización es responsable de determinar qué información documentada se necesita conservar, el periodo de tiempo por el que se va a conservar y qué medios se van a utilizar para su conservación.

En la norma ISO 9000:2015 se define un proceso como un conjunto de actividades que interactúan para transformar elementos de entrada en elementos de salida. Estos procesos pueden "definirse, medirse y mejorarse". Por otra parte, en la norma se especifica que un procedimiento se refiere a "la forma específica de llevar a cabo una actividad o un proceso", es decir, cuando un proceso tiene pasos establecidos y ordenados para obtener un resultado.

En otras palabras, un proceso es *lo que hacemos* (proceso de explotación de la información) y un procedimiento es *cómo lo hacemos*.

Los procedimientos pueden estar documentados o no. Muchas organizaciones no establecen en sus políticas la documentación de procedimientos, sin embargo, contar con los mismos favorece la realización de actividades futuras que impliquen tareas

similares. Un procedimiento permite documentar la forma en que se deben realizar las actividades rutinarias o repetitivas de una organización; su principal objetivo es estandarizar la realización de las mismas, a fin de minimizar errores, desvíos y variaciones.

La Norma ISO/TR 10013 "Directrices para la documentación del sistema de gestión de la calidad", proporciona pautas para desarrollar y mantener la documentación necesaria para el S.G.C. En esta norma se especifica que la organización debería definir la estructura y formato de los procedimientos documentados; estos pueden presentarse en formato texto, diagramas de flujo, tablas o una combinación de estos, de acuerdo a las necesidades de la organización. En esta norma, se aporta una jerarquización de la documentación en tres niveles [38], en los que especifica:

a)- En un primer nivel (Nivel A), se refiere al *Manual de gestión de la calidad*. El manual de calidad es el documento que describe las características básicas del sistema de gestión de calidad (SGC) de acuerdo a la normativa aplicada (por ejemplo, ISO 9001, o la que corresponda).

b)- En un segundo nivel (Nivel B), hace referencia a los procedimientos del Sistema de Gestión de Calidad (SGC): cada organización debe establecer cuáles son los procedimientos que necesita documentar. Si esto no está especificado en las políticas de la organización, la Norma ISO 10013 no lo exige.

c)- En un tercer nivel (Nivel C), se refiere a instrucciones de trabajo, registros y otros documentos del SGC.

En la Norma ISO 10013 se especifica la estructura de los procedimientos documentados, que son documentos complementarios que responden a ¿qué?, ¿quién lo hace?, ¿cuándo?, ¿dónde?.

Según esta norma, los procedimientos de calidad deben incluir los siguientes elementos expuestos en la **Tabla 7:** Estructura de los procedimientos documentados según la Norma ISO 10013 [38].

Tabla 7: Estructura de los procedimientos documentados según la Norma ISO 10013 [38]

	Elementos	Descripción
1	**Título**	Titulo que identifica al procedimiento.
2	**Propósito**	Describe el propósito del procedimiento.
3	**Alcance**	Permite explicar el alcance del procedimiento, que áreas cubre y cuáles no.
4	**Responsabilidades y funciones**	Permite describir las responsabilidades y funciones de todas las personas/cargos incluidos en cualquier parte del procedimiento.
5	**Los registros**	Define y lista los registros que resultan de las actividades descritas en el procedimiento.
6	**Control de documentos**	Identifica los cambios, la fecha de revisión, la aprobación y versión del documento debería ser incluida en cada documento de acuerdo a lo establecido en el control de documentos.
7	**Descripción de actividades**	Esta es la parte principal del procedimiento. Permite describir los demás elementos del procedimiento, es decir, qué debería realizarse, por quién y cómo, cuándo y dónde. En algunos casos el "por qué" también debería definirse. Además, las entradas y salidas de las actividades deben ser explicadas, incluyendo los recursos que sean necesarios.
8	**Anexos**	Se pueden incluir anexos, en caso de ser necesario.

Una vez identificados y definidos los procesos se documentan los procedimientos. Un documento es la información (datos que poseen significado) y su medio de soporte (procedimiento, documento, dibujo, informe, etc.) según la norma ISO 9001, un documento es la forma en que se provee la información que se requiere en una organización para desempeñar las actividades.

Para la realización de diagramas que se incluyen en el procedimiento de explotación de información, se pueden usar los siguientes estándares:

- UML (*Unified Modeling Language* o Lenguaje de Modelado Unificado): Es un estándar ISO; se presenta en [39] como ISO/IEC 19501:2005, donde se describe al UML como "un lenguaje gráfico para visualizar, especificar, construir y documentar los artefactos de un sistema intensivo en software". UML permite modelar casi cualquier tipo de aplicación, plataformas de implementación y se puede usar con cualquier proceso de desarrollo [40].
- BPMN (*Business Process Model and Notation* o Modelo y Notación de procesos de Negocio): Es un estándar global para modelar procesos de negocios. BPMN 2.0 es parte de los estándares de mejora de procesos de OMG (*Object Management Group* o Grupo de Gestión de Objetos) [41], que difieren del lenguaje de modelado unificado (UML) utilizado en el diseño de software. La

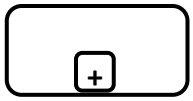
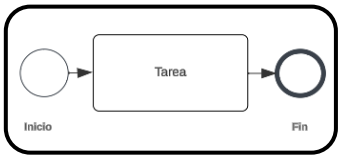
especificación BPMN 2.0.1 de OMG se ha publicado como norma internacional ISO/IEC 19510:2013 [42].

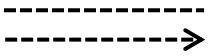
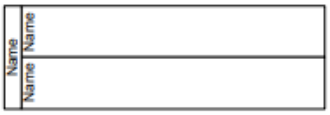
2.7.1. Notación grafica estandarizada para procesos de negocio BPMN

BPMN es una notación gráfica para el modelado de procesos que utiliza el formato de flujo de trabajo. El principal objetivo es proveer una notación estándar, fácilmente legible, que permita coordinar la secuencia de los procesos o actividades y los mensajes que fluyen entre los responsables de las diferentes actividades [43].

Para diagramar un procedimiento es esencial entenderlo, es decir, saber cuál es su objetivo y el camino o los pasos necesarios para lograrlo. Los elementos básicos de la notación BPMN se exponen en la siguiente **Tabla 8:** Elementos de notación BPMN [44].

Tabla 8: Elementos de notación BPMN [44].

1-Objetos de flujo: Actividades, eventos y puertas de enlace. Son los principales elementos gráficos que definen el comportamiento de los procesos.		
Actividades	Describen el trabajo desarrollado dentro de un proceso de negocio, pueden ser atómicas o compuestas. Se usan para modelar tareas y subprocesos. Pueden ser iterativas.	
Subproceso *colapsado*	Un subproceso es un conjunto de actividades incluidas dentro de un proceso. Se usa cuando los detalles del subproceso no pueden visualizarse. El signo (+) indica que la actividad es un *subproceso* y que tiene un nivel más bajo de detalle (tareas), esta asociado a un solo rol.	
Subproceso *expandido*	Los detalles del mismo pueden visualizarse, es decir, esta en el mismo nivel de detalle del proceso y tiene un evento de inicio y fin de proceso; puede estar asociado a uno o varios roles.	
Eventos	Representan algo que sucede durante el desarrollo de un proceso y que afecta su flujo. La frontera determina el tipo de evento. Según el momento en que ocurran, un evento puede ser *inical, intermedio o final*. Funcionan como un disparador para iniciar o completar un proceso.	
Compuertas	Son controles de secuencia de flujo en un proceso. Determinan bifurcaciones o combinaciones de rutas, en relación a las condiciones expresadas.	

2-Objetos de conexión: Los *objetos de flujo* están conectados entre sí mediante *objetos de conexión*. Estos son los diferentes tipos de objetos de conexión de árbol.		
Flujo de secuencia	Un *flujo de secuencia* se usa para enlazar dos elementos. Muestra la secuencia en que las actividades se realizan. Se representa como una línea recta con una flecha,	
Flujo de mensajes	Un *flujo de mensajes* indica el envío de un mensaje entre dos elementos ubicados en piscinas distintas. No puede ser usado para conectar actividades en una misma piscina. Se representa con una línea discontinua con un círculo al principio y una flecha al final.	
Asociación	Se utiliza para asociar *artefactos* o *elementos de datos* a un *objeto de flujo*.se representa por una línea punteada.	
3-Canales (*Swimlanes*): Representan los responsables de las actividades en un proceso. Por ejemplo: Organizaciones, roles, áreas funcionales o sistemas.		
Piscina	Identifica cada uno de los principales participantes en un proceso. Puede contener uno o más carriles. Puede ser abierta o cerrada (muestra o esconde los detalles internos.	
Carril	Muestra un rol o área funcional dentro de una piscina, se usa para organizar y categorizar las actividades de acuerdo a funciones o roles de las personas o áreas involucradas en un proceso.Se utilizan para organizar y categorizar *actividades*.	
4. Artefactos: Elementos de documentación para hacer más comprensibles los diagramas.		
Agrupación	Se usa para agrupar diferentes actividades pero no afecta al flujo dentro de un diagrama. Un *grupo* es una agrupación de elementos gráficos que están dentro de la misma *categoría*.	
Anotación de texto	Las *anotaciones de texto* permite que un modelador proporcione información de texto adicional para el lector de un diagrama BPMN.	
5.Datos: Representan archivos de datos, objetos de datos o documentos que son producidos y/o consultados por un proceso o actividad. Pueden ser: dato de entrada, dato de salida, dato de tipo de objeto, colección de objetos de datos, almacén y mensaje.		
Objeto de datos	Muestra qué datos se requieren para una actividad. Proporcionan información sobre lo que requieren las actividades para ser realizadas (datos de *entrada*) y/o lo que producen (datos de *salida*).	
Almacén de datos	Base de datos o conjunto de datos almacenados en algún medio de soporte.	

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Capítulo 3

Metodología

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

3. Metodología

En este capítulo del TFM se expone la metodología usada en el desarrollo de este trabajo. Se presenta una única sección (sección 3.1), en la que se expone el marco metodológico usado, que se estructura en tres etapas.

3.1. Marco metodológico

Este TFM inicia con el planteamiento del problema y objetivos que buscan solucionar el mismo, descriptos en el capítulo 1. La propuesta metodológica seguida en este trabajo se estructura en tres etapas:

Etapa 1: A fin de lograr los objetivos propuestos, se realizó inicialmente una Revisión Sistemática de la Literatura (RSL) relacionada con explotación de información y minería de datos, con la finalidad de delimitar el campo de estudio y generar información que sustente el trabajo realizado.

A partir de la revisión de la bibliografía existente sobre explotación de información y minería de datos se definió el objeto de estudio como el diseño de un procedimiento de explotación de información, que permita sistematizar las tareas a realizar en un proyecto de explotación de la información mediante una adaptación de CRISP-DM, que ayude a detectar o predecir las problemáticas laborales y sus factores de incidencia en un conjunto de datos estructurados de PEA.

En base al estudio de distintas investigaciones en este campo y en relación con la opinión de diversos autores (expuestas en el capítulo 2), que coinciden en que CRISP-DM es adecuada para aplicar a proyectos de explotación de información, se eligió esta metodología para el diseño de un procedimiento de explotación de la información con la finalidad de obtener resultados de calidad en este tipo de proyectos. El análisis de la literatura posibilitó además, la elaboración del estado de la cuestión de la temática elegida en este TFM, que se presentó en el capítulo 2. La revisión de la bibliografía se realizó en forma continua durante el desarrollo de este trabajo a fin de actualizar el mismo con nuevos hallazgos en el campo y con la finalidad de mejorar el procedimiento definido inicialmente en este TFM.

Etapa 2: En esta etapa, se diseñó el procedimiento de explotación de la información que es una adaptación de la metodología CRISP-DM para proyectos de explotación de información.

A partir del análisis de estándares para el diseño de procedimientos, expuesto en el capítulo anterior (estado de la cuestión), se diseñó un *Modelo general de procedimiento de explotación de información*, que se usó en los procedimientos

creados. Además, se realizó una propuesta de aplicación del procedimiento de explotación de la información (que consta de un procedimiento principal y procedimientos específicos), que es una representación gráfica que muestra como se aplican los procedimientos creados en un proyecto de explotación de información.

Se analizaron distintos estándares para el diseño del procedimiento de explotación de la información y se consideraron varias definiciones expuestas en la Norma ISO [1] para la elaboración del mismo. En dicha norma se expone que una vez identificados y definidos los *procesos* se puede documentar o escribir lo que se va a realizar, por ejemplo, mediante *procedimientos*. Es decir, un documento es la información y su medio de soporte puede ser un procedimiento, un dibujo, un informe, etc.; consiste en la forma en cómo se provee la información que en la organización requieren para desempeñar las actividades. Para diseñar un procedimiento es fundamental *entender* el mismo, es decir, saber qué objetivo se desea lograr y los pasos necesarios para realizarlo. En este sentido, para el diseño del procedimiento de explotación de la información se realizó una adaptación del modelo de proceso CRISP-DM, que presenta el ciclo de vida de minería de datos en varias fases (Fig. 1), a fin de ordenar el trabajo realizado en un proyecto de explotación de información.

En cuanto al diseño de los procedimientos específicos (que se ejecutan dentro del procedimiento principal), se tienen en cuenta los problemas de minería de datos que se quieren resolver en el proyecto; se definen los objetivos de minería de datos para resolver esos problemas y se determina el proceso de explotación de información a aplicar según las definiciones dadas en [4], para el logro de los objetivos de minería de datos (que dan solución al objetivo general de este TFM). Entonces, para cada proceso de explotación de información identificado, se generó un procedimiento específico en el que se define el objetivo del procedimiento, su alcance, la descripción del mismo y se realizó un diagrama del proceso de explotación de información, con la notación BPMN, en el que se muestra la entrada, el proceso y la salida del mismo. Al identificar el proceso de explotación de información a aplicar, se puede determinar las técnicas o algoritmos a usar; esta elección va a estar dada por el tipo de problemas y el tipo de datos que se este trabajando.

En la elaboración de procedimientos de explotación de la información se consideran los siguientes estándares:

1)- ISO 9000:2015 [1] y 9001:2015 [37]: Estas normas se utilizan en la definición de conceptos fundamentales a tener en cuenta al diseñar el procedimiento.

2)- ISO 10013: Esta norma proporciona *Directrices para la documentación de procedimientos*; incluye la descripción de los elementos que debe contener un procedimiento y aporta información sobre la estructura del mismo. La misma está explicada en [38].

3)- BPMN (*Business Process Model and Notation o Modelo y Notación de procesos de Negocio*) [42]: Es un estándar para el modelado de procesos de negocio y servicios web. Es una notación gráfica que permite diseñar las actividades de un procedimiento, que se explica en [43] y [44].

El diseño del procedimiento se realizó mediante tablas y diagramas. Para los diagramas se usó la notación BPMN, que permite representar las tareas asociadas a cada procedimiento. La elección de esta notación se debe a que luego del estudio realizado, se considera a BPMN como la más adecuada para este tipo de proyectos.

Etapa 3: En esta etapa se aplicó el procedimiento de explotación de la información (adaptación de CRISP-DM), a un caso de estudio. Es decir, se realizó la validación del procedimiento de explotación de la información diseñado, con datos reales, obtenidos en el relevamiento sociodemográfico realizado en el Barrio Industrial de la ciudad de Corrientes. Esta validación se efectuó en base a la propuesta de aplicación de procedimientos, presentada en el capítulo 4 (Fig. 2).

En la validación del procedimiento de explotación de la información se describen, las tareas a realizar en un proyecto de explotación de la información mediante la aplicación del procedimiento principal (adaptación de CRISP-DM) y los procedimientos específicos que se ejecutan dentro del mismo, en relación a los objetivos de minería de datos definidos y procesos de explotación de información determinados. Se pretende que el procedimiento diseñado facilite la tarea de encontrar patrones de comportamiento o conocimiento en una base de datos y ayuden a detectar las problemáticas laborales y sus factores de incidencia, en PEA pertenecientes a zonas urbanas. Asimismo, constituye una guía en la ejecución de las tareas y actividades que se desarrollan en el proyecto de explotación de información, requeridas para el logro de los objetivos establecidos.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Capítulo 4

Solución Propuesta

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

4. Solución Propuesta

En este capítulo del TFM se explican los aspectos generales de la solución propuesta (sección 4.1); en la siguiente sección se presenta un modelo general a usar en diseño del procedimiento de explotación de la información (sección 4.2); luego, se presenta una propuesta de aplicación de procedimientos a un proyecto de explotación de información, que resume la solución a la problemática planteada (sección 4.3), esta propuesta consta de un procedimiento principal a ser aplicado al proyecto de explotación de la información que es una adaptación de CRISP-DM (sección 4.4) y procedimientos específicos diseñados en relación a cada objetivo de minería de datos definido y proceso de explotación de información identificado (sección 4.5.); finalmente, se valida el procedimiento diseñado, en un caso de estudio (sección 4.6.), se presenta el caso de estudio sobre el que se realiza la validación (sección 4.6.1) y seguidamente, se realiza la validación del procedimiento de explotación de la información (sección 4.6.2.), mediante la validación del procedimiento principal propuesto en este TFM y los procedimientos específicos asociado a cada objetivo de minería de datos definido.

4.1. Aspectos generales de la solución propuesta

La solución propuesta en este TFM consiste en el diseño de un procedimiento de explotación de información, que permita sistematizar las actividades realizadas en un proyecto de explotación de información, servirá como guía en la realización de este tipo de proyectos y permitirá documentar el trabajo realizado.

El procedimiento de explotación de la información planteado consta de:

- Un procedimiento principal, que es una adaptación de la metodología CRISP-DM, que permite guiar la ejecución del proyecto de explotación de información. Esta metodología describe de forma detallada y precisa las actividades a realizar en un proyecto de explotación de información, por lo que el uso de CRISP-DM ayuda a encontrar patrones válidos y potencialmente útiles a partir de un trabajo ordenado.
- Procedimientos específicos, diseñados para resolver los problemas de minería de datos. Dado que un proyecto de explotación de la información puede involucrar varios problemas de minería de datos, se definen los objetivos de minería de datos y los procesos de explotación de información que permitan dar solución a los problemas detectados. Se diseñó un procedimiento específico en relación con

cada proceso de explotación de información identificado, con el fin de resolver los problemas de minería de datos y cumplir con el objetivo general de este TFM.

Se pretende que el procedimiento de explotación de la información diseñado facilite la tarea de encontrar patrones de comportamiento o conocimiento en una base de datos y ayude a detectar las problemáticas laborales y sus factores de incidencia, en una PEA perteneciente a zonas urbanas. Asimismo, constituye una guía en la ejecución de las tareas y actividades que se desarrollan en el proyecto de explotación de la información, requeridas para el logro de los objetivos establecidos.

Al identificar el proceso de explotación de información a aplicar, se pueden determinar las técnicas o algoritmos a usar; esta elección va a estar dada por el tipo de problema y el tipo de datos que se esté trabajando.

4.2. Modelo de procedimiento de explotación de información

En esta sección se propone un modelo general que se aplica al diseño del procedimiento de explotación de la información, que permite sistematizar las actividades a realizar en un proyecto de este tipo, con la metodología CRISP-DM. Este modelo se organiza según las especificaciones expuestas en la Norma ISO 10013, explicada en [38], la que establece la estructura para los procedimientos.

Las partes que presenta el procedimiento de explotación de la información propuesto son:

Encabezado: Si bien la Norma ISO 9001 no exige una codificación para los procedimientos, es recomendable hacerlo por motivos de organización de documentos ya que facilita su búsqueda y actualización. El encabezado debe aparecer en cada página del procedimiento y en general consiste en un recuadro de identificación con los siguientes elementos:

a)- **Logo de la organización** (en el lado izquierdo del recuadro). Incluir una imagen representativa y aprobada.

b)- **Nombre del procedimiento (**en la parte central del recuadro).

c)- A la derecha del recuadro se puede incluir el código del procedimiento, el número de revisión, la fecha en que entrará en vigencia y el número de páginas. También se puede optar por adicionar el proceso al que pertenece.

c.1)- Código: PR-XXX-YY

Los procedimientos diseñados en este TFM se codificarán con el código: PR-XXX-YY. El código de procedimiento a usar en el encabezado del procedimiento se establece en la **Tabla 9:** Código de procedimiento a usar en el encabezado.

Tabla 9: Código de procedimiento a usar en el encabezado.
Fuente: Elaboración propia

Notación		Descripción
PR		Procedimiento
XXX:		Identifica al tipo de procedimiento, principal o especifico.
	GEN	Procedimiento principal de explotación de la información.
	EI1	Procedimiento específico de explotación de información, asociado al objetivo de minería de datos N° 1
	EI2	Procedimiento específico de explotación de información, asociado al objetivo de minería de datos N° 2
	EI3	Procedimiento específico de explotación de información, asociado al objetivo de minería de datos N° 3
	EI(..)	Procedimiento específico de explotación de información, asociado al objetivo de minería de datos N°(..)
YY:		Número de orden del documento (procedimiento)
	00	Corresponde al procedimiento principal
	01	Procedimiento1, asociado al objetivo de minería de datos N° 1
	02	Procedimiento2, asociado al objetivo de minería de datos N° 2
	03	Procedimiento3, asociado al objetivo de minería de datos N° 3
	(..)	Procedimiento(..), asociado al objetivo de minería de datos N° (..)

c.2)- Revisión: El número de revisión del procedimiento se expone en la **Tabla 10:** Número de revisión del procedimiento a usar en el encabezado.

Tabla 10: Número de revisión del procedimiento a usar en el encabezado.
Fuente: Elaboración propia

Notación		Descripción
Revisión:		La letra "A" corresponde a un borrador. A partir de la primera revisión se escribe 1,2, y así sucesivamente.
X	**A**	Versión inicial (borrador) del procedimiento.
	1, 2,(..)	Luego de las distintas revisiones, las mismas se enumerarán en forma correlativa (1,2,3…).

c.3)- Fecha: dd/mm/aaaa. Ingresar día, mes y año en que se establece el procedimiento.

c.4)- Página: 1 de xx. Ingresar el número consecutivo de las hojas utilizadas en el procedimiento y número total de páginas (xx).

Recuadro de control de emisión: Se situará únicamente al pie de la última hoja del procedimiento. En el mismo se colocará los responsables por la *elaboración*, *revisión* (en caso de que la organización lo disponga asi) y la *autorización* del documento con la respectiva fecha y firma. A diferencia del encabezado, no es necesario que se repita en todo el procedimiento. También es recomendable agregar la función o cargo de los responsables.

Los procedimientos fueron diseñados según el modelo dado en la siguiente **Tabla 11:** Modelo de procedimiento [38], cuya estructura se basa en lo establecido en la Norma ISO 10013, explicada en [38].

Tabla 11: Modelo de procedimiento [38]
Fuente: Elaboración propia

<table>
<tr><td rowspan="3">Logo</td><td rowspan="2">Nombre del Procedimiento</td><td>Código: PR-XXX-YY</td></tr>
<tr><td>Revisión: X</td></tr>
<tr><td>Fecha : dd/mm/aaaa</td></tr>
<tr><td colspan="2">Normas: ISO 9001:2015; ISO 10013; BPMN.</td><td>Página: 1 de xx</td></tr>
<tr><td colspan="3">Objetivo del procedimiento
Alcance
Descripción del procedimiento
Diagrama del procedimiento</td></tr>
<tr><td colspan="3">Control de Emisión</td></tr>
<tr><td>Elaboró

Función/Cargo:
AyN:
Fecha:
Firma:</td><td>Revisó

Función/Cargo:
AyN:
Fecha:
Firma:</td><td>Autorizó

Función/Cargo:
AyN:
Fecha:
Firma:</td></tr>
</table>

4.3. Propuesta de aplicación de procedimientos a un proyecto de explotación de la información

Como solución a la problemática planteada en este TFM, se realiza una propuesta de aplicación del procedimiento de explotación de información, que se visualiza en la **Fig. 2:** Propuesta de aplicación del procedimiento de explotación de información.

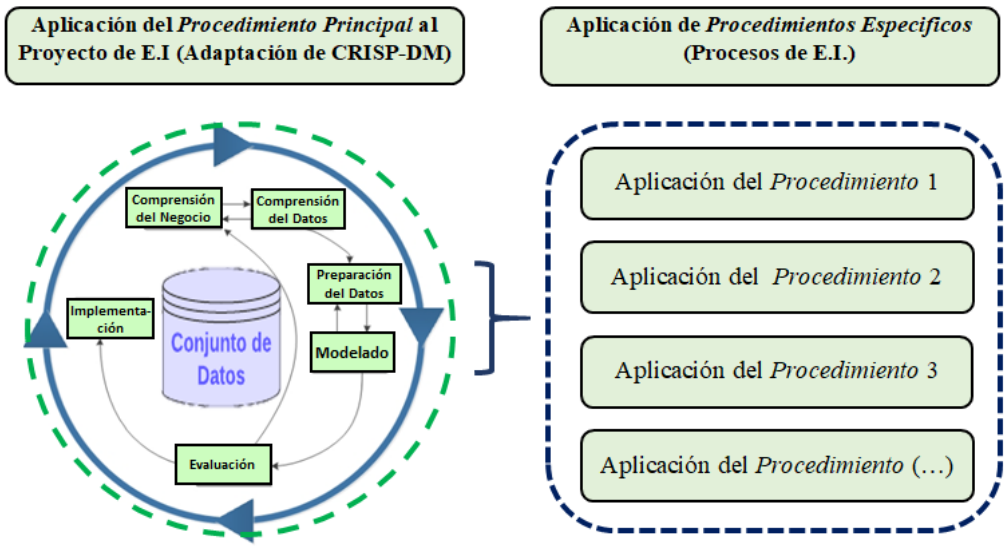


Fig. 2: Propuesta de aplicación del procedimiento de explotación de información.
Elaborado en base a [29].

Esta propuesta (Fig. 2) es una adaptación de CRISP-DM y resume la solución propuesta a la problemática planteada. El procedimiento de explotación de la información diseñado está constituido por:

1)- Un procedimiento principal correspondiente al proyecto de explotación de información, basado en CRISP-DM, que permite sistematizar las actividades a realizar en un proyecto de explotación de la información con CRISP-DM. En la Fig. 2 se puede observar, que el diagrama de la izquierda muestra la aplicación del procedimiento principal a un proyecto de explotación de información, en el que se realiza una adaptación de la metodología CRISP-DM. Como se explica en el capítulo 2, las fases en CRISP-DM no son secuenciales, sino que se puede volver a fases anteriores según lo requiera el trabajo de explotación de datos.

2)- Procedimientos específicos, diseñados en relación con cada objetivo de minería de datos definido y proceso de explotación de información identificado para resolver el problema. A la derecha de la Fig. 2, se visualiza cómo se aplican los procedimientos específicos en el desarrollo del proyecto. Cada uno de estos procedimientos, a su vez, representa otra secuencia de procesos a ejecutar, con

sus correspondientes entradas y salidas de datos, que serán descriptas y diagramadas en el correspondiente procedimiento.

4.4. Procedimiento principal de explotación de la información

El procedimiento principal de explotación de la información se plantea ante la necesidad de sistematizar las actividades a realizar en un proyecto para la explotación de datos de población y el logro del objetivo general de este TFM.

El propósito de este documento es establecer las actividades que abarquen todo el proyecto y sirvan como marco de referencia para proyectos de explotación de la información con datos estructurados de población. A fin de aportar calidad y eficacia al proceso, se seleccionó la metodología CRISP-DM.

Las actividades realizadas en el procedimiento se describen y modelan en un diagrama según el estándar BPMN.

En el diagrama del procedimiento principal, se muestra la representación gráfica de las fases de CRISP-DM. Las tareas a realizar en el proyecto de explotación de la información se representan como un *subproceso* (colapsado), según se indica en [44], y se presentan en otro diagrama con el *subproceso* correspondiente (expandido), con sus entradas, el proceso y las salidas correspondientes.

A continuación se muestra el procedimiento de explotación de la información diseñado, en la **Tabla 12:** Diseño del procedimiento principal de explotación de información.

Tabla 12: Diseño del procedimiento principal de explotación de información.
Fuente: Elaboración propia

<table>
<tr><td rowspan="3">Logo</td><td rowspan="3">Procedimiento de Explotación de la Información
(adaptación de CRISP-DM)</td><td>Código: PR-GEN-00</td></tr>
<tr><td>Revisión: A</td></tr>
<tr><td>Fecha : dd/mm/aaaa</td></tr>
<tr><td>Normas: ISO 9001:2015; ISO 10013; BPMN</td><td>Página: 1 de 6</td></tr>
<tr><td colspan="3">Objetivo: El objetivo de este procedimiento es establecer las actividades necesarias para la ejecución de un proyecto de explotación de la información, mediante la metodología CRISP-DM, que guie en la realización de actividades de explotación de datos estructurados de población, para la detección de problemáticas laborales y sus factores de incidencia y garantice la calidad en los resultados obtenidos.

Alcance: Este procedimiento es aplicable a proyectos de explotación de información, con datos estructurados de población. Permitirá guiar o sistematizar las actividades realizadas en este tipo de proyectos, hasta descubrir patrones útiles o conocimiento en el conjunto de datos de población mediante técnicas de minería de datos, que serán seleccionadas acorde a la problemática planteada y al conjunto de datos disponible.</td></tr>
</table>

Logo	**Procedimiento de Explotación de la Información** (adaptación de CRISP-DM)	Código**:** PR-GEN-00
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 2 de 6

Descripción del procedimiento para un proyecto de explotación de información:

Ante la necesidad de mejorar los resultados obtenidos en un proyecto de explotación de la información se propone la aplicación de este procedimiento, que permite ordenar las actividades a realizar en mismo, mediante el uso de la metodología CRISP-DM (modelo de proceso). En un proyecto de este tipo se puede identificar:

- **Entrada**. La entrada es una base de datos o un conjunto de datos estructurados de población.
- **Proceso**: Para realizar la explotación de datos en un proyecto de explotación de información, se desarrollan las distintas fases de la metodología CRISP-DM (descriptas en el capítulo 2), diagramadas en este procedimiento principal de explotación de la información como *subprocesos colapsados*, debido a que cada *subproceso* involucra varias tareas que serán especificadas y representadas mediante el diagrama de *subproceso expandido* correspondiente.
- **Salida:** Conocimiento. Los modelos obtenidos mediante la aplicación de CRISP-DM y los resultados e informes obtenidos en cada fase que serán documentados a lo largo del proceso (indicadas como objetos de datos o bases de datos en los diagramas).

La metodología CRISP-DM se divide en seis fases que se describen de la siguiente manera:

1)- **Comprensión del negocio**: En esta etapa se define el problema que se va a resolver, se establece el objetivo del negocio (proyecto explotación de la información o investigación) y los objetivos de minería de datos.

2)- **Comprensión de los datos**: Se realiza una exploración detallada de los datos (de *entrada*) para comprender su contenido, calidad y estructura. Se identifican los datos relevantes para el problema definido en la etapa anterior, los datos faltantes o inconsistentes. Se realizan reportes de recopilación, descripción, exploración y calidad de los datos (*salida*).

3)- **Preparación de los datos**: En esta etapa se lleva a cabo la limpieza de los datos (para tratar valores perdidos, valores atípicos y datos duplicados), la selección de los atributos relevantes y la transformación de los datos a un formato adecuado para el análisis.

4)- **Modelado**: En relación al problema que se quiere resolver y los objetivos de minería de datos establecidos, se aplica el procedimiento específico correspondiente, según el proceso de explotación de información determinado y las técnicas de modelado adecuadas.

5)- **Evaluación**: Se evalúan los modelos obtenidos en la fase anterior para determinar su eficacia y calidad. Se analizan los resultados, se detectan posibles limitaciones del modelo y se realiza una revisión del proceso de modelado a fin de identificar oportunidades de mejora.

6)- **Implementación**: Se implementa el modelo para la toma de decisiones, se realiza un seguimiento y mantenimiento del mismo para garantizar su calidad y eficacia. Esta etapa puede reducirse a la documentación y presentación de los resultados del proyecto explotación de la información a las partes interesadas (cliente).

Registro de actividades: Se conserva un registro detallado de todas las actividades llevadas a cabo en cada fase de la metodología CRISP-DM.

Logo	**Procedimiento de Explotación de la Información** (adaptación de CRISP-DM)	Código**:** PR-GEN-00
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 3 de 6

Se documentan las actividades realizadas y los resultados obtenidos. Se archivan los datos originales y los resultados del análisis.

Revisión y mejora continua: Se revisa periódicamente el procedimiento para identificar oportunidades de mejora y se implementan cambios si es necesario, para asegurar su eficacia y eficiencia.

Anexos: Se incluyen los anexos necesarios, como los resultados del análisis y los documentos de soporte.

Aprobación: El procedimiento es aprobado por el equipo de análisis de datos.
El procedimiento puede variar según las necesidades y características del proyecto y los datos disponibles. Es decir, debe ser adaptado a cada problema o caso de estudio.

Diagrama del procedimiento principal de un proyecto de explotación de la información:
Las tareas generales (subprocesos) a realizar en un proyecto de explotación de la información con la metodología CRISP-DM, se grafican de la siguiente manera:
En el diagrama de la **Fig. 3:** Proyecto de explotación de información. Fases de la metodología CRISP-DM, se visualizan las fases de CRISP-DM, aplicadas a un proyecto de explotación de información.

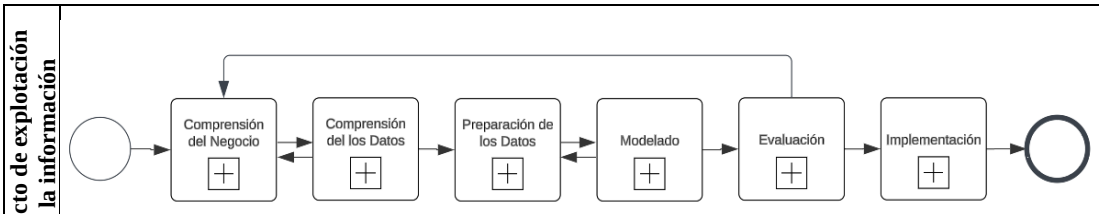


Fig. 3: Proyecto de explotación de información. Fases de la metodología CRISP-DM

Cada fase se muestra como un *subproceso colapsado*, debido a que en cada una de ellas se realizan varias tareas (que se diagraman en este procedimiento mediante el *subproceso expandido* correspondiente). En las siguientes Figuras (4-9), se observan los diagramas de los *subprocesos expandidos*, correspondientes a cada una de las fases de CRISP-DM (Fig.3), con sus tareas y salidas:

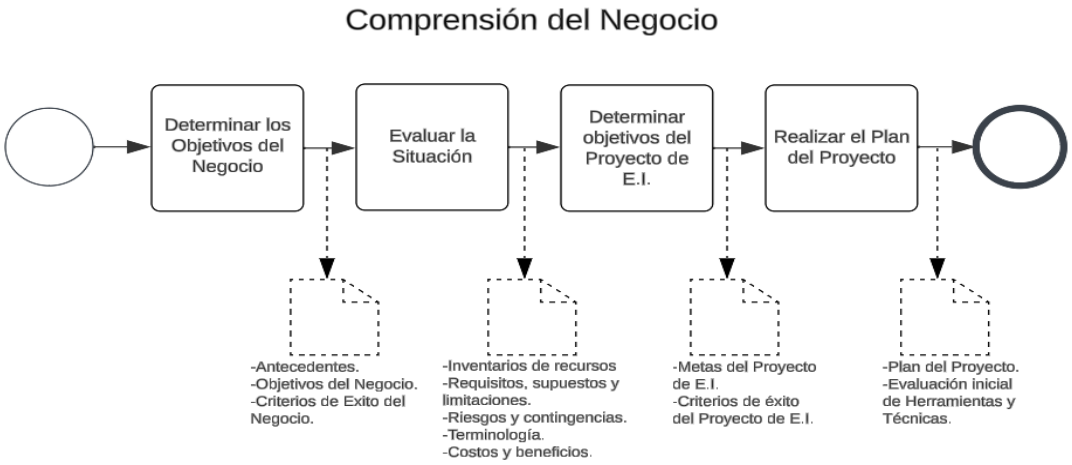


Fig. 4: Subproceso expandido de la fase 1 ‘Comprension del Negocio’

Logo	**Procedimiento de Explotación de la Información** (adaptación de CRISP-DM)	Código**:** PR-GEN-00
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 4 de 6

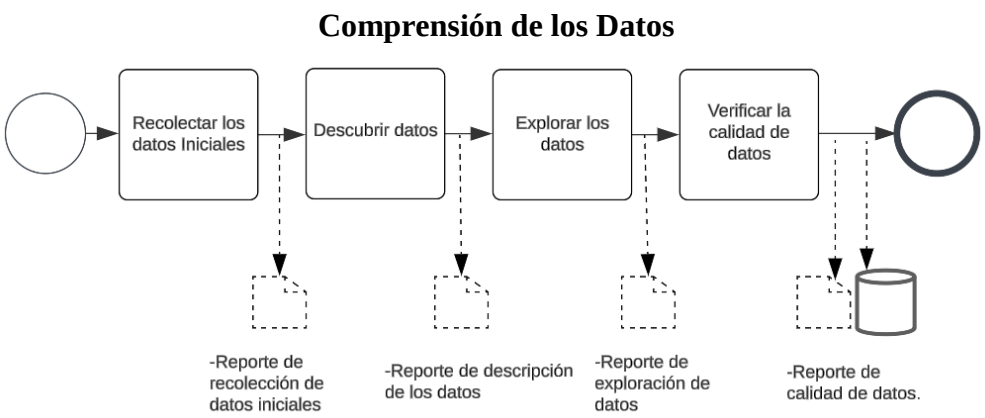


Fig. 5: Subproceso expandido de la fase 2 ‘Comprension de los datos’.

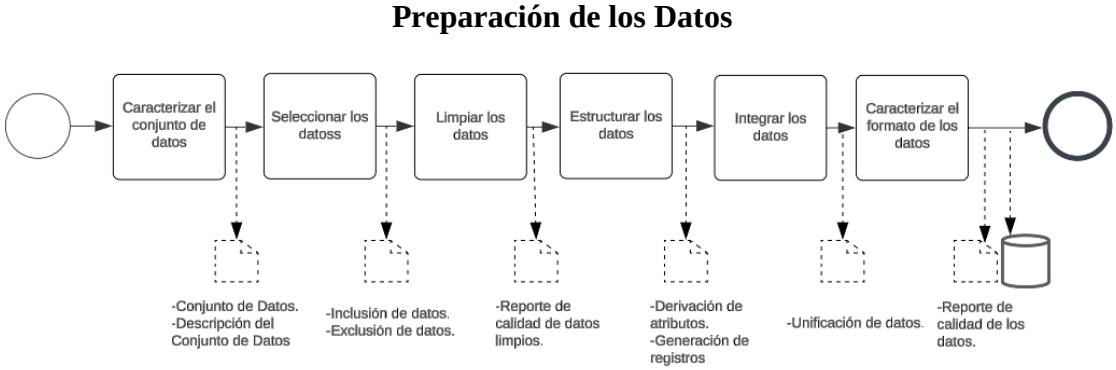


Fig. 6: Subproceso expandido de la fase 3 ‘Preparación de los Datos’.

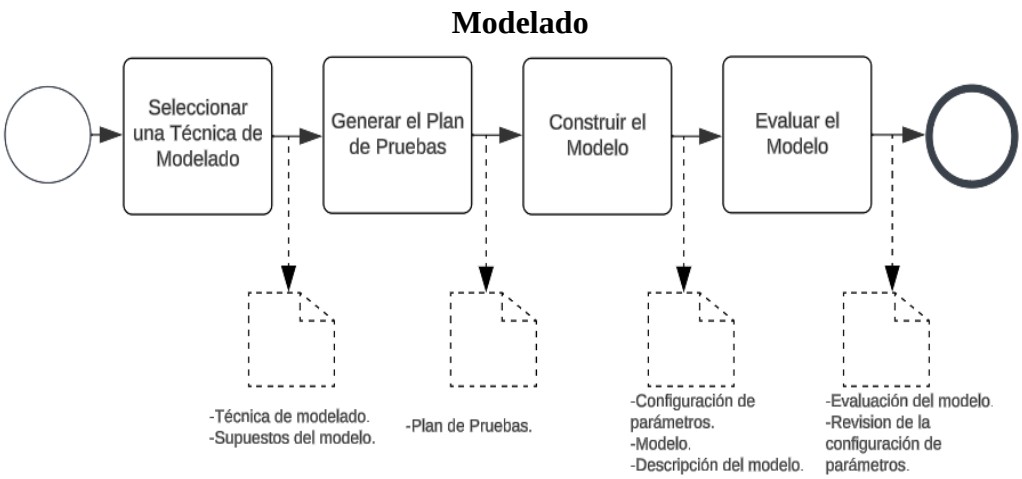


Fig. 7: Subproceso expandido de la fase 4 ‘Modelado’.

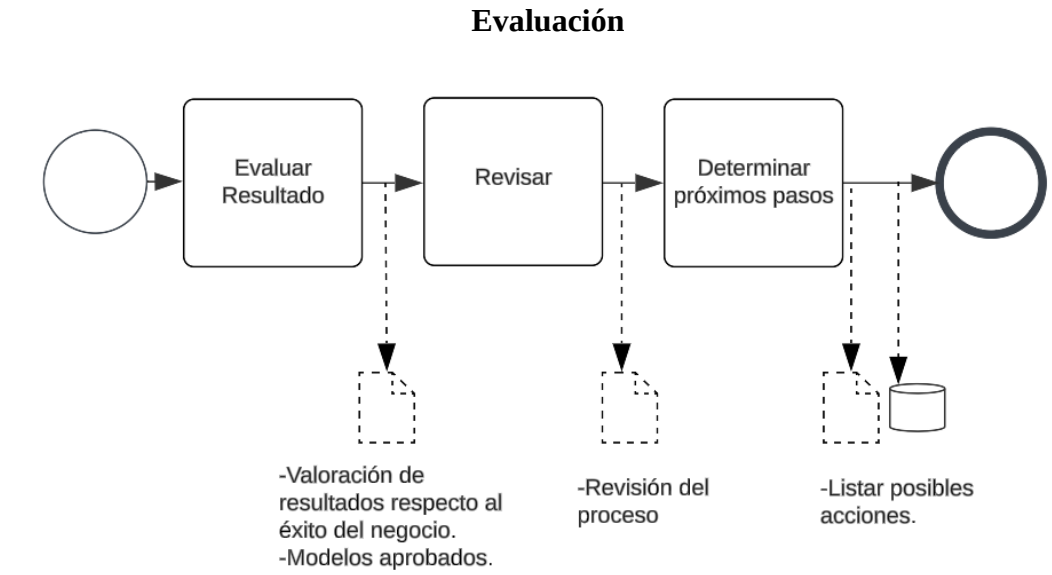


Fig. 8: Subproceso expandido de la fase 5 'Evaluación'.

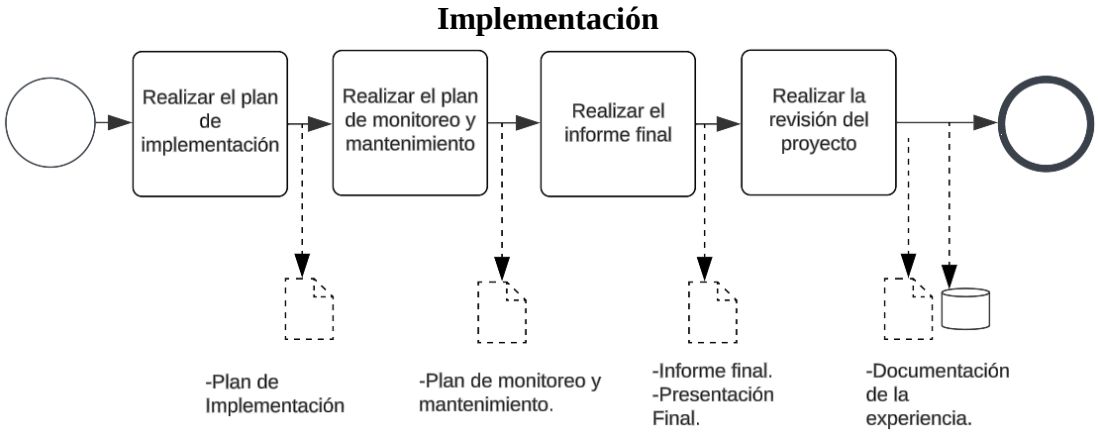


Fig. 9: Subproceso expandido de la Fase 6 'Implementación'.

En los diagramas de los subprocesos expandidos correspondientes a cada una de las fases de CRISP-DM, se visualizan las salidas obtenidas luego de cada tarea o actividad, representadas como objetos de datos o base de datos, que son utilizadas en cada etapa subsiguiente del proceso.

Las salidas de cada fase consisten en documentos digitales o datos en archivos que se modifican en el proceso, según corresponda (y sirven como información de entrada en la siguiente etapa).

Logo	**Procedimiento de Explotación de la Información** (adaptación de CRISP-DM)	Código**:** PR-GEN-00
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 6 de 6

Control de Cambios: se describen los cambios que tuvo el documento (detalle de modificaciones, fecha, revisión, responsables, etc.). El siguiente cuadro muestra el historial de la documentación del procedimiento:

Versión	**Fecha de vigencia**	**Apartado modificado**	**Modificación realizada**	**Responsable**
00	Año 2021	Inicio del documento	Creación del Documento	Analista de explotación de información
01	Año 2022	Todos los apartados	Revisión General	Analista de explotación de la información (con revisión de expertos)
02	Año 2023	Todos los apartados	Revisión General (con experto)	Analista de explotación de la información (con revisión de expertos)

Control de emisión		
Elaboró	**Revisó**	**Autorizó**
Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma:

Para el diseño del procedimiento se usaron contenidos de las normas ISO 9000:2015 [1], ISO 9001:2015 [37], ISO 10013 [38] y la notación BPMN [44].

4.5. Procedimientos específicos

Luego del ordenamiento de actividades a realizar en el proyecto de explotación de la información mediante CRISP-DM, que permitió la definición del procedimiento principal asociado al objetivo de este TFM, en esta sección se diseñan los procedimientos específicos del proyecto de explotación de información, asociados a cada objetivo de minería de datos definido (en relación con cada problema de minería de datos detectado), necesarios para lograr el objetivo general del proyecto.

En base a los problemas de minería de datos detectados, se definen los objetivos de minería de datos (objetivos del proyecto en términos técnicos) y se determina el proceso de explotación de información a aplicar para el logro de cada uno de estos objetivos. Además, según el tipo de problema definido, es posible establecer las técnicas (o algoritmos) a utilizar en cada proceso de explotación de información [4].

Para cada proceso de explotación de información definido, asociado a la

problemática detectada, se diseña un procedimiento específico, mediante el uso de tablas y diagramas con la notación BPMN [44]. En este procedimiento, se establece la tarea a realizar (por ejemplo agrupamiento, clasificación, etc.) en el proyecto de explotación de la información para la obtención de modelos que son analizados e interpretados con la ayuda de expertos.

Se debe tener en cuenta, que la performance de los algoritmos, depende en gran medida de las características del conjunto de datos a analizar y también por los valores de los ejemplos del conjunto de datos. El desafío consiste en hallar los algoritmos que mejor describan el conjunto de datos a analizar.

4.5.1. Descripción del modelo de procedimientos específicos

A fin de diseñar los procedimientos específicos, se deben determinar los problemas de minería de datos y definir los objetivos de minería de datos para solucionarlos. En base a estos objetivos, se determina el proceso de explotación de información a aplicar, según lo definido en [4].

A partir de un conjunto de datos estructurados de población, sobre el que no se tiene conocimiento previo, se propone aplicar inicialmente un enfoque *no supervisado* con el fin de establecer los grupos (clases) y entenderlos. Luego de hallar los grupos, se pueden aplicar enfoques *supervisados* que permitan caracterizar a esos grupos o conocer las características de las personas que los conforman y cómo se relacionan con la clase.

A partir de estas necesidades, se establecen los siguientes objetivos de minería de datos:

Objetivo de minería de datos N° 1: Determinar los distintos *grupos* asociados a conjunto de datos de población y realizar un análisis exploratorio de los mismos, a fin de entenderlos.

Objetivo de minería de datos N° 2: Descubrir las *reglas de comportamiento* que identifiquen a cada grupo (o clase) en un conjunto de datos estructurados de población, a fin de caracterizar dichos grupos.

Objetivo de minería de datos N° 3: Determinar los factores que poseen mayor incidencia o las características distintivas de las personas que pertenecen a un grupo (clase) en un conjunto de datos estructurados de población. Es decir, determinar en qué medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase.

En relación a los objetivos de minería de datos establecidos, en la **Tabla 13**: Propuesta de procedimientos específicos, asociados a los procesos de explotación de información**,** se describen los procedimientos específicos, que presentan la solución a las problemáticas planteadas:

Tabla 13: Propuesta de procedimientos específicos, asociados a los procesos de explotación de información
Fuente: Elaboración propia

Procedimiento	**Proceso de explotación de información**	**Técnica de modelado** (Técnica de minería de datos)
Procedimiento1	En relación al *objetivo de minería de datos N°1*, se quiere identificar *grupos* de personas con características similares, en el conjunto de datos de población y lograr un mayor conocimiento sobre los datos (realizar un análisis exploratorio de los datos). Proceso de explotación de información a aplicar: "Descubrimiento de grupos".	Aprendizaje *no supervisado*: Conjunto de datos sin atributo *clase* u *objetivo* (target) Algoritmos de *segmentación* o *agrupamiento.* Técnicas no supervisadas y herramientas de Visualización.
Procedimiento2	En relación al *objetivo de minería de datos N°2*, se quiere *descubrir las reglas de comportamiento de cada grupo* (clase), que permitan caracterizar a cada grupo. Proceso de explotación de información a aplicar: "Descubrimiento de reglas de comportamiento".	Aprendizaje *supervisado:* Conjunto de datos con atributo *clase* u *objetivo* (target) Algoritmo Árbol de Decisión
Procedimiento3	En relación al *objetivo de minería de datos N°3*, se quiere ponderar en qué medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase. Proceso de explotación de información a aplicar: "Ponderación de interdependencia de atributo".	Aprendizaje *supervisado:* Conjunto de datos con atributo *clase* u *objetivo* (target) Algoritmo Red Bayesiana (Naive Bayes)

En base a los procesos de explotación de información identificados, se pueden determinar los algoritmos a utilizar en cada caso. Los datos se preprocesan para que se puedan aplicar las técnicas elegidas, y se obtienen los modelos asociados a cada procedimiento, los que serán analizados e interpretados con la ayuda de expertos en el dominio.

4.5.2. Procedimiento N° 1

En relación con el objetivo de minería de datos N° 1 y partir de un conjunto de datos estructurados de población, en el que no se conocen las etiquetas o grupos (clases), se requiere aplicar inicialmente un enfoque no supervisado. En este caso, no se tiene

conocimiento previo sobre el conjunto de datos (o no se dispone de ningún criterio de agrupamiento o segmentación de los datos).

Se determinó aplicar el proceso de explotación de información para el descubrimiento de grupos en este procedimiento. Este proceso de explotación de información es útil en los casos en que se necesita identificar una partición en el conjunto de información disponible, ya que permite la identificación de grupos de personas con características similares en el conjunto de datos de población, sobre el dominio de un problema [16]. Para el descubrimiento de grupos se propone el uso de algoritmos de agrupamiento (o segmentación), técnicas de reducción de dimensión y visualización de la información. La aplicación de este procedimiento permitirá además, realizar un análisis exploratorio de los datos a fin de entenderlos.

En esta sección, se presenta el **procedimiento N° 1** para detectar grupos en conjuntos de datos estructurados de población, sin un atributo objetivo (clase), como muestra la siguiente **Tabla 14:** Diseño del procedimiento N° 1 "Descubrimiento de grupos".

Tabla 14: Diseño del procedimiento N° 1 "Descubrimiento de grupos".

<table>
<tr><td rowspan="3">Logo</td><td rowspan="3">Nombre del procedimiento 1:
"Descubrimiento de grupos"
(asociado al objetivo de minería de datos N° 1)</td><td>Código: PR-EI1-01</td></tr>
<tr><td>Revisión: A</td></tr>
<tr><td>Fecha : dd/mm/aaaa</td></tr>
<tr><td>Normas: ISO 9001:2015; ISO 10013; BPMN</td><td>Página: 1 de 2</td></tr>
<tr><td colspan="3">Objetivo de minería de datos N°1: Determinar los distintos grupos asociados a un conjunto de datos de población.
Ante la necesidad de identificar o establecer grupos de personas con las mismas características o con características similares, asociados al conjunto de datos estructurados de población a estudiar, se selecciona el proceso de descubrimiento de grupos para dar solución al problema detectado.

Alcance: Este procedimiento guía las actividades en la aplicación del proceso de descubrimiento de grupos, hasta determinar grupos en el conjunto de datos de población mediante técnicas de minería de datos, seleccionadas acorde a la problemática planteada y al conjunto de datos disponible.

Descripción del procedimiento: En relación con el proceso de explotación de información establecido, se determinan los algoritmos a utilizar y se preprocesan los datos para adaptarlos de manera que puedan ser procesados por los algoritmos. Este procedimiento debe ser adaptado según el conjunto de datos utilizado y el problema a resolver.

Proceso de explotación de información "Descubrimiento de grupos":

Entrada: Conjunto de datos (estructurados) de población.

Proceso: Actividades realizadas por el algoritmo o técnica seleccionada, para el logro del objetivo de minería de datos planteado. Se aplicará una técnica no supervisada que permita el agrupamiento (o segmentación) o visualización de datos, con el fin de detectar los grupos en el conjunto de datos de población y realizar un análisis exploratorio de los datos.
El proceso de descubrimiento de grupos, se resume en los siguientes pasos:
1)- Integración del conjunto de datos, a fin de que se puedan aplicar los algoritmos de agrupamiento o técnicas de reducción de dimensión y visualización de la información en la identificación de grupos.</td></tr>
</table>

Logo	**Nombre del procedimiento 1: "Descubrimiento de grupos"** (asociado al objetivo de minería de datos N° 1)	Código**:** PR-EI1-01
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 2 de 2

2)- Actividades realizadas mediante el algoritmo o técnica seleccionada, para el logro del objetivo de minería de datos planteado. Aplicación de la técnica de agrupamiento o técnica de reducción de dimensión y visualización de la información al conjunto de datos estudiado para la identificación de grupos.

3)- Generación el archivo con los grupos descubiertos (o visualización de los grupos identificados).

Salida: Como resultado del proceso se obtendrá un archivo con los grupos indentificados y un reporte con el análisis exploratorio que incluya visualizaciones e interpretaciones de los modelos obtenidos, analizados e interpretados con la ayuda de expertos en el dominio y en la disciplina.

Diagrama del procedimiento N° 1:

En relación con el objetivo de minería de datos N° 1 y el proceso de explotación de información elegido, el procedimiento N° 1 se representa mediante el diagrama de la **Fig. 10:** Diagrama del proceso "Descubrimiento de grupos":

Proceso: Descubrimiento de Grupos

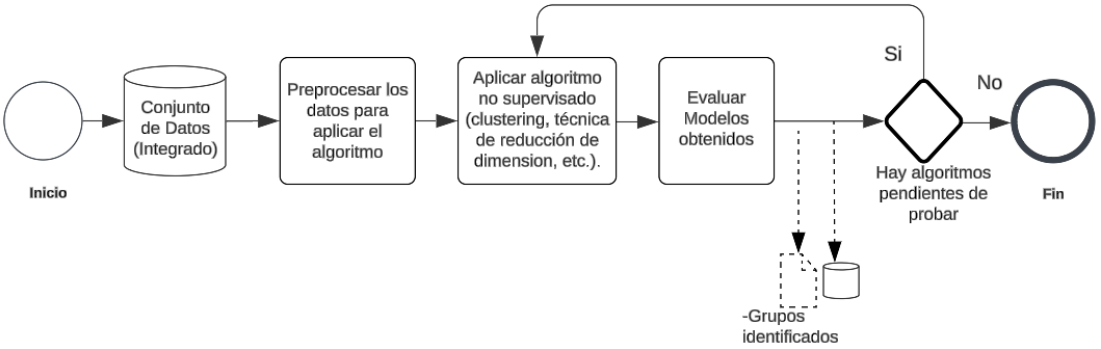


Fig. 10: Diagrama del proceso "Descubrimiento de grupos"

En el diagrama del procedimiento N° 1 (Fig. 10), la salida se identifica como un objeto de datos (reporte o informe) o como una base de datos (archivo con los grupos descubiertos o identificados). El diagrama del procedimiento N°1 es la representación gráfica de la secuencia de actividades a realizar para el logro del objetivo N° 1.

Control de Emisión		
Elaboró	**Revisó**	**Autorizó**
Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma:

4.5.3. Procedimiento N° 2

En relación al objetivo de minería de datos N° 2 y partir de un conjunto de datos estructurados de población, en el que se conocen las etiquetas o grupos (clases), se requiere aplicar un enfoque supervisado, a fin de descubrir las características que identifican a un grupo o establecer las características de las personas que conforman

esos grupos y obtener modelos que los representen. Se determinó entonces, aplicar el proceso de explotación de información para el *descubrimiento de reglas de comportamiento*. Este proceso se usa cuando se quiere identificar las características que determinan un resultado específico en el dominio del problema. Por ejemplo, se puede usar para establecer las características que determinan (o caracterizan) a las personas que conforman un grupo en un conjunto de datos de población. Es decir, la aplicación de este proceso permitirá descubrir las reglas de comportamiento o características que identifican a un grupo.

En esta sección, se presenta un procedimiento para descubrir reglas de comportamiento en conjuntos de datos estructurados de población, con un atributo objetivo (clase), el cual se muestra en la **Tabla 15:** Diseño del procedimiento N° 2 'Descubrimiento de reglas de comportamiento'.

Tabla 15: Diseño del procedimiento N° 2 'Descubrimiento de reglas de comportamiento'.

<table>
<tr><td rowspan="3">Logo</td><td rowspan="3">Nombre del procedimiento 2:
"Descubrimiento de reglas de comportamiento"
(asociado al objetivo de minería de datos N° 2)</td><td>Código: PR-EI2-02</td></tr>
<tr><td>Revisión: A</td></tr>
<tr><td>Fecha : dd/mm/aaaa</td></tr>
<tr><td>Normas: ISO 9001:2015; ISO 10013; BPMN</td><td>Página: 1 de 2</td></tr>
<tr><td colspan="3">Objetivo: Descubrir las reglas de comportamiento que identifiquen a cada grupo (o clase) en un conjunto de datos estructurados de población (PEA), a fin de caracterizar a dichos grupos.
Se determina el proceso de explotación de información "Descubrimiento de reglas de comportamiento" para dar solución al problema de minería de datos detectado, asociado al objetivo de minería de datos N° 2 definido.

Alcance: Este procedimiento es aplicable a proyectos de explotación de la información con datos estructurados de población (PEA). Permite guiar o sistematizar las actividades a realizar hasta descubrir reglas de comportamiento o patrones útiles, mediante la aplicación de técnicas de minería de datos, que serán seleccionadas acorde a la problemática planteada y al conjunto de datos disponible.

Descripción del procedimiento: En relación con el proceso de explotación de información, se determina que se va a utilizar el algoritmo de Árbol de Decisión. Se preparan los datos para que estos puedan ser procesados por el algoritmo elegido. La aplicación de esta técnica permitirá determinar las características de las personas que conforman cada grupo; como resultado, se obtendrán modelos (árbol de decisión), que serán analizados e interpretados con la ayuda de expertos en el dominio y en la disciplina.

Proceso de explotación de información "Descubrimiento de reglas de comportamiento":

Entradas: Conjunto de datos (estructurados) de población, con grupos identificados (clase).

Proceso: Actividades a realizar por el algoritmo de árbol de decisión, que permitirá dar solución al problema de minería de datos detectado. Es decir, se hará la explotación de datos sobre el conjunto de datos que se adaptó para poder aplicar la técnica seleccionada e identificar las reglas de comportamiento de cada atributo clase que permita caracterizar a los grupos.</td></tr>
</table>

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Logo	**Nombre del procedimiento 2: "Descubrimiento de reglas de comportamiento"** (asociado al objetivo de minería de datos N° 2)	Código**:** PR-EI2-02
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 2 de 2

El proceso de descubrimiento de reglas de comportamiento, se resume en los siguientes pasos:

1)- Integración y preparación del conjunto de datos (estructurados) de población, para la aplicación del algoritmo árbol de decisión, asociado al proceso de explotación de información definido.

2)- Identificación del atributo clase.

3)- Aplicación de la algoritmo árbol de decisión sobre el conjunto de datos con un atributo clase.

Salida: Como resultado del proceso se obtendrá el modelo que será analizado e interpretado con la ayuda de expertos en el dominio y en la disciplina. Se pueden obtener varios modelos, a partir de la aplicación del algoritmo, con distintos parámetros.

Diagrama del procedimiento: En relación con el problema detectado y al objetivo de minería de datos N° 2, se presenta el diagrama del procedimiento N°2 que se muestra en la

Fig. 11: Diagrama del proceso "Descubrimiento de reglas de comportamiento", que es la representación gráfica de la secuencia de actividades a realizar en el proceso de descubrimiento de reglas de comportamiento para el logro del objetivo de minería de datos N° 2, sus entradas, los procesos y las salidas correspondientes:

Proceso: Descubrimiento de reglas de comportamiento

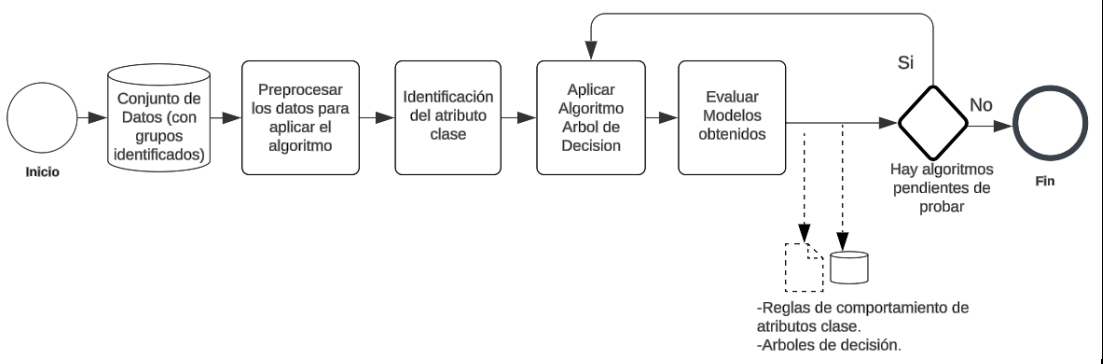


Fig. 11: Diagrama del proceso "Descubrimiento de reglas de comportamiento"

En el diagrama del procedimiento N° 2 se identifica la salida como un objeto de datos (reporte o informe) o como una base de datos (archivo), correspondiente a las reglas de comportamiento de cada atributo clase y árbol de decisión generados.

Control de emisión

Elaboró	**Revisó**	**Autorizó**
Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma

4.5.4. Procedimiento N° 3

A partir de un conjunto de datos estructurados de población, en el que se conocen los grupos (o clases), se propone aplicar un enfoque *supervisado*, a fin de determinar en qué medida la variación de los valores de una variable incide sobre la

variación del valor del atributo clase (o identificar las características distintivas de las personas que pertenecen a un determinado grupo).

Se determinó entonces, la necesidad de aplicar el proceso de explotación de información "*Ponderación de interdependencia de atributos"* (o descubrimiento de atributos significativos). Este proceso es útil para determinar en qué medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase. Es decir, permite identificar los factores o atributos significativos que sean distintivos de las personas que pertenecen a un grupo (o clase), en un conjunto de datos estructurados de población. En este proceso, se propone el uso de Redes Bayesianas.

En esta sección, se presenta un procedimiento para realizar el proceso de ponderación de interdependencia de atributos en conjuntos de datos estructurados, con un atributo objetivo, que se muestra en la **Tabla 16:** Diseño de procedimiento 'Proceso de ponderación de interdependencia de atributos'.

Tabla 16: Diseño de procedimiento 'Proceso de ponderación de interdependencia de atributos'.

Logo	**Nombre del procedimiento 3: "Ponderación de interdependencia de atributos"** (asociado al objetivo de minería de datos N° 3)	Código**:** PR-EI3-03
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 1 de 2

Objetivo: Determinar los factores que poseen incidencia o las características distintivas de las personas que pertenecen a un grupo o clase en un conjunto de datos estructurados de población (PEA).

Ante la necesidad de determinar los factores que poseen incidencia o las características distintivas de las personas que pertenecen a un grupo (o clase) en un conjunto de datos estructurados de población, se selecciona el proceso de explotación de información "Ponderación de interdependencia de atributos" para dar solución al problema detectado.

Alcance: Este procedimiento se aplica a proyectos de explotación de información, con datos estructurados de población. Permite guiar o sistematizar las actividades realizadas para la aplicación del proceso "Ponderación de Interdependencia de Atributos", hasta descubrir patrones útiles o conocimiento mediante la técnica de minería de datos seleccionada.

Descripción del procedimiento: En relación con el proceso de explotación de información establecido, se determina que se utilizará el algoritmo de Redes Bayesianas (Naive Bayes); se preparan los datos para adaptarlos y que puedan ser procesados por el algoritmo. Este procedimiento debe ser adaptado según el conjunto de datos utilizado y el problema a resolver.

Proceso de explotación de información "Ponderación de interdependencia de atributos":

Entrada: Conjunto de datos (estructurados) de población.

Proceso: Actividades a realizar por el algoritmo de redes bayesianas (Naive Bayes), que permitirá dar solución al problema detectado.

El proceso de ponderación de interdependencia de atributos, se resume en los siguientes pasos:

1)- Integración y preparación del conjunto de datos (estructurados) para la aplicación del algoritmo Naive Bayes (red bayesiana) en el proceso de explotación de información definido.

Logo	**Nombre del procedimiento 3: "Ponderación de interdependencia de atributos"** (asociado al objetivo de minería de datos N° 3)	Código**:** PR-EI3-03
		Revisión**:** A
		Fecha **:** dd/mm/aaaa
	Normas: ISO 9001:2015; ISO 10013; BPMN	Página**:** 2 de 2

2)- Identificación del atributo clase.
3)- Aplicación del algoritmo Naive Bayes sobre el conjunto de datos con un atributo clase.

Salida: Como salida del proceso se obtendrá el modelo obtenido a partir de la aplicación del algoritmo de redes bayesianas (Naive Bayes), que será analizado e interpretado con la ayuda de expertos. Se pueden obtener varios modelos, con distintos parámetros.

Diagrama del procedimiento 3: En relación con el problema detectado y el objetivo de minería de datos N° 3 definido, se presenta el diagrama del **procedimiento 3**, que se muestra en la **Fig. 12:** Diagrama del proceso "Ponderación de interdependencia de atributos", que es la representación gráfica de la secuencia de actividades a realizar para el logro del objetivo de minería de datos N° 3, mediante el proceso de explotación de información elegido, sus entradas, los procesos y las salidas (mostrada como base de datos o un objeto de datos).

Proceso: Ponderación de interdependencia de atributos

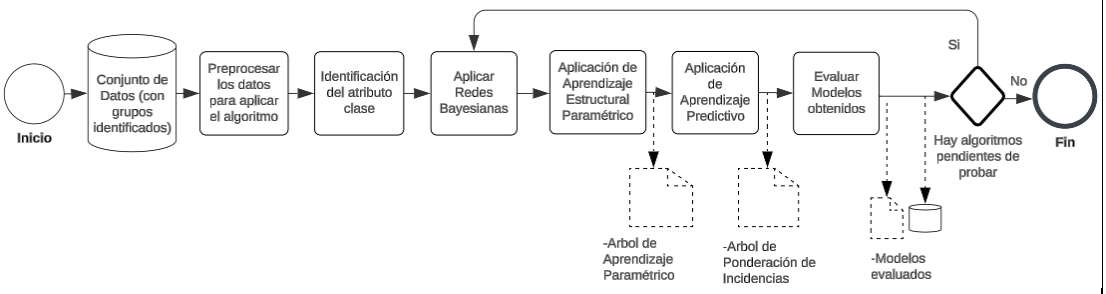


Fig. 12: Diagrama del proceso "Ponderación de interdependencia de atributos"

Control de emisión

Elaboró	**Revisó**	**Autorizó**
Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma:	Función/Cargo: AyN: Fecha: Firma:

4.6. Validación del procedimiento de explotación de información

En el capítulo anterior se presentó una propuesta metodológica que consiste en el diseño de un procedimiento de explotación de la información, fundamentado en la necesidad de establecer una sistematización de actividades a realizar en un proyecto de explotación de la información.

Este procedimiento consta de un procedimiento principal que establece una forma ordenada de ejecución de proyectos de explotación de la información. Además, como en el desarrollo o ejecución de un proyecto de explotación de la información se pueden identificar diferentes problemas de minería de datos, se establecen los objetivos de minería de datos y se diseña un procedimiento específico asociado a cada objetivo de minería de datos definido y proceso de explotación de información

identificado para resolver el problema [4]. Es decir, estos procedimientos específicos se aplican dentro del procedimiento principal para resolver los problemas de minería de datos, que permiten dar solución al objetivo general de este TFM.

En este capítulo se valida el procedimiento de explotación de la información, con sus procedimientos principal y específicos.

4.6.1. Caso de estudio

La validación del procedimiento propuesto se realizó mediante la experimentación con datos reales, obtenidos en el relevamiento realizado en el Barrio Industrial de la ciudad de Corrientes. En el relevamiento sociodemográfico de población y viviendas, realizado en este barrio de la zona norte de la ciudad de Corrientes, se obtuvo información sobre la población, como ser, características habitacionales, situación laboral de la misma, nivel educativo, composición familiar de los residentes del barrio, entre otras.

El relevamiento estuvo a cargo del equipo técnico de la Dirección de Estadística y Censos en cooperación con la Dirección de Sistemas y Tecnologías de Información, ambas pertenecientes a la Subsecretaría de Sistemas y Tecnologías de Información (SUSTI), dependiente del Ministerio de Hacienda y Finanzas de la Provincia de Corrientes, en colaboración con la Carrera de Licenciatura en Relaciones Laborales dependiente de la Facultad de Ciencias Económicas, de la Universidad Nacional del Nordeste (UNNE). En el marco de este operativo, alumnos de la carrera antes mencionada realizaron sus trabajos finales de carrera.

El operativo se llevó a cabo en el mes de octubre del año 2016 y fue diseñado en base al Censo Nacional de Población, Hogares y Viviendas desarrollado por el INDEC en el año 2010. Según este censo, la cantidad de viviendas en el Barrio Industrial era de 478, que comprendían 498 hogares particulares, con una población total del barrio de 1.809 habitantes (869 varones y 940 mujeres).

Se eligió este barrio porque en el mismo existe una heterogeneidad de sectores sociodemográficos y económicos, motivo por el cual, con la información recabada del mismo se puede reflejar las características y necesidades de barrios similares de la ciudad de Corrientes. El cuestionario usado para dicho relevamiento se adjunta en el Anexo B de este trabajo.

Para la validación, se tuvo en cuenta la propuesta de aplicacion de procedimientos, presentada en el capítulo 4 (Fig. 2).

En la validación del procedimiento de explotación de la información, se describen las tareas a realizar en un proyecto mediante la aplicación del procedimiento principal (adaptación de CRISP-DM) y los procedimientos específicos que se ejecutan dentro del mismo.

4.6.2. Validación del procedimiento principal

Objetivo: Establecer las tareas necesarias para la explotación de datos estructurados de población correspondiente al relevamiento sociodemográfico realizado en el Barrio Industrial de la ciudad de Corrrientes, mediante el uso de la metodología CRISP-DM. Esta metodología permite organizar las tareas de hallar patrones y obtener conocimiento de los datos, lo cual ayuda a detectar las problemáticas laborales en la población estudiada (PEA) y sus factores de incidencia.

Alcance: Este procedimiento es aplicable a proyectos de explotación de la información con datos estructurados. Se validó este procedimiento, con los datos del relevamiento sociodemográfico realizado en el Barrio Industrial de la ciudad de Corrientes, donde se estudió la PEA en dicho conjunto de datos. El procedimiento principal diseñado servirá como guía de las actividades a realizar en el proyecto de explotación de la información, hasta descubrir patrones útiles o conocimiento en el conjunto de datos, mediante técnicas de minería de datos.

Las técnicas a utilizar son seleccionadas, en relación a las problemáticas planteadas y al conjunto de datos disponible. Es decir, la aplicación de este procedimiento puede variar según las necesidades y características del proyecto y los datos disponibles, por lo que debe ser adaptado a cada caso de estudio.

Anexos: en el anexo de este TFM se incluye información sobre el mercado de trabajo en Argentina, el relevamiento sociodemográfico realizado en el “Barrio Industrial” y los cuestionarios usados en el mismo, se incluye un glosario con terminología utilizada en este TFM, se exponen algunos enfoques en explotación de información e información sobre el conjunto de datos obtenido en el relevamiento realizado en el Barrio Industrial.

Aprobación: El procedimiento es aprobado por el equipo de análisis de datos.

Este procedimiento organiza la ejecución de un proyecto de explotación de la información según CRISP-DM, con:

- **Entrada**: Conjunto de datos estructurados del relevamiento sociodemográfico realizado en el Barrio Industrial de la ciudad de Corrientes.

- **Proceso**: en el proyecto de explotación de la información se desarrollan las distintas fases de la metodología CRISP-DM mostradas en el diagrama de la Fig. 3 como subprocesos colapsados, representados con sus subprocesos expandidos en las Fig. 4-9.
- **Salida:** Conocimiento. Los modelos obtenidos mediante la aplicación de CRISP-DM y los resultados o reportes obtenidos en cada fase que serán documentados a lo largo del proceso (indicados como objetos de datos o bases de datos en los diagramas de subprocesos expandidos).

Descripción de las tareas a realizar en el procedimiento de explotación de la información:

En el procedimiento principal de explotación de la información propuesto en este TFM, se presenta el diagrama de un proyecto de explotación de la información (Fig. 3) basado en CRISP-DM, donde cada una de las etapas se muestra como un subproceso colapsado, debido a que cada una incluye varias tareas. Estos subprocesos se representan en el procedimiento mediante sus correspondientes diagramas de subprocesos expandidos (Fig. 4 hasta Fig. 9).

A continuación, se describen las tareas del proyecto de explotación de la información (adaptación de CRISP-DM), mediante la aplicación del procedimiento principal de explotación de la información, los procedimientos específicos diseñados, y se documentan los resultados obtenidos:

Fase 1: Comprension del Negocio:

En el diagrama presentado en la Fig. 4 se muestra el subproceso expandido de la fase de comprension del negocio, donde se visualizan las tareas generales de CRISP-DM y sus salidas, presentadas en el diagrama como objetos de datos (o bases de datos, según corresponda). Este subproceso se desarrolla a continuación para el caso en estudio:

1.1. Determinar los objetivos de negocio: Esta primer tarea del subproceso se refiere a entender y establecer los objetivos que se quieren lograr (cliente), desde el punto de vista de esta investigación (negocio).

Negocio: TFM (investigación).

Cliente: Departamento de Relaciones del Trabajo, Facultad de Ciencias Económicas, Universidad Nacional del Nordeste.

Caso: Barrio Industrial. Población a estudiar (PEA de zonas urbanas).

Para lograr esto, se describen las siguientes tareas específicas asociadas (salidas):

1.1.1. Antecedentes: Para entender el negocio (investigación), es necesario analizar la información disponible sobre la situación actual de la población estudiada, a partir del conjunto de datos del relevamiento sociodemográfico del Barrio Industrial.

Ante los problemas que existen en el ámbito laboral, siendo la desocupación uno de ellos, es fundamental detectar y estudiar las problemáticas relacionadas a la situación laboral de las personas, a fin de entenderlas y atenderlas desde diversos ámbitos, entre ellos el académico.

Sin embargo, ante la escasez de estudios actuales en la región, realizados específicamente sobre la Población Económicamente Activa mediante el uso de explotación de la información, en este TFM se busca analizar la relevancia de estas herramientas de explotación de información para poder implementarlas en investigaciones o temáticas relacionadas al campo laboral. La investigación en este campo es particularmente impulsada por el Departamento de Relaciones del Trabajo (de la FCE, de la UNNE).

En este contexto, se considera a la explotación de información como una herramienta fundamental para el análisis de datos de población, por lo que se aplica y valida el procedimiento de explotación de la información diseñado, al caso de estudio del Barrio Industrial. Se busca encontrar patrones de comportamiento en el conjunto de datos del relevamiento sociodemográfico realizado en este barrio, a fin de conocer esta población y detectar las problemáticas que se presentan en relación a la situación laboral de la población estudiada.

En este caso de estudio, la comprensión del negocio se realizó a partir de los datos obtenidos del relevamiento que fueron otorgados por el Departamento de Relaciones del Trabajo (de la FCE, de la UNNE) para su uso en este TFM. Los datos del relevamiento fueron aportados a dicho departamento por la Dirección de Estadística y Censos, con la finalidad de que sean usados en investigaciones, trabajos de fin de carrera y tesis de posgrado realizadas en el ámbito de la carrera de Licenciatura en Relaciones Laborales, de la Facultad de Ciencias Económicas, de la Universidad Nacional del Nordeste (Anexo B: Información sobre Relevamiento B° Industrial).

A partir de un análisis inicial se determinan y evalúan las características del proyecto de explotación de la información a realizar sobre el caso de estudio del Barrio Industrial, los recursos humanos involucrados y la estimación de posibilidad de éxito de este proyecto.

1.1.2. Objetivos de negocio

Se identifican los objetivos principales del cliente (Departamento de Relaciones del Trabajo, de la FCE, de la UNNE) desde una perspectiva no técnica.

En relación con objetivo de este TFM, en esta investigación se diseñó un procedimiento para sistematizar las tareas a realizar en un proyecto de explotación de la información mediante una metodología. Este procedimiento se aplica para el análisis de los datos del relevamiento del Barrio Industrial, con el fin de obtener conocimiento sobre la PEA que habita en el mismo, mediante técnicas de minería de datos.

Es necesario validar el procedimiento de explotación de la información diseñado a partir de los datos estructurados de población del relevamiento sociodemográfico del Barrio Industrial, a fin de verificar que dicho procedimiento guie y facilite la ejecución de actividades en el proyecto de explotación de datos, permita obtener conocimiento y revele patrones de comportamiento en los datos.

1.1.3. Criterios de éxito del negocio:

En esta investigación, el criterio de éxito se evaluará a partir de los resultados obtenidos con la aplicación de este procedimiento de explotación de la información, resultados que serán analizados y valorados por expertos en el dominio, expertos en el área de Sistemas y en el área de Relaciones del Trabajo.

Se considera como criterio de éxito para el proyecto de explotación de la información:

- Que los procedimientos generados faciliten y ordenen las tareas de explotación de la información mediante una metodología, a fin de asegurar la obtención de resultados de calidad en el proyecto, aplicado al estudio de la población del Barrio Industrial. Es decir, se espera obtener conocimiento de los datos o patrones de comportamiento (modelos), que permitan conocer las problemáticas laborales y sus factores de incidencia en la PEA estudiada.

- Que los Procedimientos diseñados puedan ser adaptados y usados en el estudio de otras poblaciones similares.
- Que estos procedimientos sirvan de antecedente para otros trabajos de investigación en el campo de las Relaciones del Trabajo y que promuevan el uso de herramientas de análisis de datos en estudios de población.
- Que los conocimientos obtenidos en este TFM, sirvan como insumo o posibiliten que, desde el Departamento de Relaciones del Trabajo, se planifiquen acciones para mejorar la situación de la población, en relación con los problemas detectados en el ámbito laboral o mejorar la empleabilidad de esta población mediante cursos, talleres de oficios y planificación de capacitaciones ofrecidas a los habitantes del Barrio Industrial.
- En el ámbito académico (Departamento de Relaciones del Trabajo, de la FCE, de la UNNE), que sirva de antecedente para la actualización del plan de estudios de la carrera de Relaciones Laborales, en relación con el uso de herramientas, métodos y propuestas de posgrados.

1.2. Evaluar la situación: Se analizan las restricciones y factores que se deben tener en cuenta en el proyecto. Para lograr esto, se realizan las siguientes tareas específicas asociadas (salidas):

1.2.1. Inventario de recursos: Los recursos con los que se cuenta para este estudio (recursos de hardware, software, de fuentes de datos y humanos) se exponen en la siguiente **Tabla 17:** Recursos del proyecto de explotación de la información.

Tabla 17: Recursos del proyecto de explotación de la información
Fuente: Elaboración propia

Recursos	Descripción
Recursos de Hardware	• Una notebook con acceso a internet, procesador AMD, 8GB RAM. • Una notebook con acceso a internet, procesador Pentium, 4GB RAM. • Una impresora multifunción Epson.
Recursos de Software	• Se trabajará con herramientas de software libre para minería de datos: software Orange, con paquetes instalados para la explotación y modelado de datos. • Herramientas de trabajo colaborativo y software de ofimática para la elaboración de reportes. • Sistema Operativo Windows.
Recurso de Datos/ Información	• Acceso a repositorios para la búsqueda de información. • Fuente de información: Datos del relevamiento sociodemográfico realizado en el Barrio Industrial, de la ciudad de Corrientes. Se cuenta con la información en un archivo de formato .xls, conformado por varias hojas de cálculo, con la información recabada en el relevamiento realizado en año 2016 (el formulario o cuestionario usado se expone en el Anexo B).

Recursos	Descripción
Recursos humanos	• Un Analista de explotación de información • Expertos en el dominio, en la disciplina y en el área de Relaciones del Trabajo: ✓ Dra. en Matemáticas y Computación. Master Universitario en Matemáticas y Computación, Especialidad Minería de Datos y Sistemas Inteligentes. Docente en la Carrera Licenciatura en Relaciones Laborales. Facultad de Ciencias Económicas, UNNE. Ex. Personal Técnico de la Delegación Regional INDEC NEA. Integrante del equipo técnico del operativo del Relevamiento Barrio Industrial. ✓ Experto en Estadística y Computación. Jefe de la Delegación Regional INDEC NEA. Ex. Director a Cargo de la Dirección de Estadística y Censos de la ciudad de Corrientes. ✓ Licenciada en Relaciones Laborales. Docente en la Carrera Licenciatura en Relaciones Laborales. Facultad de Ciencias Económicas, UNNE.

1.2.2. Requisitos, supuestos y limitaciones:

- **Requerimientos**: Los resultados del proceso se representan de una forma clara, simple y entendible para los usuarios, con visualizaciones y descripciones de los resultados obtenidos.
- **Supuestos**:
  - ✓ Los datos del relevamiento (caso de estudio), son reales.
  - ✓ Las fuentes de datos son accesibles en todo momento.
- **Restricciones (del proyecto):**
  - ✓ Algunas de las limitaciones pueden estar dadas por la características del conjunto de datos (la generalización es relativa a los datos y al problema de minería de datos). El conjunto de datos a analizar corresponde al año 2016 (fecha del relevamiento). No se dispone de datos actuales.
  - ✓ Se identifican muchos datos sobre los que no se tienen respuestas o no corresponde que contesten algunos encuestados. Por lo que en estos campos no hay valores, lo que puede representar una limitación para el software.

    Se decidirá en cada caso, si corresponde que estos datos sean imputados.

1.2.3. Riesgos y contingencias: Se determinan los posibles riesgos que se pueden dar en el proyecto, el impacto que genera y las acciones de contingencia propuestas en cada caso, que se exponen en la siguiente **Tabla 18:** Riesgos y contingencias del proyecto de explotación de la información.

Tabla 18: Riesgos y contingencias del proyecto de explotación de la información.
Fuente: Elaboración propia

Riesgo	Descripcion	Impacto	Acciones de contingencia
1	Que el tipo de datos del relevamiento dificulte la aplicación de las técnicas necesarias.	Medio	• Preparar o adaptar los datos para que puedan ser usados por los algoritmos.
2	La baja calidad de los datos dificulta la aplicación de técnicas de minería de datos.	Alto	• Mejorar la calidad de los datos. • Preprocesar los datos de modo que representen lo mejor posible la realidad del problema.
3	Los patrones encontrados no logran alcanzar los objetivos propuestos.	Alto	• Utilizar distintas técnicas de minería de datos en la fase de modelado. • Usar distintos parámetros a fin de mejorar los modelos obtenidos.
4	Los patrones detectados no son comprendidos.	Medio	• Realizar un análisis exploratorio de los datos, a fin de entenderlos (con el experto en el dominio). • Obtener distintas representaciones o visualizaciones de los datos a fin de que los patrones encontrados sean entendibles.
5	Los datos disponibles no son lo suficientemente actuales.	Bajo	Realizar distintas visualizaciones a fin de comprender los resultados obtenidos y poder aplicarlos e interpretarlos en otros contextos.

1.2.4. Terminología: al inicio de este TFM se incluye una lista de abreviaturas usadas en este informe. Además, como parte de la comprensión de esta investigación, se agrega y describe en el Anexo C un glosario con terminologías utilizadas en este TFM.

1.2.5. Análisis de costos y beneficios del proyecto:

Costos:

- El proyecto de explotación de la información se desarrolla en el contexto de este TFM (trabajo de investigación), por lo que los costos son los propios del uso de la PC e Internet.
- No se generan costos adicionales en este proyecto, debido a que el software usado es de código abierto y los datos pertenecen al un organismo público (Departamento de Relaciones del Trabajo, de la FCE, de la UNNE).

Beneficios:

- El estudio realizado servirá para adquirir mayor conocimiento sobre la condición laboral de la PEA del Barrio Industrial y se usará como base para promover estudios desde la Universidad (Departamento de Relaciones del Trabajo, de la FCE, de la UNNE), que utilicen nuevas herramientas de análisis como la explotación de información y la minería de datos.

- La información generada en este proyecto, servirá para orientar la toma de decisiones en el Departamento de Relaciones del Trabajo en cuanto a la actualización de contenidos del nuevo plan de estudios de la carrera de Relaciones Laborales. Asimismo, a partir del conocimiento obtenido en este TFM, se podrán estudiar y proponer nuevas líneas de investigación o cursos de posgrado.

1.3. Determinar los objetivos del proyecto de explotación de la información:

La primera fase de análisis del problema, incluye la comprensión de los objetivos del proyecto desde una perspectiva de la investigación, con el fin de convertirlos en objetivos de minería de datos (en términos técnicos) y en una planificación [22].

Para lograr esto, se realizan las siguientes tareas específicas asociadas (salidas):

1.3.1.Objetivos de minería de datos:

A partir del análisis de los objetivos de este trabajo y el repositorio de datos, se delimita el alcance del proyecto en un conjunto de objetivos de minería de datos que se abordan mediante la aplicación de un procedimiento específico asociado a un proceso de explotación de información y con los algoritmos de minería de datos correspondientes. Los problemas de minería de datos representan los objetivos del proyecto en términos técnicos (objetivos de minería de datos), y la respuesta que se obtiene de cada problema de minería de datos permite lograr los objetivos de este TFM [5].

Al analizar la fuente de información disponible, es muy importante determinar la estructura y naturaleza de los datos, debido a que esta comprensión permite determinar el tipo de solución a proponer.

En este contexto, se definen los siguientes objetivos de minería de datos para este proyecto de explotación de la información. En relación a cada objetivo, se aplica el procedimiento específico correspondiente:

Objetivo de minería de datos N° 1: determinar grupos en el conjunto de datos del relevamiento del Barrio Industrial.

Ante la necesidad de identificar o establecer los grupos de personas con las mismas características o con características similares en el conjunto de datos sociodemográfico correspondiente al relevamiento realizado en el Barrio Industrial, se selecciona el proceso de explotación de información "Descubrimiento de grupos".

Para aplicar este proceso de explotación de información se ejecuta el *Procedimiento n° 1.* Se trabajó con un enfoque no supervisado a fin de conocer los datos mediante un análisis exploratorio inicial y construir modelos que permitan detectar los grupos o segmentar los datos en el conjunto de datos del relevamiento del Barrio Industrial.

Objetivo de minería de datos N° 2: Determinar las reglas de comportamiento de cada clase o identificar las características (o variables) que distinguen a las personas de cada grupo.

Ante la necesidad de determinar las reglas de comportamiento de cada grupo en el conjunto de datos sociodemográfico del relevamiento del Barrio Industrial, se seleccionó el proceso de explotación de información "Descubrimiento de reglas de comportamiento". Para aplicar este proceso, se ejecutó el *Procedimiento N° 2*, en el que se propone la aplicación de un enfoque supervisado.

Objetivo de minería de datos N° 3: determinar o identificar los factores o atributos con mayor incidencia (o frecuencia de ocurrencia) sobre la clase (objetivo).

Ante la necesidad de determinar la frecuencia de incidencia de los atributos (variables explicativas) sobre la variable explicada u objetivo, se seleccionó el proceso de explotación de información "Ponderación de interdependencia de atributos". Para aplicar este proceso de explotación de información, se ejecutó el *Procedimiento N° 3,* en el que se propone la aplicación de un enfoque supervisado.

1.3.2. Criterios de éxito de explotación de la información

Se definen los siguientes criterios de éxito para el proyecto de explotación de la información (desde el punto de vista técnico), que especifican, en relación con cada objetivo de minería de datos definido, las condiciones bajo las cuales se aceptarán los resultados obtenidos:

Criterio de éxito para el objetivo de minería de datos N° 1: se espera que, con la aplicación del *Procedimiento N° 1* (que propone un enfoque no supervisado), se logre detectar grupos u obtener una segmentación de los datos que permita identificar y entender los distintos grupos en el conjunto de datos sociodemográfico del Barrio Industrial. Asimismo, la aplicación de este procedimiento permitirá realizar un análisis exploratorio de los datos.

Criterio de éxito para el objetivo de minería de datos N° 2: se espera que, con la aplicación del *Procedimiento N° 2* (enfoque supervisado), se obtengan modelos con una buena capacidad predictiva, con una tasa de acierto superior al 70%, que permitan establecer reglas de comportamiento para caracterizar a cada grupo.

Criterio de éxito para el objetivo de minería de datos N° 3: se espera que, con la aplicación del *Procedimiento N° 3* (enfoque supervisado), se obtengan modelos con una buena capacidad predictiva, con una tasa de acierto superior al 70%, que permitan determinar los factores (variables explicativas o independientes) con mayor incidencia en la determinación de la clase (variable explicada) en el conjunto de datos del Barrio Industrial.

1.4. Crear el plan del proyecto para minería de datos:

Se realiza la planificación para el proyecto de explotación de la información en relación a los objetivos planteados.

1.4.1. Plan del proyecto (Salida)

En este plan, se listan las tareas que se ejecutan en el proyecto de explotación de la información y la duración de las mismas; es un documento dinámico, que se revisa y ajusta durante el desarrollo del proyecto.

El siguiente plan de trabajo muestra la duración de ejecución del proyecto de explotación de la información, que fue estimado inicialmente en 24 meses como se visualiza en la **Tabla 19:** Plan de proyecto.

Tabla 19: Plan de proyecto
Fuente: Elaboración propia

Proyecto de explotación de la información (adaptación de CRISP-DM)	Meses																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Entendimiento del negocio	X	X	X	X																				
Comprensión de los datos			X	X	X	X																		
Preparación de los datos					X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
Modelado											X	X	X	X	X	X	X	X	X	X	X			
Evaluación													X	X	X	X	X	X	X	X	X	X		
Implementación (elaboración de reportes)											X	X	X	X	X	X	X	X	X	X	X	X	X	X

1.4.2. Evaluación inicial de herramientas y técnicas

En este apartado de realiza la evaluación inicial de herramientas y técnicas de minería de datos que se pueden utilizar en el proyecto para el logro de los objetivos de minería de datos y, de este modo, cumplir con los objetivos de este TFM.

El proyecto se desarrolló con herramientas de software libre. Para la fase de análisis de datos, modelado y visualización se utilizó el software Orange.

En función de los problemas detectados y los objetivos de minería de datos definidos, se determinaron los procesos de explotación de información a aplicar. Se debe tener en cuenta que, para ciertos tipos de problemas, algunas técnicas son más apropiadas que otras.

En los procedimientos específicos diseñados se establecen, para cada proceso de explotación de información identificado, las técnicas de minería de datos a utilizar.

Las técnicas empleadas en cada procedimiento específico, en relación con el objetivo de minería de datos definido y el proceso de explotación de información identificado, fueron:

Procedimiento 1: En este procedimiento se aplicó un enfoque no supervisado que, mediante el proceso de explotación de información de "Descubrimiento de Grupos", permitió identificar los grupos en el conjunto de datos del relevamiento del Barrio Industrial.

En el conjunto de datos del relevamiento del Barrio Industrial no se conoce la salida (clase); por lo tanto, con la aplicación de este procedimiento se identificaron los grupos o clases.

Mediante la segmentación se pueden visualizar los grupos en el conjunto de datos, en los cuales todos los elementos del grupo comparten características comunes. Las técnicas apropiadas para segmentar son técnicas de agrupación o segmentación y visualización.

Se realizó un análisis exploratorio de los datos a partir de la aplicación de una técnica descriptiva y distintas visualizaciones que provee la herramienta Orange, con el objetivo de obtener una primera visión de los datos. La exploración y el análisis inicial de los datos pueden ayudar a entender la naturaleza de los mismos y determinar los grupos en el conjunto de datos del relevamiento.

Procedimiento 2: En este procedimiento se aplicó un enfoque supervisado, que mediante el proceso de "Descubrimiento de Reglas de Comportamiento", permitió hallar las reglas de comportamiento de cada grupo (clase) en el conjunto de datos del relevamiento del Barrio Industrial (PEA).

La técnica aplicada es la clasificación (modelado predictivo). El objetivo es lograr modelos de clasificación (como el árbol de decisión) que clasifiquen correctamente la clase ante objetos no previstos anteriormente.

Los rótulos de las clases pueden ser definidos por el usuario o derivados de la segmentación. Se seleccionó esta técnica debido a que el conjunto de datos del relevamiento, a partir de la aplicación del procedimiento específico 1, cuenta con un conjunto de observaciones en las que se conoce la salida (o clase) requerida para aplicar esta técnica.

Para seleccionar las técnicas de clasificación, es importante considerar la naturaleza de los datos. En el conjunto de datos analizado, las variables explicativas son en su mayoría categóricas o binarias (dicotómicas, con valores 0 o 1 que indican ausencia o presencia), solo dos son numéricas y la variable explicada (dependiente) es categórica. Por ello, las técnicas candidatas son el Árbol de Decisión y las Redes Bayesianas (aplicadas en el Procedimiento 3).

En este procedimiento se usó el algoritmo árbol de decisión y además, distintas visualizaciones que provee el software Orange, con el fin de interpretar los resultados obtenidos.

Procedimiento 3: En este procedimiento se aplicó un enfoque supervisado, que mediante el proceso de "ponderación de interdependencia entre atributos", permitió identificar las características distintivas (factores de incidencia) de las personas que pertenecen a un grupo o clase en el conjunto de datos del relevamiento sociodemográfico del Barrio Industrial (PEA).

Como se explicó anteriormente, la selección de técnicas de clasificación depende de la naturaleza de los datos y una de las técnicas candidatas para este conjunto de datos fue la red bayesiana. Por ello, en este procedimiento se usó el algoritmo Naive Bayes, además de distintas visualizaciones que provee el software Orange, con el fin de presentar e interpretar los resultados obtenidos.

Fase2: Comprensión de los datos

En esta etapa se realizaron acciones para familiarizarse con los datos, a fin de entender los mismos. Las tareas generales de esta fase con sus salidas son:

2.1. Recolectar los datos iniciales:

En el caso de estudio los datos porvienen del relevamiento realizado en el Barrio Industrial en el año 2016.

2.1.1. Informe inicial de recopilación de los datos (*salida*):

En esta fase se definió el conjunto de datos que se va a utilizar en el proyecto.

En la exploración se observó que el conjunto de datos cuenta con los registros de cada formulario que se registró en el relevamiento realizado en el Barrio Industrial a partir de los cuestionarios publicados en el Anexo B.

La fuente de datos del relevamiento del Barrio Industrial, se encuentra en varias planillas de cálculo. Se realizó la integración de datos con la ayuda de un experto en el dominio; en esta integración se consideraron las variables más relevantes para el problema en estudio, a fin de poder trabajar con ellas en los siguientes pasos (ver Anexo E: Datos del relevamiento del Barrio Industrial). El archivo resultante es una planilla con una sola hoja de cálculo (en formato .xls), que cuenta con 703 filas (ejemplos) y 16 columnas (variables o características) de las cuales dos son numéricas y las demás categóricas.

2.2. Describir los datos:

El análisis de datos del relevamiento sociodemográfico del Barrio Industrial se inicia con la descripción de las variables. Como se visualiza en la **Fig. 13:** Variables del conjunto de datos, el set de datos cuenta con 703 registros (ejemplos) con 16 variables (características), de las cuales 14 variables son categóricas y 2 variables son numéricas. Se puede notar que la mayoría de las variables son categóricas. Además, se observa que existen valores perdidos en el conjunto de datos.

Info

703 instances
16 features (23.5% missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	Sexo	categorical	feature	1, 2
2	Años	numeric	feature	
3	**CoberturaSalud**	**categorical**	**feature**	
4	JubilacionPensi...	categorical	feature	1, 2
5	**AsisteEstEducat**	**categorical**	**feature**	
6	**NivelEducat**	**categorical**	**feature**	
7	Computadora	categorical	feature	1, 2
8	Internet	categorical	feature	1, 2
9	ConviveEnPareja	categorical	feature	1, 2
10	Trabajo1hr	categorical	feature	1, 2
11	EnlasUlt4semBT	categorical	feature	1, 2
12	ChangaVenta	categorical	feature	1, 2

Reset

703

Fig. 13: Variables del conjunto de datos

2.2.1.Informe de descripción de los datos (salida):

En este informe se describen los datos de la planilla de cálculo. En la **Tabla 20:** Variables del conjunto de datos del relevamiento del Barrio Industrial, se muestran las variables con su descripción, tipo y valores posibles.

Tabla 20: Variables del conjunto de datos del relevamiento del Barrio Industrial

Variable	Descripción	Tipo	Valores
Sexo	Sexo de la persona encuestada	Categórica	**1**=Varon **2**=Mujer
Años	Edad de la persona encuestada	Numérica	1-99
CoberturaSalud	Cobertura de Salud de la persona encuestada	Categórica	**1**=Obra social (Incluye Pami) **2**=Prepaga a través de obra social. **3**=Prepaga solo por contratación voluntaria. **4**=Programas o planes estatales de salud. **5**=No tiene obra social, prepaga o plan estatal.
JubilacionPension	¿Recibe Jubilación o Pensión?	Categórica	**1**=Si **2**=No
AsisteEstEducat	¿Asiste o Asistió a un establecimiento educativo?	Categórica	**1**=Asiste **2**=Asistió **3**=Nunca Asistió
NivelEducat	Nivel educativo de la persona encuestada	Categórica	**1**=Inicial (Jardín, Pre-escolar) **2**=Primario **3**=EGB **4**=Secundario **5**=Polimodal **6**=Superior no Universitario **7**=Universitario **8**=Post Universitario **9**=Educación Especial
Computadora	¿Utiliza computadora?	Categórica	**1**=Si **2**=No
Internet	¿Utiliza Internet?	Categórica	**1**=Si **2**=No
ConviveEnPareja	¿Convive en pareja o matrimonio?	Categórica	**1**=Si **2**=No
Trabajo1hr	Durante la semana pasada, ¿trabajó por lo menos una hora?	Categórica	**1**=Si **2**=No
BTrab4Ultsem	En las últimas 4 semanas ¿estuvo buscando trabajo: contestó avisos, consultó amigos/parientes, puso carteles, hizo algo para ponerse por su cuenta?	Categórica	**1**=Si **2**=No
ChangaVta	En esa semana ¿hizo alguna changa, algo para vender fuera o ayudo a un familiar/amigo en una chacra o negocio?	Categórica	**1**=Si **2**=No

Variable	Descripción	Tipo	Valores
LicVacaciones	En esa semana ¿tenía trabajo, pero estuvo de licencia por vacaciones, enfermedad, suspensión, conflicto laboral, etc.?	Categórica	1=Si 2=No
Dtojubilacion	En ese trabajo, ¿le descuentan para la jubilación?	Categórica	**1**=Si **2**=No **3**=Ignorado
AportaIndep	En ese trabajo, ¿aporta por si mismo para la jubilación?	Categórica	**1**=Si **2**=No **3**=Ignorado
CantHijos	Cantidad de hijos vivos	Numérica	**1-9**

2.3. Explorar los datos:

Se realiza la exploración de los datos a fin de observar la distribución y comportamiento de las variables con mayor relevancia para este estudio.

Mediante visualizaciones se detectan atributos claves o conjuntos de datos interesantes para el logro de los objetivos de minería de datos del proyecto.

2.3.1. Reporte de exploración de los datos (salida):

En este reporte se realizó el análisis detallado de los datos del relevamiento del Barrio Industrial con el fin de detectar los atributos relevantes o conjuntos de datos que sean interesantes para el logro de los objetivos del proyecto de explotación de la información.

Se exponen los resultados de la exploración de los datos realizada en el software Orange que permite analizar la distribución de datos y comportamiento de las variables más relevantes para este estudio. La **Fig. 14:** Distribución de las variables del conjunto de datos, muestra la salida de Orange donde se puede analizar la distribución de las variables del conjunto de datos y la cantidad de valores ausentes en el mismo (ver Anexo E):

Fig. 14: Distribución de las variables del conjunto de datos

En los histogramas que se muestran en la imagen (Fig. 14) se observa la distribución de las variables (categóricas y numéricas) en relación con la variable *Trabajo1hr* (“*Durante la semana pasada, ¿trabajó por lo menos una hora?”*), que es la variable más importante para la determinación de la situación laboral “ocupada” de una persona [2]. Se puede notar que:

- El color celeste representa la frecuencia de las personas que ***sí*** *trabajaron* al menos *una hora durante la semana anterior al relevamiento.*
- El color rojo representa la frecuencia de las personas que ***no*** *trabajaron* al menos *una hora durante la semana anterior al relevamiento.*

Además, en el caso de las variables categóricas, se muestra la distribución de frecuencias. Por ejemplo, si se analiza el histograma correspondiente a la variable *Sexo***,** se observa que:

- Hay más mujeres que varones en el conjunto de datos del relevamiento, lo cual se evidencia por el valor de la moda = 2, que se muestra en la imagen.
- Además, para esta variable no hay valores faltantes (o perdidos).

El análisis de la distribución de frecuencias, proporciona una idea inicial sobre las variables del conjunto de datos.

Respecto a la variable *Años* (*edad de la persona encuestada*), que es una variable numérica, se observa que la edad promedio en el conjunto de datos es de 37 años (donde la edad mínima es de 10 años y máxima es de 75 años). Además, la edad más frecuente es 21 años.

Si se realiza una segmentación de la variable *Trabajo1hr* (cuyas respuestas posibles son: **1**= "Si"; **2**= "No"; **" "**=No Contesta), en relación con la variable *Sexo*, se puede observar en la **Fig. 15:** Segmentacion de datos (respecto a la variable *Trabajo1hr*), las agrupaciones de los datos:

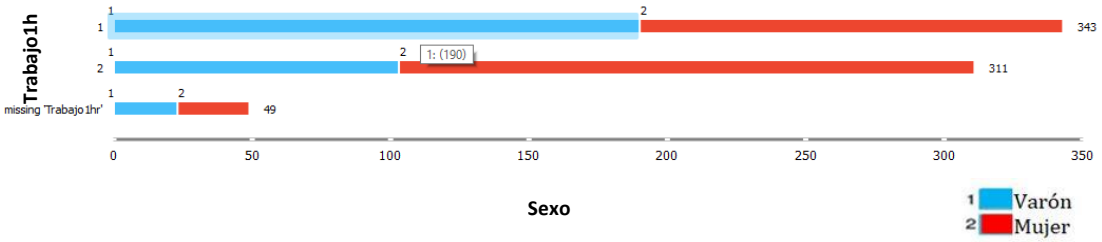


Fig. 15: Segmentacion de datos (respecto a la variable *Trabajo1hr*)

Si se analiza la Fig.15, se puede ver que hay más varones *ocupados* que mujeres (190 casos de 343).

Además, en el grupo de personas que contestaron que *no trabajaron una hora la semana anterior al relevamiento*, la mayoría son mujeres (208 casos de 311). También, se visualiza que existe un grupo importante de valores perdidos (49 valores), que son las personas que no respondieron la pregunta sobre si trabajaron una hora la semana anterior al relevamiento.

Estos registros que poseen valores ausentes o no poseen información sobre si la persona trabaja o no, se analizaron con los expertos, respecto a otras variables para determinar su relevancia para este estudio.

Las variables numéricas del conjunto de datos son *Años* (*Edad de la persona*) y *CantHijos* (*Cantidad de hijos de la persona*).

En la **Fig. 16:** Distribución de los datos (respecto a Edad y Cantidad de hijos), se puede visualizar cómo están distribuidos los datos en relación con estas dos variables.

Las edades de las personas van desde los 10 hasta los 75 años y se muestran agrupadas en intervalos de 10 años, representados en distintos colores. En el conjunto de datos las personas tienen entre 10 y 75 años.

Además, los varones se ven representados con un círculo (1=Varón) y las mujeres con una "x" (2= Mujer), cuyas referencias se muestran en la esquina superior derecha de la Fig. 16.

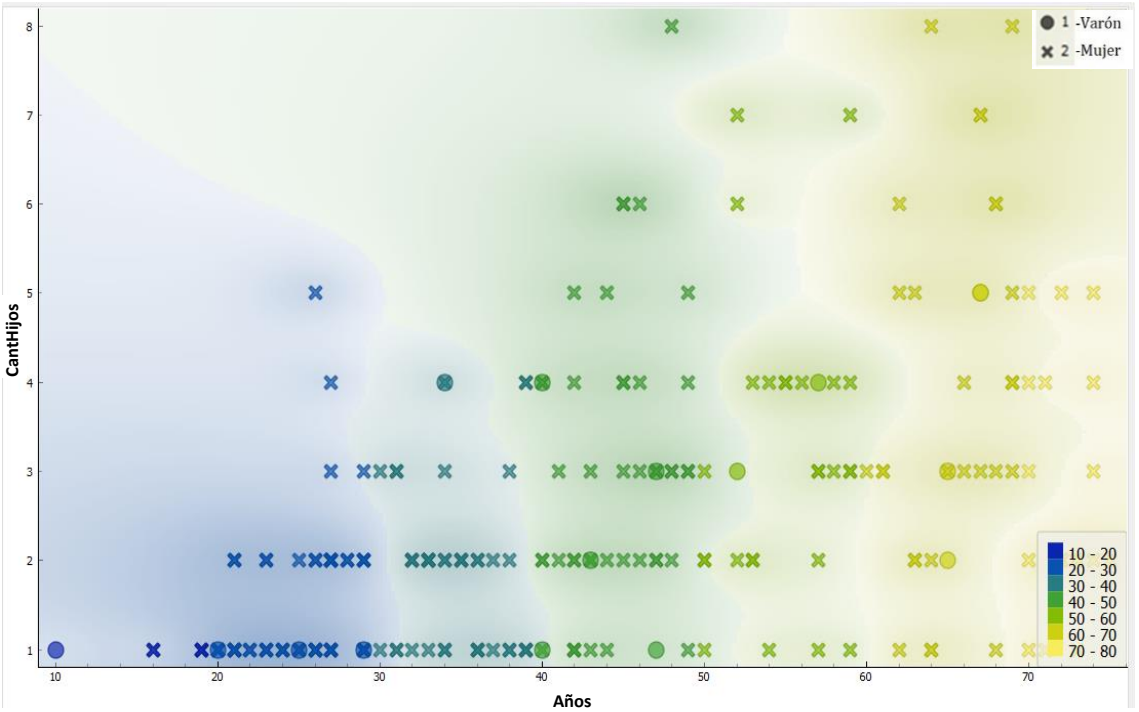


Fig. 16: Distribución de los datos (respecto a Edad y Cantidad de hijos)

2.4.Verificar la calidad de los datos.

Se examinó la calidad de los datos del relevamiento y se analizaron los mismos en relación a su completitud, los errores que se presentan en los mismos y los datos ausentes.

2.4.1.Informe de calidad de los datos (salida):

En este reporte se explica la calidad de los datos, como ser la completitud, los errores en los datos y los valores faltantes. Se documenta el análisis de calidad realizado y las posibles soluciones a los problemas encontrados.

Se detectaron varios problemas en el conjunto de datos, como ser, datos ingresados con errores, abreviaturas o símbolos, datos expresados de manera diferente, los que fueron analizados y corregidos manualmente a fin de unificar los formatos y posibilitar o mejorar su tratamiento por el software de minería de datos (ver Anexo E).

Algunas variables fueron eliminadas del conjunto de datos, debido a que no aportan información relevante para el estudio realizado, por ejemplo:

- Variables que aportan información sobre los encuestados, como: *¿Qué te gustaría hacer al terminar el colegio secundario?* o *¿Tenés acceso a Internet en forma fluida?* (correspondientes al Bloque VI: "Expectativas laborales de jóvenes de 15 a 19 años que asistan a colegio secundario", del cuestionario presentado en el Anexo B).
- Variables relacionadas con las preguntas: *Donde Ud. Trabja ¿lo evalúan de alguna forma?* o *¿en su futuro laboral ¿Qué le gustaría hacer?* (correspondientes al Bloque VIII: "Empleabilidad" del cuestionario presentado en el Anexo B).

Entre los problemas más frecuentes detectados están los valores ausentes, resultantes de la falta de respuesta de las personas en el relevamiento o en otros casos, no correspondía que contesten una determinada pregunta (ver Anexo E).

Se realizó un análisis detallado del conjunto de datos del relevamiento, con la ayuda de expertos, para decidir qué hacer con los valores faltantes. Se decidió su inclusión en el conjunto de datos, si es que son relevantes para este estudio; o su eliminación si no aportan información relevante.

En cuanto a los datos anómalos, las variables numéricas Edad y Cantidad de hijos, no presentan datos extremos.

Fase3: Preparación de los Datos

Esta fase tiene como objetivo obtener la vista minable de los datos del relevamiento realizado en el Barrio Industrial. Es decir, se realizaron tareas de limpieza, formateo e integración de los datos recolectados, con el fin de aplicar sobre los mismos las técnicas de modelado.

El trabajo sobre los datos se realizó varias veces sin seguir un orden determinado, para elegir los registros (ejemplos) y variables (características) que necesitan ser

transformados o limpiados, con el objetivo de adaptar el conjunto de datos para que se puedan aplicar las técnicas de modelado.

3.1. Especificar el conjunto de datos

En esta instancia, se describe el conjunto de datos del relevamiento del Barrio Industrial, que se va a utilizar para el modelado y para el análisis en el proyecto de explotación de la información.

3.1.1.Conjunto de datos (salida)

Se analizaron con expertos las distintas variables del conjunto de datos según su relevancia. En este trabajo se consideraron las variables que aportan información importante al estudio realizado, con el fin de determinar el conjunto de datos que se utilizará en el modelado y el análisis principal del proyecto.

3.1.2. Descripción del conjunto de datos (salida)

Como en este estudio se analiza la PEA del Barrio Industrial, es importante incluir información sobre la situación laboral de cada persona, información que no se tiene en el conjunto de datos.

La variable *Trabajo1hr* (*Durante la semana pasada, ¿trabajó por lo menos una hora?*), es fundamental para determinar si una persona está *ocupada*. Esta variable, junto con la variable *BTrab4Ultsem* **(***En las últimas 4 semanas ¿estuvo buscando trabajo: contestó avisos, consultó amigos/parientes, puso carteles, hizo algo para ponerse por su cuenta?),* son muy importantes para determinar la situación de desocupación de una persona en el conjunto de datos estudiado [2]. Se incluyeron en el conjunto de datos todas las variables necesarias para definir si una persona está *ocupada* o *desocupada*.

3.2. Seleccionar los datos:

En esta instancia, se seleccionaron los datos que se usaron en el análisis. En esta etapa, se debe definir con qué atributos o variables (columnas) y con qué observaciones (filas o registros) se va a trabajar. Esta elección se realizó en función de la importancia que poseen para cumplir con los objetivos del proyecto y la calidad de los mismos; selección que se justificará en cada caso.

3.2.1. Justificación de la inclusión /exclusión de los datos (salida).

En este reporte, se explican las causas por las cuales se incluyeron y excluyeron los datos para el análisis, según la importancia que poseen para el cumplimiento de los objetivos del proyecto y la calidad de los mismos.

Inclusión de datos: Los datos que se incluyeron en el análisis son aquellos que aportan información relevante para este estudio. Como se realizó un estudio sobre la PEA, se incluyeron todas las variables importantes que permitieron estudiar dicha población.

La PEA, por definición [2] esta formada por las personas "ocupadas" y "desocupadas". La variable *Trabajo1hr***,** es una variable fundamental para la determinación de la situación laboral "ocupada" o "desocupada" de una persona, que es la situación que interesa fundamentalmente en este estudio. Entonces, todas las variables que aporten información sobre la PEA o la determinación de la situación laboral de una persona fueron incluidas en el conjunto de datos del relevamiento sociodemográfico del Barrio Industrial, estudiado en este TFM.

Asimismo, como en el conjunto de datos no existe información sobre la situación laboral de las personas, con los expertos se consideró importante incluir la variable *SitLaboral* (situación laboral de una persona) al conjunto de datos.

La variable *Trabajo1h* tiene 49 datos ausentes, entre estos registros (o ejemplos), se consideró relevante conservar los datos de las personas que:

- Realizaron sus aportes jubilatorios, por lo que tienen un trabajo formal y forman parte de la PEA.
- Los que realizaron changas o algún trabajo informal, por lo que se considera que tienen un trabajo informal.

Exclusión de Datos: A partir del análisis del conjunto de datos, se observó que existen muchos valores ausentes o faltantes, detectados por el software Orange como valores perdidos (ver Anexo E). Los registros que no poseen información relevante para este estudio fueron excluidos del conjunto de datos.

Esta exclusión e inclusión de datos se realizó a partir de un análisis detallado del conjunto de datos del relevamiento y con la ayuda de expertos en el dominio.

3.3. Limpieza de datos: Esta etapa que tiene como objetivo mejorar la calidad de los datos, por lo que se tomaron decisiones respecto a los problemas de calidad encontrados en los mismos, como datos ausentes o datos anómalos.

3.3.1. Informe de limpieza de los datos (salida): En este reporte se incluyeron las decisiones tomadas sobre los problemas de calidad de los datos.

Existen valores ausentes en el conjunto de datos, que pueden afectar negativamente al modelo resultante, motivo por el cual se determinó, junto con los expertos, la forma de tratarlos. En el caso de datos ausentes (valores perdidos) que no aportan información al estudio que se está realizando, se eliminan del conjunto de datos.

En el caso de los valores que aportan información para este estudio, se consideró la posibilidad de realizar una imputación de los valores ausentes.

Las variables que tienen mayor porcentaje de valores ausentes son: *AportaIndep* (69%), *CantHijos* (67%), *BTrab4Ultsem* (59%), *DtoJubilacion* (57%), *LicVacaciones* (54%), *ChangaVenta* (50%), algunas de las cuales se explican en el Anexo E.

Estas variables son importantes para definir la situación laboral de una persona; por ello, junto con los expertos se decidió conservar estos datos y generar nuevos atributos a partir de los mismos.

3.4. Construir los datos: En esta fase se lleva a cabo la construcción de nuevos datos, derivados de los disponibles, que son importantes para el análisis.

3.4.1. Atributos derivados (salida) Estos atributos se calculan a partir de otros atributos del mismo registro.

Como se expuso anteriormente, para el estudio de la PEA es necesario considerar la situación laboral de una persona y el tipo de trabajo que realiza en relación con su formalidad o informalidad, variables que no existen en el archivo. Por este motivo, a partir del análisis realizado y con la ayuda de los expertos, se decidió generar estas dos nuevas variables a partir de los datos del relevamiento.

En este sentido, se consideró necesario agregar información al conjunto de datos, a fin de determinar la vista minable sobre la que se trabajará. Se agregaron entonces las siguientes variables:

- Variable *SitLaboral*: aporta información sobre la situación laboral de una persona, que puede ser *ocupado, desocupado* o *inactivo*. Esta variable se utilizó como variable *clase* (objetivo).

  A partir de la generación de esta variable se pudo determinar el conjunto de datos de PEA (ocupados y desocupados).

- Variable *CarPEA*: se determinó a partir del análisis de las características del trabajo de las personas en relación con su formalidad (por ejemplo, si realizaban aportes jubilatorios) o informalidad (por ejemplo, si no hacían aportes pero realizaban changas o algún trabajo informal).

  Esta se utilizó como variable *meta* en Orange, es decir, una variable que no se usó en el estudio, pero que aporta información útil para caracterizar a esta población en relación con la formalidad o informalidad de su trabajo e interpretar los modelos obtenidos.

  Estos atributos derivados que se generaron a partir de uno o más atributos existentes en el conjunto de datos, son necesarios para la aplicación de las técnicas de modelado y han sido evaluadas con la ayuda de expertos en el dominio y en Relaciones Laborales.

3.4.2. Registros generados. No se generan nuevos registros en el conjunto de datos.

3.5. Integrar los datos:

3.5.1. Unificación de datos: El conjunto de datos final unificado, está formado por 703 ejemplos y 14 variables, cuyas distribuciones se muestran en el Anexo E.

3.6. Caracterizar el formato de los datos:

En esta etapa se realizó el cambio en el formato del conjunto de datos del relevamiento del Barrio Industrial (pero no en relación a su significado), a fin de adaptarlos a los requisitos de las técnicas de modelado elegidas.

3.6.1. Reporte de calidad de datos: Luego de realizar la limpieza en el conjunto de datos, se aplica el formato adecuado y unificado a los mismos para continuar con el análisis de datos. Se realizan las modificaciones de formato sobre los datos que son necesarias para adecuarlos a la herramienta de análisis a utilizar.

Fase4: Modelado

En esta fase de Modelado se eligieron las técnicas de modelado o algoritmos de minería de datos adecuados para ser aplicados sobre el conjunto de datos o vista minable que fueron preparados en la fase anterior.

Se determinaron además, los parámetros óptimos, los que se ajustan al modelo. Pueden existir distintas técnicas para un mismo problema de minería de datos y a su

vez, cada una de ellas tiene ciertos requisitos sobre los datos, motivo por el cual suele ser necesario volver a la fase de preparación de los datos.

Las tareas (generales) de la fase 'Modelado' de la metodología CRISP-DM (con sus salidas) son:

4.1. Seleccionar una técnica de modelado:

En esta instancia, se seleccionaron las técnicas de minería de datos a utilizar.

4.1.1. Técnica de modelado (salidas):

Se aplicaron los procedimientos específicos diseñados en relación a cada proceso de explotación de información determinado. Se documentó la técnica de modelado específica asociada a dicho proceso explotación de información:

Procedimiento1: Proceso de "Descubrimiento de grupos"

En este procedimiento se planteó la aplicación de un enfoque no supervisado.

Este procedimiento permitió realizar un análisis exploratorio de los datos sociodemográficos del relevamiento del Barrio Industrial. A partir de este análisis se realizaron descripciones que proporcionaron información específica y fidedigna en relación con el conjunto de datos estudiado.

Para la selección de las técnicas a utilizar, resulta importante la naturaleza de los datos. En el conjunto de datos del relevamiento, la mayoría de las variables son categóricas (solo dos variables son numéricas), por lo que se consideró adecuado aplicar la técnica no supervisada de Análisis de Correspondencia (AC), que provee Orange. Esta es una técnica similar a PCA, pero calcula la transformación lineal en datos discretos en lugar de continuos.

Procedimiento N° 2: Proceso de "Descubrimiento de reglas de comportamiento*"*

En este procedimiento, se planteó la aplicación de un enfoque supervisado. Se seleccionó la técnica de Árbol de Decisión, que permitió construir un modelo predictivo, para descubrir las reglas de comportamiento de las personas que pertenecen a cada clase, a fin de poder caracterizar a los grupos.

Procedimiento N° 3: Proceso de "Ponderación de interdependencia de atributos"

En este procedimiento, se planteó la aplicación de un enfoque supervisado. La técnica seleccionada fue Naive Bayes, que aprende un modelo bayesiano ingenuo a partir de los datos. Esta técnica permitió construir un modelo predictivo, que ayudó a identificar cuáles son las variables explicativas (o

independientes) con mayor incidencia sobre la clase (variable explicada), en el conjunto de datos sociodemográfico del Barrio Industrial.

4.1.2. Supuestos del modelo.

Los datos de la base de datos del relevamiento del Barrio Industrial están disponibles para ser analizados. Además, son representativos para el estudio que se va a realizar.

4.2. Generar el plan de pruebas.

El plan de pruebas se generó en los procedimientos 1 y 2 (supervisados).

4.3. Construir el modelo

En esta etapa, se aplicó la técnica elegida sobre el conjunto de datos del relevamiento sociodemográfico del Barrio Industrial para generar uno o más modelos. Seguidamente, se explica la generación de modelos para cada procedimiento.

Procedimiento1: Proceso de "Descubrimiento de grupos"

Mediante la técnica no supervisada de Análisis de Correspondencia, se buscó determinar *grupos* de personas con características similares en el conjunto de datos estructurado de población, sin un atributo objetivo (clase).

Luego de aplicar la técnica Análisis de Correspondencia, se obtiene el modelo que se muestra en la **Fig. 17:** Análisis de Correspondencia. En la imagen se puede ver, en cada uno de los cuadrantes, las agrupaciones realizadas respecto a los componentes 1 y 2.

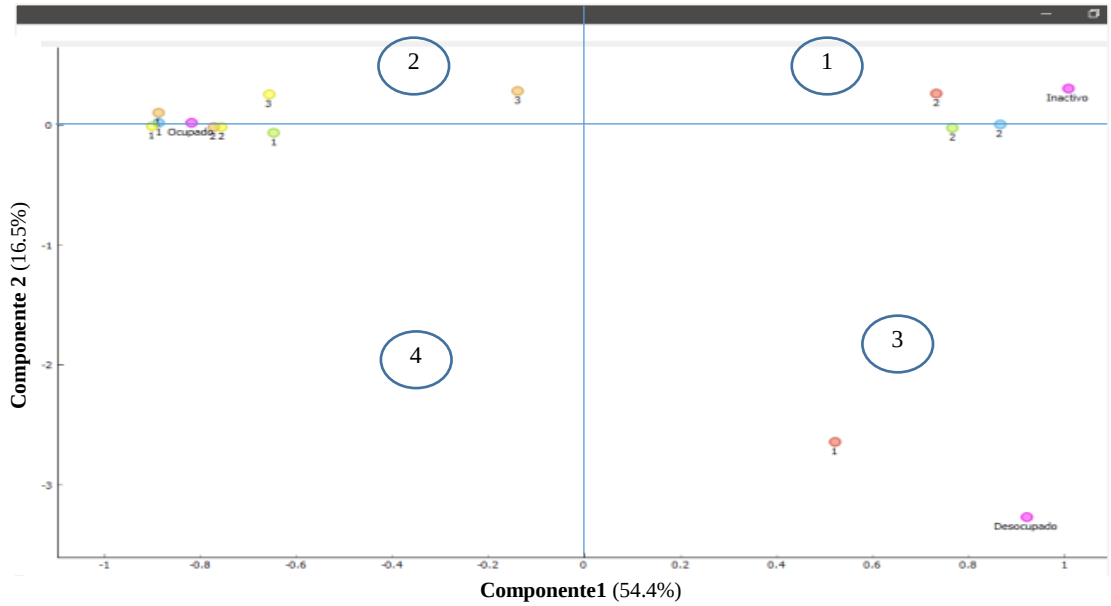


Fig. 17: Análisis de Correspondencia

En la Fig.17 se observa:

- La línea horizontal representa al *Componente 1*, que muestra hacia arriba y hacia abajo, si hay agrupación de objetos similares. El componente que explica mejor el modelo es el *Componente1*, cuyo valor de inercia es mayor (54.4%).

  En la **Fig. 18:** Componente 1 (Análisis de Correspondencia), se puede ver que la variable más relevante es *SitLaboral inactivo*, luego, la *SitLaboral desocupado* con menor relevancia.

Data Table (6) - Orange

File Edit View Window Help

	Variable	Value	Component 1
14	SitLaboral	Inactivo	1.00844
13	SitLaboral	Desocupado	0.921857
2	Trabajo1hr	2	0.866011
6	ChangaVta	2	0.767074
4	BTrab4Ultsem	2	0.73298
3	BTrab4Ultsem	1	0.521116
9	Dtojubilacion	3	-0.139748
5	ChangaVta	1	-0.647504
12	AportaIndep	3	-0.658042
11	AportaIndep	2	-0.757167
8	Dtojubilacion	2	-0.773565
15	SitLaboral	Ocupado	-0.820503
1	Trabajo1hr	1	-0.887452
7	Dtojubilacion	1	-0.889372
10	AportaIndep	1	-0.902126

Fig. 18: Componente 1
(Análisis de Correspondencia)

- La línea vertical representa al *Componente 2*, y muestra a la izquierda o derecha si hay alguna agrupación de objetos similares.

  En la **Fig. 19:** Componente 2 (Análisis de Correspondencia), se observa que para el *Componente 2*, la variable más relevante es la *SitLaboral inactivo*, luego la *SitLaboral ocupado* con menor relevancia.

Data Table (6) - Orange

File Edit View Window Help

	Variable	Value	Component 2
14	SitLaboral	Inactivo	0.304698
9	Dtojubilacion	3	0.279534
4	BTrab4Ultsem	2	0.25871
12	AportaIndep	3	0.253103
7	Dtojubilacion	1	0.0987197
15	SitLaboral	Ocupado	0.0172093
1	Trabajo1hr	1	0.0155493
2	Trabajo1hr	2	0.00431037
10	AportaIndep	1	-0.0112792
11	AportaIndep	2	-0.0186111
8	Dtojubilacion	2	-0.0211172
6	ChangaVta	2	-0.0232053
5	ChangaVta	1	-0.0696274
3	BTrab4Ultsem	1	-2.6398
13	SitLaboral	Desocupado	-3.26568

Fig. 19: Componente 2
(Análisis de Correspondencia)

Es decir, se determinaron 3 grupos: *inactivo, ocupado y desocupado.*

Si se analizan los componentes, respecto a las variables:

1)- En el primer cuadrante de la Figura 17, se identifica a las personas de situación laboral *inactivo* y se puede notar que las variables más relevantes que representan a este *grupo* o las variables que caracterizan a las personas que forman este grupo de *inactivos*, son:

- Personas que contestaron que *No trabajaron* una hora la semana anterior al relevamiento (Trabajo1hr=2)
- Personas que contestaron que *No hicieron changas* o *alguna actividad informal* (ChangaVta=2)
- Personas que contestaron que *No buscaron trabajo en las ultimas 4 semanas antes del relevamiento* (BTrab4Ultsem =2).

2)- En el segundo cuadrante se puede ver que se agrupan las personas con situación laboral *ocupada* y se puede notar que las variables más relevantes que representan a este grupo o las personas que forman este grupo, son:

- Personas que contestaron que *Trabajaron una hora la semana anterior al relevamiento* (Trabajo1hr=1)
- Además, estan muy relacionadas con este grupo las personas que contestaron que:
  - ✓ *Le hacen descuentos jubilatorios en su trabajo* (DtoJubilacion=1)
  - ✓ *Aportan por si mismos para su jubilación* (AportaIndep=1), es decir que son *trabajadores formales*.
- Además se relacionan con este grupo las personas que contestaron que:
  - ✓ *Hacen changas o alguna actividad informal* (ChangaVta=1)
  - ✓ *No hacen aportes jubilatorios* (AportaIndep=2), es decir que son *trabajadores informales*.

3)- En el tercer cuadrante se puede ver que se agrupan las personas con situación laboral *desocupada* y se puede notar que las variables más relevantes que representan a este grupo de personas son:

- Personas que contestaron que *no trabajaron una hora la semana anterior al relevamiento* (Trabajo1hr=2)
- Pero, estas personas contestaron que *sí buscaron trabajo en las ultimas 4 semanas antes del relevamiento* (BTrab4Ultsem =1).

Es decir, la situación laboral *desocupado* está muy relacionada con la variable *BTrab4Ultsem =1* (*que indica que una persona sí buscó trabajo en un periodo determinado, específicamente en las cuatro últimas semanas antes del relevamiento*), lo cual coincide con las definiciones dadas sobre desocupados [2].

La aplicación de este procedimiento permitió además, realizar un análisis exploratorio de los datos a fin de entenderlos. Este análisis exploratorio incluyó

distintas visualizaciones que provee la herramienta Orange. El análisis de estas visualizaciones permitió determinar o entender los grupos y la PEA.

Si se realiza una segmentación respecto a la variable *SitLaboral* (situación laboral), y se analiza la variable *CarPEA (desocupado, estudiante, jubilado, inactivo, trabformal y trabinformal)*, se puede observar, en la **Fig. 20:** Segmentacion de datos (respecto a la variable "SitLaboral"), que en el grupo cuya situación laboral es *ocupado*, la variable *CarPEA* toma dos valores (*TrabFormal y TrabInformal*). Si bien hay muchas personas *ocupadas*, la mayoría de estas personas tienen trabajos informales.

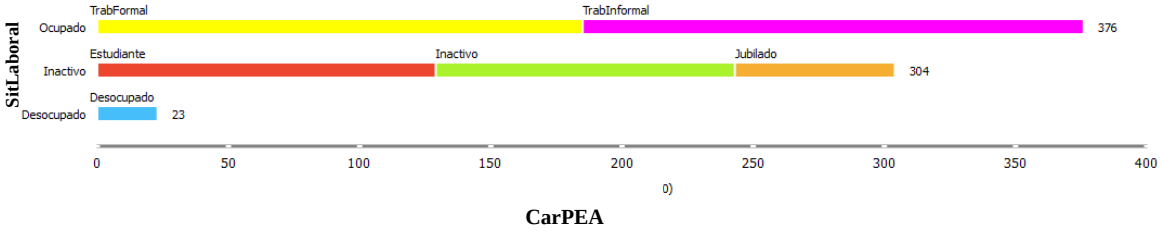


Fig. 20: Segmentacion de datos (respecto a la variable "SitLaboral")

Además, se observa que el grupo de *inactivos* tiene 304 ejemplos y esta formado por *estudiantes, inactivos* (amas de casa, por ejemplo) *y jubilados*. Este grupo de personas no forma parte de la PEA.

El grupo de *desocupados* tiene solo 23 ejemplos. Este grupo Forma parte de la PEA junto con el grupo de *ocupados*.

Otra de las visualizaciones que provee el Software Orange para el tratamiento de datos categóricos, es la *Matriz* o *Diagrama de Mosaico***,** que es la representación gráfica de una tabla de frecuencias (o tabla de contingencia), que se utiliza para visualizar datos de dos o más variables categóricas, a fin de identificar la relación entre las mismas.

Procedimiento 2: Proceso "Descubrimiento de reglas de comportamiento"

Generar plan de prueba: Al construir los modelos, se necesita un mecanismo para determinar su calidad y validez. En esta fase se divide el conjunto de datos en un grupo para entrenar el modelo (training) y otro para probarlo (test).

A partir del análisis realizado en el *Procedimiento 1*, se considera como clase, en el conjunto de datos del Barrio Industrial las que representan los grupos detectados, es decir, la variable situación laboral (SitLaboral).

En los enfoques supervisados, se entrena el modelo utilizando el conjunto de datos de *entrenamiento* y se valida el modelo utilizando el conjunto de datos de prueba.

El reporte de la tarea realizada durante esta subfase es:

Plan de pruebas: Se determina y documenta de qué forma se entrenarán y evaluarán los modelos generados. Se realiza el muestreo de los datos como muestra la **Fig. 21:** Muestreo de datos.

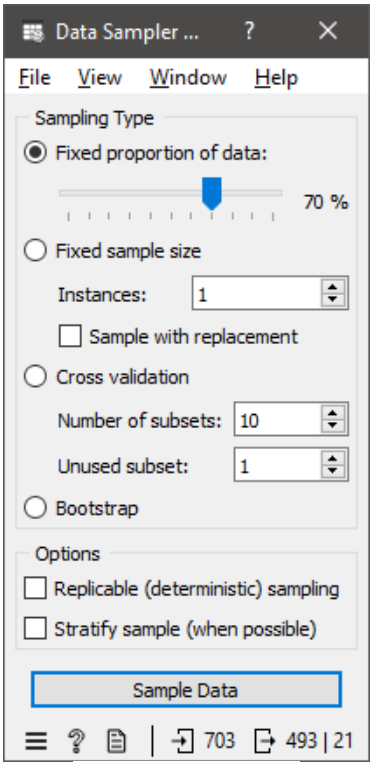


Fig. 21: Muestreo de datos

El conjunto de datos del relevamiento del Barrio Industrial se divide en:

- Conjunto de datos de entrenamiento: 70% registros.
- Conjunto de datos de prueba: 30%.

Se evaluó la capacidad del modelo en función de su matriz de confusión.

Construir el modelo: Se trabajó con algoritmos de clasificación (árboles de decisión) para construir un modelo predictivo que permita estimar, en base a los datos disponibles, las características de las personas que pertenecen a cada grupo (situación laboral).

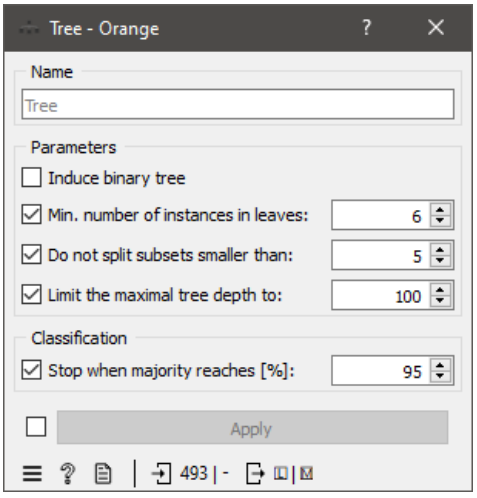


Fig. 22: Parámetros del árbol de decisión

Configuración de parámetros: en la **Fig. 22**: Parámetros del árbol de decisión, se muestran los parámetros que se proporcionan al modelo**:**

Mín. número de instancias en hojas: (6)

- **No dividir subconjuntos menores que:** (5). Prohíbe que el algoritmo divida los nodos con menos del número de instancias dado.
- **Limitar la profundidad máxima del árbol:** (100) limita la profundidad del árbol de clasificación al número especificado de niveles de nodos.
- **Detener cuando la mayoría alcance [%]:** (95). Indica que se deje de dividir los nodos después de que se alcance el umbral indicado.

Modelo: Describir los modelos reales generados por la herramienta de minería.

Como se puede ver en la **Fig. 23:** Árbol de Decisión, la variable más significativa para determinar la situación laboral de una persona es *Trabajo1hr*. A partir del árbol de decisión generado, se pueden determinar las características de las personas que conforman cada grupo. Es decir:

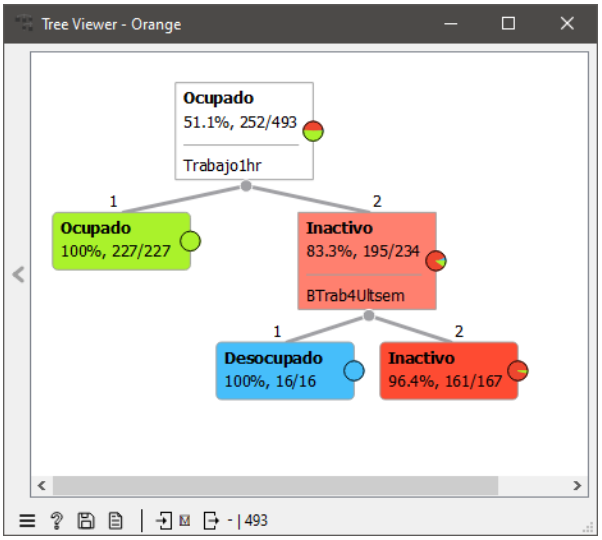


Fig. 23: Árbol de Decisión

- Si una persona contesta que trabajo una hora la semana anterior al relevamiento (Trabajan1hr=1), con seguridad su situación laboral es *"ocupado"*, con una probabilidad del 100% (para 227 casos).
- Si una persona contesta que no trabajo una hora la semana anterior al relevamiento (Trabajan1hr=2), entonces hay grandes probabilidades de que su situación laboral sea *"inactivo"* (195 casos de 234). Pero, si además la persona contesta:
  - ✓ Que buscó trabajo en las últimas 4 semanas antes del relevamiento, entonces con seguridad su situación Laboral será *"desocupado"* (16 casos).
  - ✓ Si contesta que no buscó trabajo, entonces hay grandes probabilidades de que su situación laboral sea *"inactivo"*, con una probabilidad del 96.4% (para 161 casos, sobre 167 totales).

Descripción del Modelo: Se describe el modelo resultante, mediante un informe que detalle la interpretación de los modelos y documente cualquier dificultad encontrada con su significado.

Como se puede ver en la **Fig. 24:** Predicciones del árbol de decisión, la precisión del modelo es del 93%. Si se ajusta un modelo predictivo usando una técnica de aprendizaje *supervisado*, se puede verificar qué tan bien predice el modelo la *respuesta "y"* (variable objetivo o explicada), en observaciones no utilizadas para el ajuste del modelo.

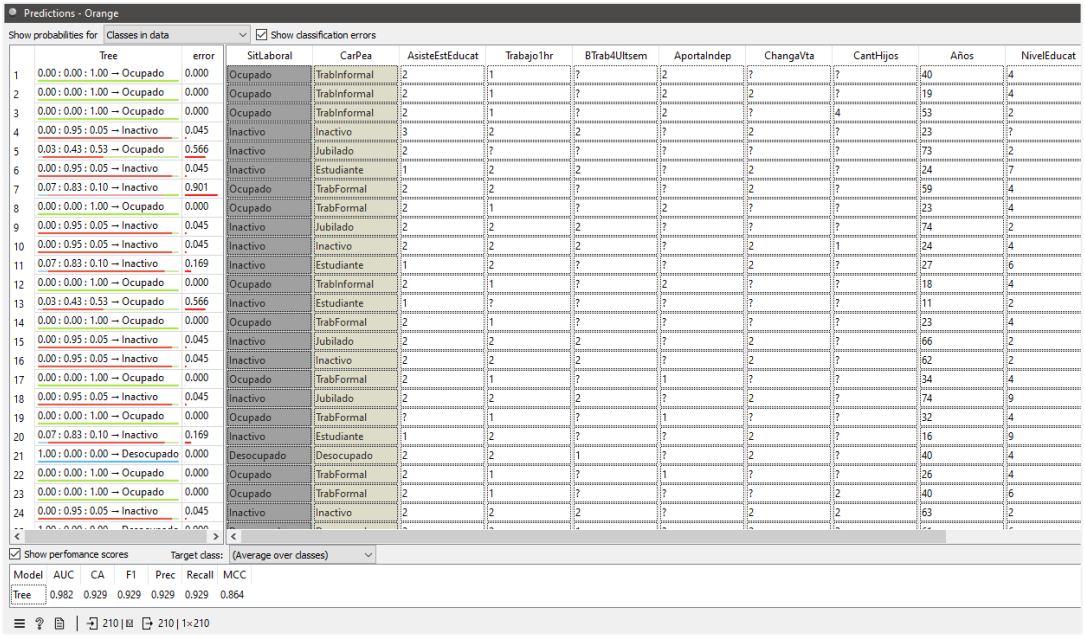


Fig. 24: Predicciones del árbol de decisión

Como en el conjunto de datos existen muchos valores ausentes, esto puede afectar los resultados obtenidos. Si bien, la presición es buena, se realiza la imputación de los valores faltantes.

Al imputar los valores, se reemplazan los valores faltantes con el valor promedio (para valores continuos) o más frecuente (para valores discretos).

Los resultados obtenidos con la imputación son los que se muestran en la **Fig. 25:** Predicciones del árbol de decisión (con imputación), la precisión mejora a un 96%. Esta precisión representa el porcentaje de acierto. A la izquierda de la imagen (Fig. 25), se visualizan las predicciones realizadas por el árbol de decisión. A la derecha de la imagen se muestran los valores reales, del conjunto de prueba.

Las métricas que representan el desempeño del modelo, se visualizan en la parte inferior izquierda de la Fig. 25.

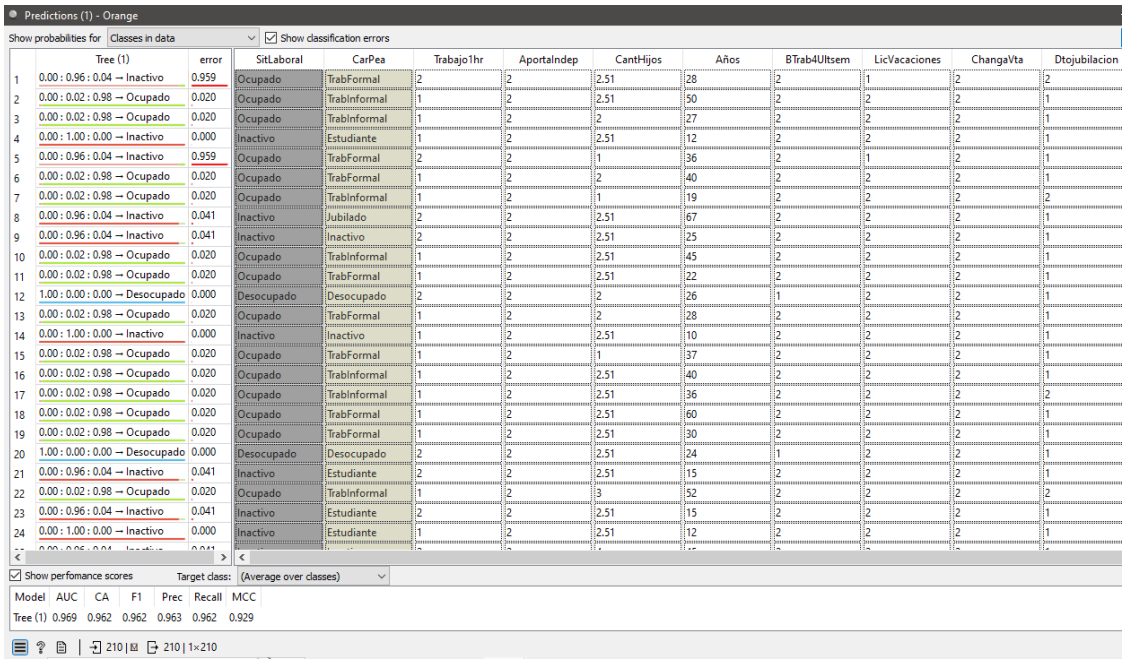


Fig. 25: Predicciones del árbol de decisión (con imputación)

- **CA (Classification accuracy)= 96%.** Indica la precisión de la clasificación.
- **Prec (Precision)=96%.** Representa la proporción de verdaderos positivos entre los casos clasificados como positivos.
- **Recall= 96%.** Es la proporción de verdaderos positivos entre todos los casos positivos en los datos.

Para medir el *desempeño*, se analiza la matriz de confusión que se muestra en la **Fig. 26:** Matriz de confusión (árbol de decisión). Esta matriz proporciona el número/proporción de instancias entre la clase prevista y la real.

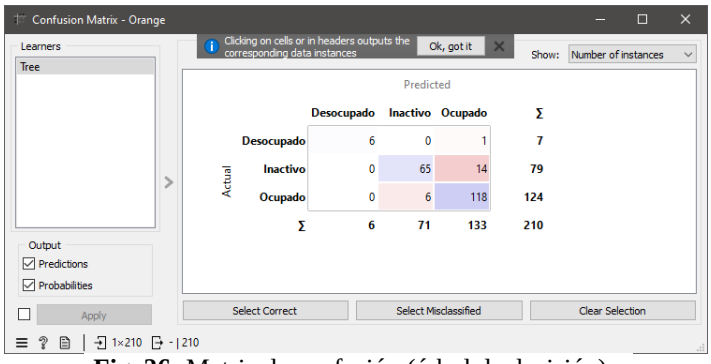


Fig. 26: Matriz de confusión (árbol de decisión)

En las filas de la matriz se pueden ver los valores *observados* y en las columnas los valores *predichos*. En la diagonal de la matriz se muestran los *aciertos*. Por fuera de la diagonal se muestran los errores:

- Existen 6 instancias en las que se predijo como “inactivo” cuando correspondia “ocupado” (en la segunda columna).
- En una instancia predice como “ocupado” cuando era “desocupado” (en la tercera columna).

- En 14 casos predice como “ocupado” cuando son “inactivos” (en la tercera columna).

Es importante tener en cuenta que en el conjunto de datos, las clases no se encuentran *balanceadas*. Es decir, la situación laboral “ocupado” es la clase mayoritaria y “desocupados” es la clase minoritaria.

En general el modelo clasifica la mayor parte de las instancias dentro de la clase mayoritaria.

Procedimiento 3: Proceso de “Ponderación de interdependencia entre atributos”

Generar plan de prueba: Se aplica un enfoque supervisado sobre un conjunto de datos estructurado, donde se conoce la clase (*SitLaboral***)**. Se divide el conjunto de datos y se entrena el modelo utilizando el conjunto de *entrenamiento* y se valida el modelo utilizando el conjunto de datos de *prueba*. El reporte de la tarea realizada durante esta subfase es:

Plan de pruebas: Se divide el conjunto de datos del relevamiento del Barrio Industrial en:

- Conjunto de datos de entrenamiento: 70% registros.
- Conjunto de datos de prueba: 30%.

Construir el modelo: En este procedimiento se trabajó con el algoritmo Naive Bayes, para construir el modelo que permitió determinar la incidencia de las variables explicativas (o independientes) sobre la variable explicada o clase (situación laboral).

Descripción del modelo: Se describe el modelo resultante, mediante un *Nomograma* que se muestra en la **Fig. 27:** Nomograma de Naive Bayes, clase "desocupado".

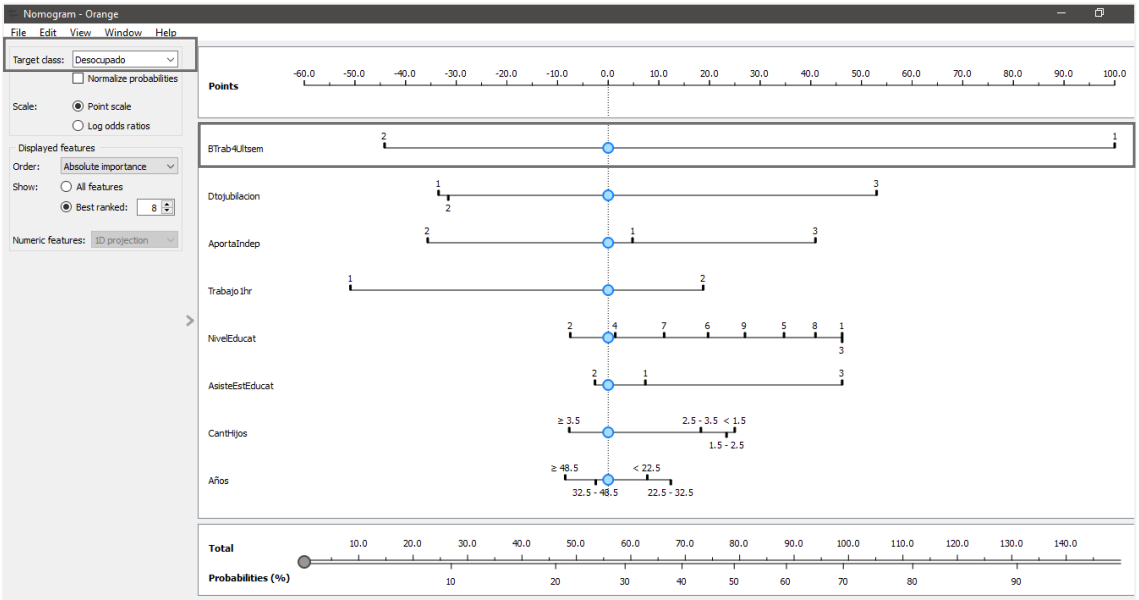


Fig. 27: Nomograma de Naive Bayes, clase "desocupado"

Un widget de *Nomograma* permite visualizar el modelo de Naive Bayes en el que se pueden ver los 8 (ocho) atributos mejor clasificados y cuánto contribuyen a la clase objetivo. Además, en el nomograma, se puede verificar cómo el cambio de los valores de los atributos afecta las probabilidades de clase mediante un widget interactivo. El primer cuadro colocado en la figura, se muestra la clase objetivo "*desocupado*". El segundo cuadro muestra el atributo más importante, la variable *BTrab4Ultsem* y su contribución a la probabilidad de la clase *situación laboral* (*SitLaboral*).

El último cuadro muestra la probabilidad total de la clase objetivo para los valores de atributos seleccionados (puntos azules), lo cual, en el caso de la variable *BTrab4Ultsem* es bastante alto. Lo mismo ocurre con los demás atributos; en la figura se puede visualizar cuánto contribuye un determinado valor a la probabilidad de una clase seleccionada.

Como se observa, la variable *BTrab4Ultsem* es la que mejor clasifica a la clase "desocupado", es un atributo altamente influyente. Le siguen en importancia las variables *Dtojubilacion, AportaIndep, Trabajo1hr*.

La variable menos influyente es *Años*, seguida de *CantHijos.*

Como se puede observar en la **Fig. 28:** Nomograma de Naive Bayes, clase "ocupado", la variable Trabajo1hr es la que mejor clasifica a la clase "*ocupado"*.

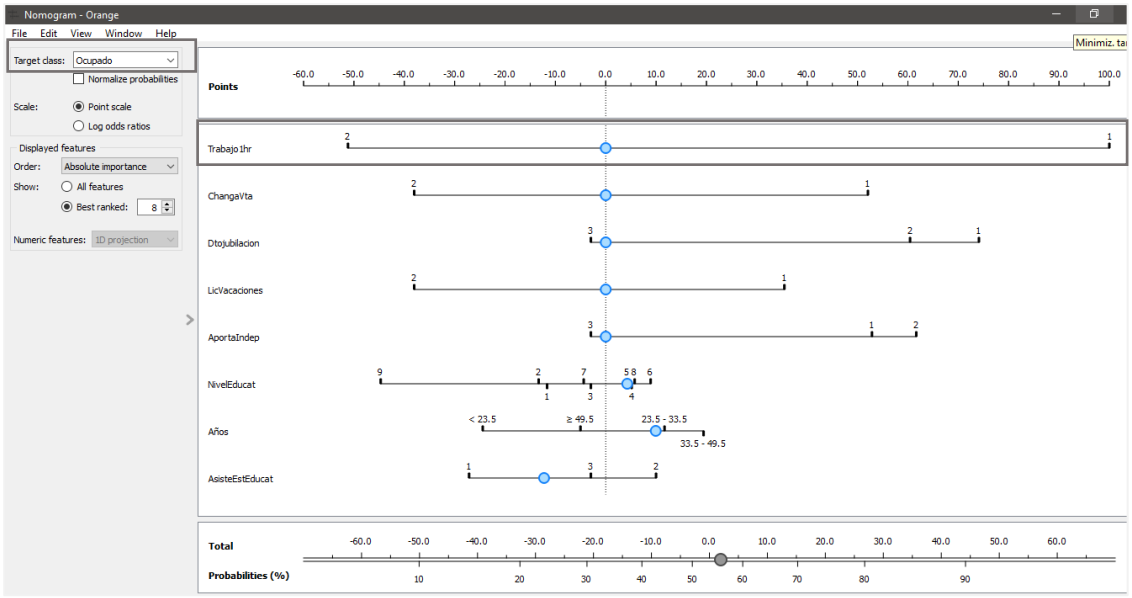


Fig. 28: Nomograma de Naive Bayes, clase "ocupado"

Es decir que, *Trabajo1hr* es un atributo altamente influyente; luego, le siguen en importancia las variables *ChangaVta. Dtojubilacion, LicVacaciones,AportaIndep.*

La variable menos influyente es la variable *AsisteEstEducat (que indica si una persona asiste o no a un establecimiento educativo)*, seguida de la variable *Años (edad de una persona).*

El primer cuadro colocado en la figura, muestra la clase objetivo "*ocupado*".

El segundo cuadro muestra el atributo más importante, la variable *Trabajo1hr* y su contribución a la probabilidad de la clase *situación laboral* (*SitLaboral*).

A partir del *entrenamiento* realizado, se evaluó el desempeño del algoritmo *Naive Bayes*.

Se realizan las predicciones en el software Orange, mediante Naive Bayes, lo que permite analizar los resultados obtenidos por el algoritmo y las métricas de desempeño del mismo.

Como se puede ver en la **Fig. 29:** Predicciones de Naive Bayes, la precisión del modelo es del 97%. Esta precisión representa el porcentaje de acierto.

A la izquierda de la imagen (Fig. 29), se visualizan las predicciones realizadas por Naive Bayes. A la derecha de la imagen se muestran los valores reales, del conjunto de prueba.

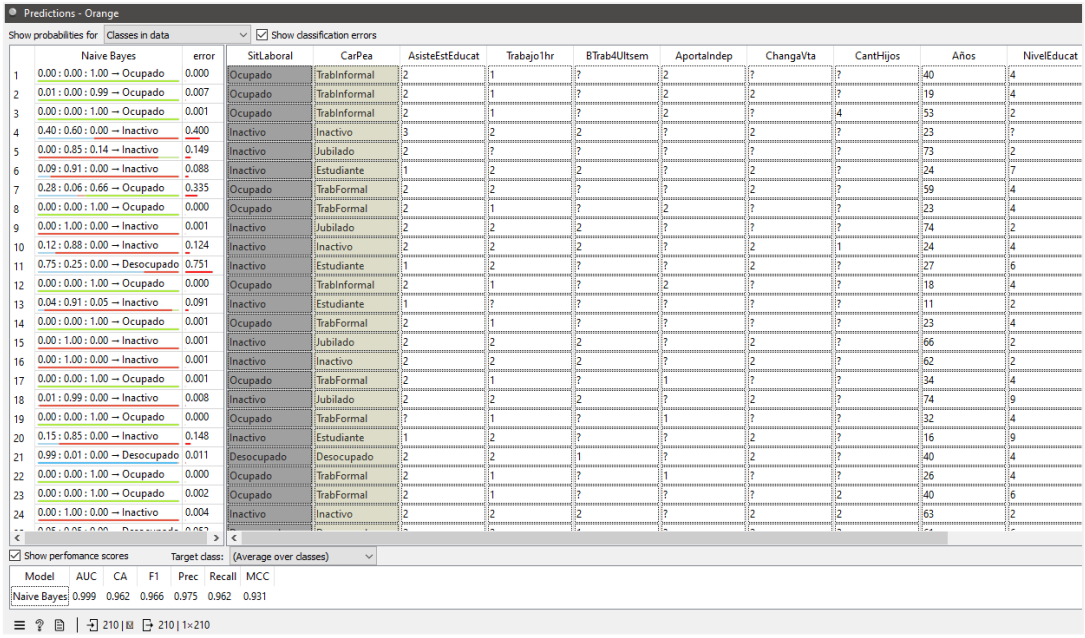


Fig. 29: Predicciones de Naive Bayes

Las métricas que representan el desempeño del modelo, se visualizan en la parte inferior izquierda de la Fig. 29.

- **CA (Classification accuracy)= 96%.** Indica la precisión de la clasificación; es la proporción de ejemplos clasificados correctamente.
- **Prec (Precision)=97%.** Representa la proporción de verdaderos positivos entre los casos clasificados como positivos (Por ejemplo, la proporción de "*ocupados*" identificados correctamente como "*ocupados*").
- **Recall= 96%.** Es la proporción de verdaderos positivos entre todos los casos positivos en los datos (por ejemplo, el número de "*desocupados*" entre todos los diagnosticados como "*desocupados*").

Se evalua la capacidad del modelo en función de su Matriz de confusión que se observa en la **Fig. 30:** Matriz de confusión (Naive Bayes).

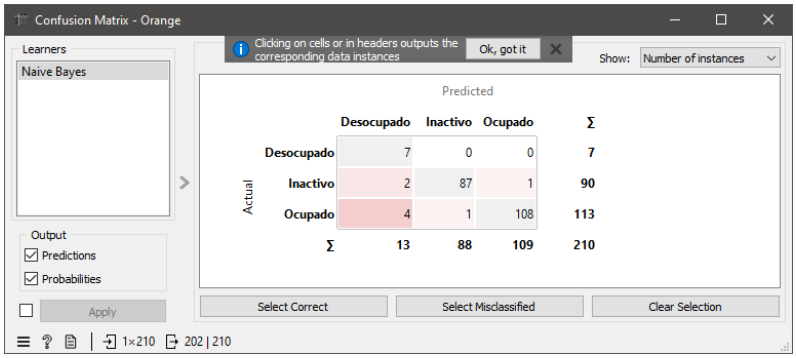


Fig. 30: Matriz de confusión (Naive Bayes)

En la diagonal de la matriz se muestran los *aciertos*. Por fuera de al diagonal se muestran los errores:

- Existen 2 instancias en las que se predijo como "*desocupado*"**,** cuando correspondia "*inactivo"* (primer columna).
- Existen 4 instancias en las que predice como *"desocupado"***,** cuando es *"ocupado"* (primer columna).
- Existe 1 instancia en las que predice como *"inactivo"* y es *"ocupado"***.**
- Existe 1 instancia en las que predice como *"ocupado"* y es *"inactivo"***.**

4.4. Evaluar el Modelo

En esta fase, se interpretó y evaluó el modelo en función del conocimiento del dominio con la ayuda de los expertos, los criterios de éxito definidos para el proyecto, a fin de evaluar el éxito de la aplicación del modelo. La técnica en general se aplica más de una vez o se utilizan técnicas alternativas a fin de generar resultados.

4.4.1. Evaluación de los modelos. Generar un reporte de evaluación de los modelos obtenidos, describiendo sus características.

A partir de la **Fig. 31:** Comparativa. Predicciones Naive Bayes y Árbol de Decisión, se pueden comparar las predicciones realizadas por ambos modelos.

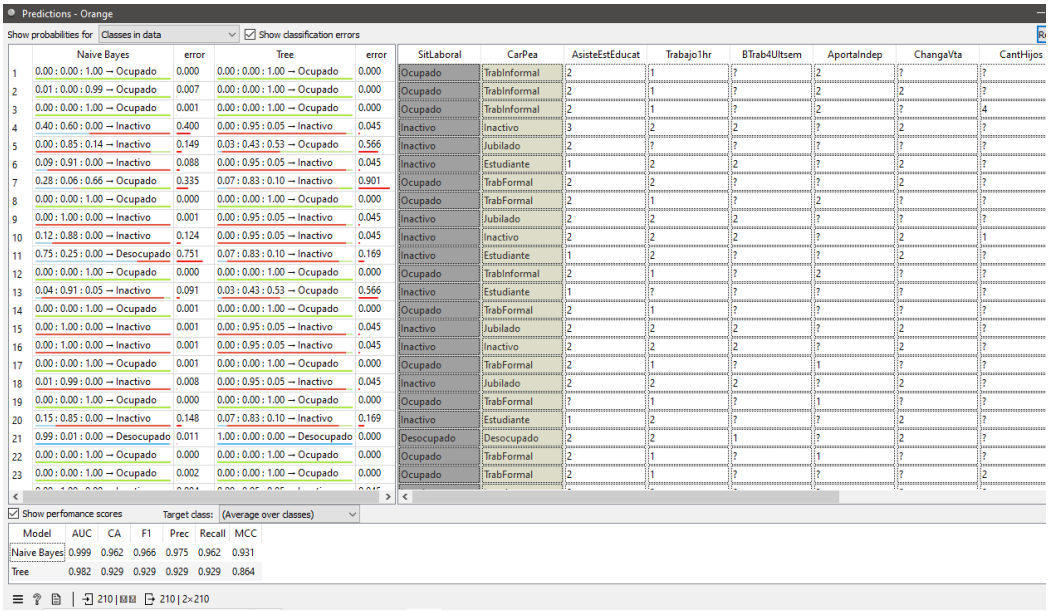


Fig. 31: Comparativa. Predicciones Naive Bayes y Árbol de Decisión

Si se analizan las métricas que representan el desempeño de los modelos, se puede notar que en relación al porcentaje de acierto, el algoritmo Naive Bayes es mejor. Como se puede ver la precisión del modelo es del 97% y la obtenida con el Árbol de Decisión es del 93%.

De igual manera, en relación a la precisión de la clasificación (accuracy), Naive Bayes arroja una mejor precisión del 96% y el Árbol de Decisión del 93%.

4.4.2.Revisión de la configuración de parámetros. En relación a la evaluación anterior, se pueden revisar los parámetros, se ajustan los mismos y se vuelve a la fase de construcción del modelo, en caso de que los modelos obtenidos no presenten un buen desempeño en relación a los criterios de éxito establecidos. Se repiten las etapas de construcción y evaluación del modelo, hasta encontrar los *mejores* modelos.

Se considera que los modelos obtenidos presentan un buen el desempeño en relación a las métricas analizadas, por lo que no se realiza la revisión de parámetros.

Fase5: Evaluación

En esta fase se evalúan los modelos y a fin de determinar si cumplen con los criterios de éxito del proyecto identificados en la primera fase, para asegurar el logro de los objetivos de negocio.

Esta fase concluye al decidir como se usarán los resultados, es decir**,** se buscan los modelos que contribuyan a una mayor calidad de análisis y que sean útiles a las necesidades de la investigación.

Se evalua cada modelo obtenido con el experto, se determina la precisión del mismo y se lo interpreta en el dominio del problema, es decir, que significan los resultados obtenidos en relación a la PEA del Barrio Industrial.

Las Tareas generales de la fase 'Evaluación' de la metodología CRISP-DM (con sus salidas) son:

5.1.Evaluar los resultados

En esta instancia se evaluaron los resultados obtenidos de la explotación de información con respecto a los criterios de éxito de este TFM.

5.1.1.Evaluar los resultados de la minería de datos:

Los resultados obtenidos de la minería de datos se evaluaron respecto a los criterios de éxito. Es decir, se evalúan los modelos para determinar su importancia en relación a los objetivos de esta investigación.

Los resultados obtenidos con la aplicación de este procedimiento de explotación de la información, son analizados y valorados por expertos en el dominio. Se considera que el procedimiento generado permite la ejecución ordenada de las tareas en un proyecto de explotación de la información. Es una guía en la realización de tareas y permite asegurar la obtención de resultados de calidad en el proyecto, aplicado un estudio de la población.

A partir de la aplicación del procedimiento de explotación de la información, se obtuvo conocimiento de los datos, se identificaron los grupos y se caracterizó a los mismos mediante la determinación de reglas de comportamiento de las personas que conforman esos grupos. Asimismo, se pudo determinar la incidencia de cada variable o los factores de incidencia en la variable clase, es decir, que variables son mas significativas para la determinación de la clase o situación laboral.

La aplicación de los procedimientos permitió obtener conocimiento de la PEA del Barrio Industrial; se pudo determinar la PEA mediante el análisis de datos de personas ocupadas y desocupadas. Asimismo, se pudo determinar que en el conjunto de datos existe una gran cantidad de personas cuya situación laboral es *Inactiva*.

En cuanto a la evaluación de los resultados obtenidos de la minería de datos en relación a los criterios de éxito, se puede decir que:

Criterio de éxito para el objetivo de minería de datos N° 1:

Se considera que se cumplió con este primer criterio, debido a que con la aplicación del **Procedimiento N° 1** se logró identificar *grupos* en el conjunto de datos del relevamiento sociodemográfico del Barrio Industrial. Asimismo mediante el análisis exploratorio se pudo entender a estos grupos.

Criterio de éxito para el objetivo de minería de datos N° 2:

Se considera que se cumplió con este segundo criterio, debido a que con la aplicación del **Procedimiento N° 2** se obtuvo un modelo de árbol de decisión cuya capacidad predictiva es aceptable, debido a que se estableció como criterio de éxito una tasa de acierto mayor al 70% y se obtuvo una tasa de acierto del 93%, por lo que se considera que se cumplió con este objetivo. Asimismo, se pudo determinar con el árbol de decisión obtenido, cuales son las características que definen a cada clase, como se explicó en el desarrollo de la validación.

Criterio de éxito para el objetivo de minería de datos N° 3:

Se considera que se cumplió con este tercer criterio, debido a que con la aplicación del **Procedimiento N° 3** se obtuvo un modelo de *Naive Bayes* cuya capacidad predictiva es aceptable, debido a que se establecio como criterio de éxito una tasa de acierto mayor al 70% y se obtuvo una tasa de acierto del 97%, por lo que se considera que se cumplió con este objetivo. Asimismo, se pudieron visualizar las predicciones realizadas por este algoritmo, además, mediante el componente nomograma, se pudo establecer para cada clase de la PEA (ocupado y desocupado), cuales son las variables más significativas que influyen en la situación laboral (o que la determinan).

Se determinó que el factor (variable explicativa) que tiene mayor incidencia sobre la determinación de la clase *"desocupado"*, en el Barrio Industrial, es la variable *BTrab4Ultsem* (es decir, que una persona busque trabajo *BTrab4Ultsem*=1), luego, le siguen en importancia las variables *Dtojubilacion, AportaIndep, Trabajo1hr* (es decir que la persona no haga *aportes jubilatorios* y tampoco esté ocupada), La variable *BTrab4Ultsem* es que mejor clasifica la clase *"desocupado"*. La variable menos influyente para la determinación de esta clase, es la variable *Años.*

Se determinó además que el factor (variable explicativa) que tiene mayor incidencia sobre la determinación de la clase *"ocupado"*, en el Barrio Industrial, es la variable *Trabajo1hr* (es decir, si una persona trabajó una hora la semana anterior al relevamiento *Trabajo1hr* =1), luego, le siguen en importancia las variables *ChangaVta. Dtojubilacion, LicVacaciones,AportaIndep* (es decir si la persona hace changas o trabajos informales, o tiene un trabajo formal, en el que aporta para la jubilación o esta de vacaciones, o aporta para su jubilación en forma independiente). La variable *Trabajo1hr* es que mejor clasifica la clase *"ocupado"*.

La variable menos influyente es la variable *AsisteEstEducat (que indica si una persona asiste o no a un establecimiento educativo)*, seguida de la variable *Años.*

5.1.2.Modelos evaluados y aprobados.

A partir de la comparación de los modelos obtenidos con Árbol de Decisión y Naive Bayes respecto a sus métricas de desempeño, se determinó que Naive Bayes arroja mejores resultados o presenta un mejor desempeño con el conjunto de datos del relevamiento del Barrio Industrial, de la ciudad de Corrientes.

5.2. Revisión del proceso

Se revisó el proceso efectuado a fin de verificar posibles errores

5.2.1. Revisión del proceso (salida):

Se revisó el proceso y se documentaron todas las actividades realizadas. Los resultados obtenidos fueron registrados y utilizados para la elaboración de este informe final.

5.3. Determinar las próximas etapas o pasos

A partir de los resultados obtenidos evaluado con expertos, se revisa el proceso realizando los ajustes que se consideren necesarios y se presentan los resultados a las partes interesadas.

Fase6: Implementación

Esta fase se suele llamar también *despliegue* y se refiere a explotar los beneficios de los modelos. Esta etapa puede reducirse a la documentación y presentación de los resultados del proyecto de explotación de la información a las partes interesadas (cliente).

En este TFM esta etapa consistió en realizar la documentación y la generación de este informe con todos los resultados obtenidos en el proyecto de explotación de información, además de la presentación de los resultados a las partes interesadas, a fin de que el conocimiento obtenido, pueda ser utilizado.

En este TFM se revisará el procedimiento de explotación de la información en forma continua, según las necesidades de las partes interesadas.

En este informe final se incluye la documentación de la experiencia adquirida durante el desarrollo del proyecto y los resultados obtenidos en la validación de los procedimientos, con el análisis de los resultados obtenidos en el proyecto de explotación de la información que será presentado a las partes interesadas (cliente) con resultados y conclusiones.

4.7. Resultados obtenidos con el procedimiento de explotación de la información

La experiencia de trabajo de este TFM, consistió en diseñar un procedimiento de explotación de la información, en el que se realiza una adaptación de la metodología o modelo de proceso CRISP-DM, a fin de ordenar la ejecución de proyectos de explotación de la información.

Mediante la aplicación de distintos algoritmos al caso de estudio, se pudo evaluar la viabilidad de los modelos y validar los mismos. Los resultados obtenidos en dicha validación se exponen en este capítulo.

A partir de los resultados obtenidos en este TFM se logró comprobar que un algoritmo funciona correctamente con un determinado set de datos y para determinado tipo de problema. En el estudio realizado, contar con una base de datos con una gran cantidad de datos categóricos, hizo que para la agrupación, por ejemplo, no se usaran algoritmos de *clustering* basados en distancias, sino que se recurriera a otras técnicas no supervisadas para el tratamiento de datos categóricos.

Se puede decir entonces, que el éxito en la aplicación del procedimiento diseñado o en cada proyecto de explotación de la información realizado, va a estar dado por la experimentación realizada en cada caso de estudio y va a estar relacionado con el problema a resolver y el conjunto de datos usado.

La aplicación del procedimiento principal de explotación de la información responde al objetivo principal de este TFM que consiste en diseñar un procedimiento que permita sistematizar el trabajo a realizar en un proyecto de explotación de la información, mediante una metodología, a fin de lograr resultados de calidad.

En el desarrollo del proyecto de explotación de la información se aplican los procedimientos específicos, como solución a los problemas de minería de datos, que se resuelven mediante la aplicación de procesos de explotación de información. Estos procesos proponen la aplicación de técnicas de minería de datos, que permiten encontrar patrones significativos en la PEA estudiada u obtener conocimiento o información de calidad sobre los datos.

4.7.1. Resultado obtenido con la aplicación del Procedimiento 1:

En este Procedimiento, se propuso la aplicación del proceso "Descubrimiento de grupo" para dar solución al objetivo de minería de datos N° 1. Como se comentó en capítulos anteriores, este conjunto de datos contiene en su mayoría variables categóricas por lo que la aplicación de técnicas de agrupamiento basado en distancias no es lo apropiado en este caso. En este contexto, se optó por realizar el análisis mediante la técnica no supervisada Análisis de Correspondencia, que permitió determinar las agrupaciones en el conjunto de datos (como se mostró en la Fig. 17).

Además de lo expuesto en la validación del procedimiento, se puede realizar un análisis exploratorio del conjunto de datos, a partir de otras visualizaciones (presentadas en el Anexo E).

Los *graficos de mosaico,* disponibles en el software Orange, permiten realizar distintas visualizaciones, que muestran (en forma análoga a una *Tabla de Contingencias*) la relación entre variables categóricas. Por ejemplo, en la **Fig. 32:** Gráfico de Mosaico, se puede visualizar la relación entre las variables categóricas *sexo* y *carpea* (característica de la población en relación a su ocupación, formalidad o informalidad en su trabajo). Es decir, se observa la proporción de personas *ocupadas* (en color verde), en relación a las personas *inactivas* (en color rojo) y la cantidad de personas *desocupadas* (color celeste).

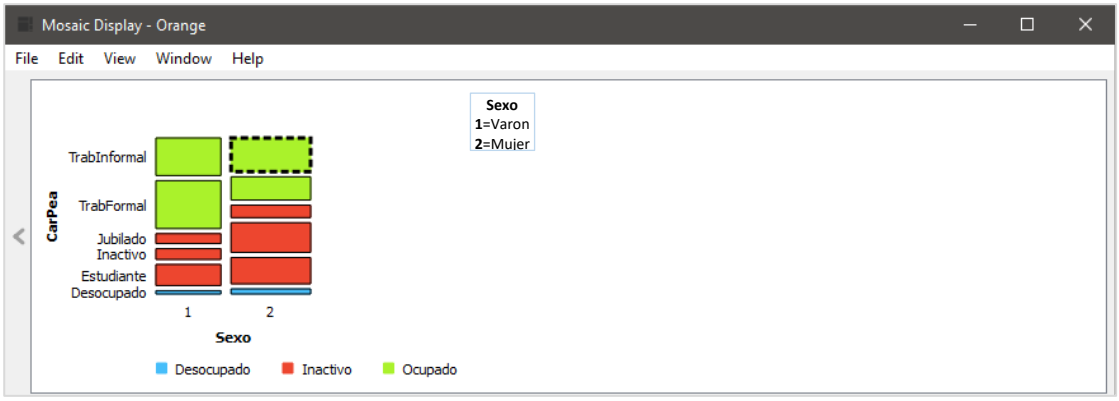


Fig. 32: Gráfico de Mosaico

Además, en la Fig. 32, se puede analizar la situación laboral por *sexo*. A la izquierda de la imagen se muestran los varones (sexo= "1") y a la derecha las mujeres (sexo= "2"). Se puede observar que en el grupo de varones hay mas personas *ocupadas*, y en este grupo de ocupados predomina el trabajo formal.

En cuanto al grupo de mujeres, existe gran cantidad de personas *inactivas* (entre ellos, amas de casa, estudiantes, personas jubiladas). En el grupo de mujeres que trabajan o están *ocupadas*, predomina el trabajo informal sobre el formal.

A partir del gráfico presentado en la Fig. 32, se puede realizar un análisis más detallado, por ejemplo, para obtener mas información relacionada al grupo de mujeres que tienen trabajos informales.

Para ello, en la **Fig. 33:** Trabajo informal (mujeres), se selecciona la porción del gráfico que corresponde al grupo de mujeres que tienen *trabajo informal* (seleccionado en la parte superior de la imagen, con línea de puntos):

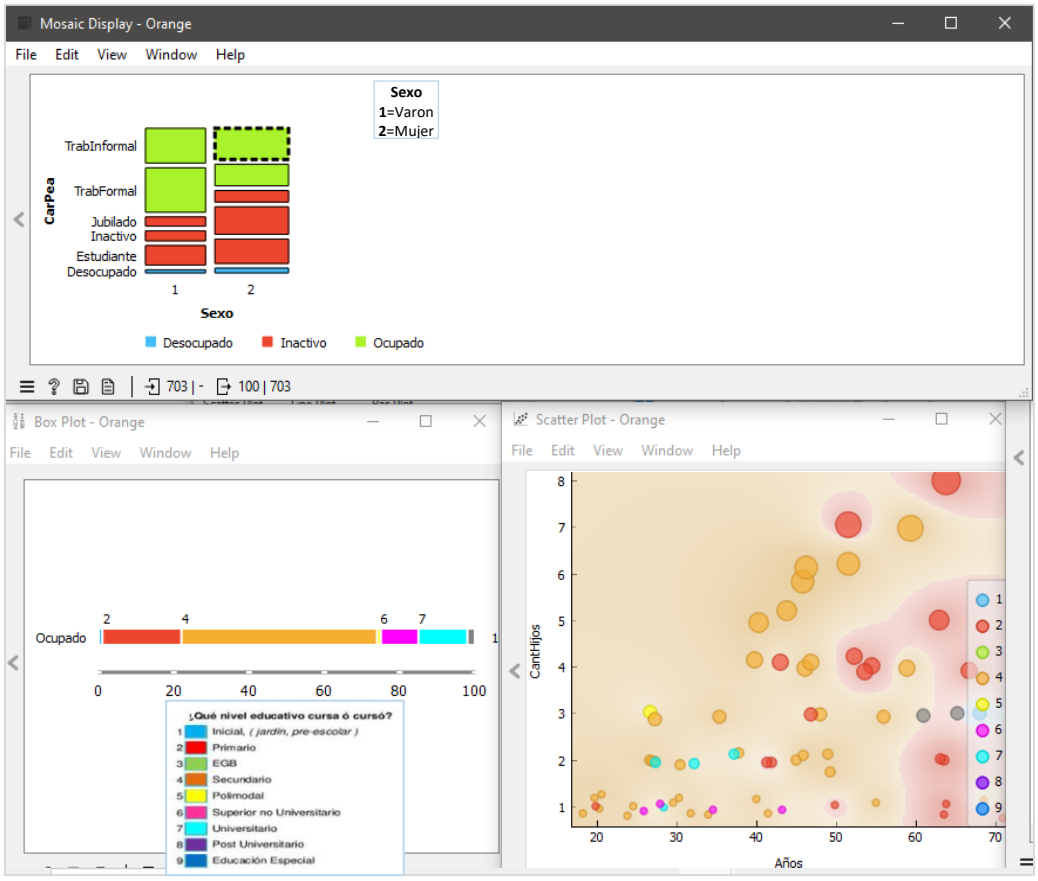


Fig. 33: Trabajo informal (mujeres)

A partir de la selección realizada, se obtienen otras visualizaciones en las que se observa:

- El *Nivel Educativo* que predomina es el *Secundario* (que corresponde aproximadamente al 52 % del total de mujeres de este grupo), le sigue en importancia el *Primario* (aproximadamente el 21 % del total de mujeres de este grupo). El 13 % tiene nivel educativo Universitario y el 10 % *Nivel Educativo Superior No Universitario.*
- Además, las mujeres de este grupo tienen edades entre 18 y 71 años.
- En cuanto a la cantidad de hijos, la mayoría tiene varios hijos (1-4).

En la **Fig. 34**: Trabajo formal (mujeres) se selecciona el grupo de Mujeres que tienen *trabajo formal* (seleccionado en el primer gráfico, con línea de puntos):

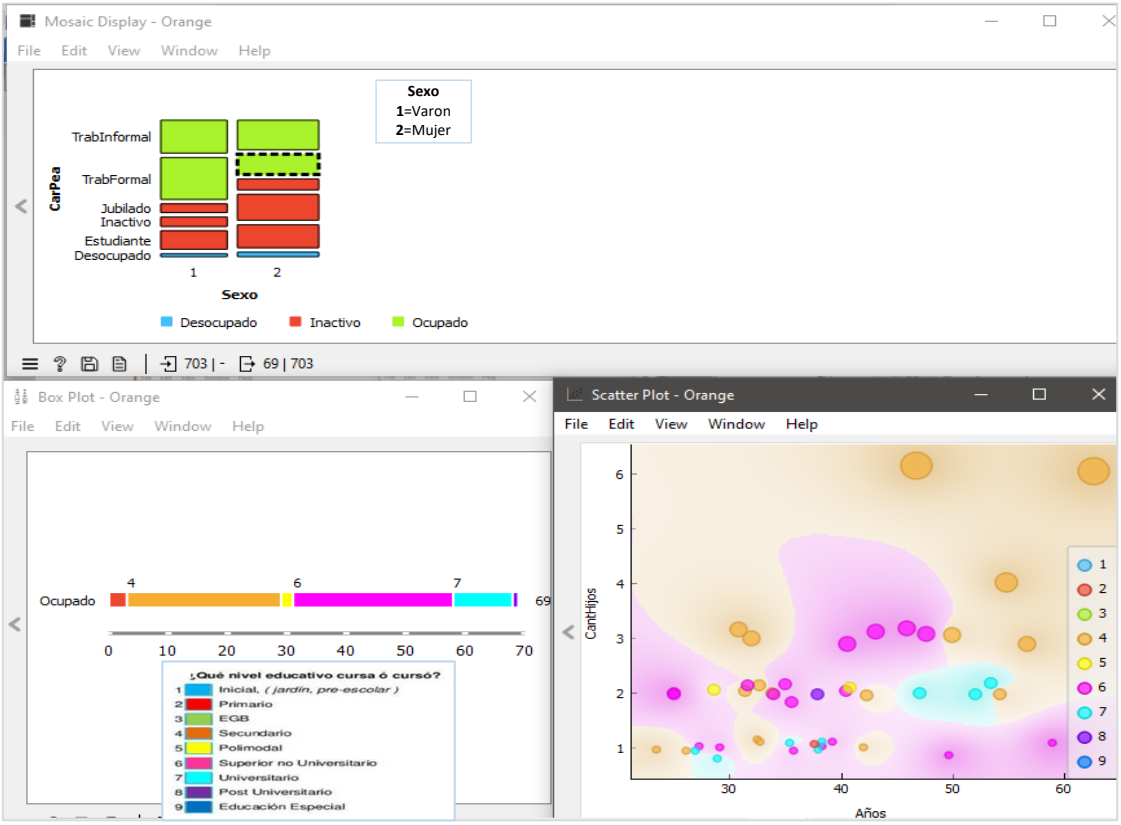


Fig. 34: Trabajo formal (mujeres)

Para el grupo seleccionado, se observa que:

El *Nivel Educativo* que predomina es el *Superior No Universitario* (que corresponde aproximadamente el 39% del total de mujeres de este grupo) y le sigue en importancia el *Secundario* (aproximadamente el 38% del total de mujeres de este grupo). El 14.5 % de las mujeres de este grupo tiene *Nivel Educativo Universitario* y el 4% un *Nivel Educativo Primario.*

- Además, estas mujeres tienen edades entre 21 hasta 62 años.
- En relación a la cantidad de hijos, la mayoría tiene pocos hijos (1-3).

En relación con los resultados obtenidos, se puede decir, en líneas generales, que en el caso de las mujeres, tener un trabajo formal requiere mayor capacitación. Esto, si se considera el bajo porcentaje de mujeres que tienen estudio *Primarios* y la mayor cantidad de mujeres con formación *Superior No Universitaria* en este grupo, respecto al grupo de mujeres con trabajo informales.

4.7.2. Resultado obtenido con la aplicación del procedimiento 2:

En este procedimiento, se propone la aplicación del proceso "Descubrimiento de reglas de comportamiento" para dar solución al objetivo de minería de datos N° 2.

A partir del conjunto de datos del relevamiento, en el que se conocen los grupos (clase), es decir, que se conoce la salida, aplicó el algoritmo árbol de decision a fin de obtener modelos que detecten las caracterizaciones de esos grupos (enfoque *supervisado).*

La precisión de árbol de decisión mostró un porcentaje de acierto de aproximadamente el 93%, se realizó una imputación sobre el conjunto de datos que permitió mejorar la precisión a un 96%.

El árbol de decisión generado muestra la importancia que tiene la variable *Trabajo1hr* (mostrado en la Fig. 23) para la determinación de las clases. Se puede notar en el mismo, la importancia de la variable *Trabajoh1hr* que es la característica principal de las personas que pertenecen al grupo de *"ocupados"* (*es decir si la persona trabajó una hora la semana anterior al relevamiento*).

Si una persona indica que no trabajo (Trabajoh1hr=2) pero si busco trabajo (BTrab4Ultsem=1) se la caracteriza como "Desocupada".

4.7.3. Resultado obtenido con la aplicación del Procedimiento 3:

En este Procedimiento, se propone la aplicación del Proceso "Ponderación de Interdependencia de Atributo (o descubrimiento de atributos significativos)", para dar solución al objetivo de minería de datos N° 3.

Este proceso de explotación de información permitió identificar los factores o variables que tienen mayor incidencia en el atributo clase (situación laboral). Es decir, permitió determinar o identificar los factores (características) que poseen mayor incidencia en la situación laboral de una persona (ocupada/desocupada). Es

decir, se trata de determinar las características distintivas de las personas respecto a la clase.

Resultado de Naive Bayes respecto a la clase *"desocupado":*

Se asoció a naive bayes un componente Nomograma, mediante el cual se pueden ver las variables más influyentes y el porcentaje de incidencia de la misma en la clase "desocupado".

Se determinó que el factor que más influye o mejor clasifica la clase "*desocupado"*, es la variable *BTrab4Ultsem* (como se mostró en la Fig. 27). Luego, le siguen en importancia las variables *Dtojubilacion, AportaIndep, Trabajo1hr.* El factor que menos incide en la clase desocupado, es la variable *Años.*

Por otra parte, se determinó que el factor que más influye o mejor clasifica la clase *"ocupado"*, es la variable *Trabajo1hr* (como se mostró en la Fig. 28). Luego, le siguen en importancia las variables *ChangaVta, Dtojubilacion, LicVacaciones, AportaIndep*. El factor que menos incide en que una persona pertenezca a la clase ocupado, es la variable *AsisteEstEducat (que indica si una persona asiste o no a un establecimiento educativo),* seguida de la variable *Años*.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Capítulo 5

Conclusiones y futuras líneas de investigación

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

5. Conclusiones y futuras líneas de investigación.

En este capítulo se exponen las conclusiones del trabajo realizado, en relación a los resultados obtenidos y los objetivos planteados en este TFM (sección 5.1), seguidamente se exponen los aportes y ámbitos de aplicación del trabajo realizado (sección 5.2) y se finaliza este apartado con las futuras líneas de investigación que surgen a partir del trabajo realizado en este TFM (sección 5.3).

5.1. Conclusiones

En los últimos años, los avances en el campo de explotación de información y la minería de datos han permitido modificar una forma de trabajo artesanal a una forma más ingenieril, al realizar los proyectos de explotación de información, mediante la aplicación de distintas metodologías. Sin embargo, la visión de las empresas se fue modificando a una realidad más dinámica, por lo que es necesario hacer frente a nuevas necesidades y requerimientos en este tipo de proyectos. En este contexto, la comunidad científica continúa trabajando para mejorar los modelos y metodologías utilizadas, a fin de optimizar los resultados obtenidos y lograr proyectos de explotación de la información exitosos.

Ante la escasez de estudios laborales en la región, en los que se utilicen herramientas de explotación de información, se considera que el procedimiento de explotación de la información diseñado es un aporte valioso en este campo y en proyectos de explotación de la información aplicado a datos estructurados, siendo estos los principales desafíos y contribuciones de este trabajo.

A continuación, se presenta un análisis de los objetivos específicos que guiaron la realización de este trabajo y los resultados obtenidos, que constituyen las aportaciones de este TFM.

En relación al primer objetivo específico planteado, se investigaron metodologías o modelos de proceso usados en explotación de información o minería de datos para datos estructurados. Se examinaron diferentes estudios (comentados el capítulo 2, Estado de la cuestión), en los que los autores explican distintos enfoques y realizan diversas contribuciones en este campo de estudio.

Luego de la investigación realizada, se considera a CRISP-DM como una metodología apta para aplicar a proyectos de explotación de la información para el análisis de datos estructurados de población, por lo que se seleccionó esta metodología para el diseño del procedimiento propuesto como solución al objetivo de este TFM. En este procedimiento se realiza una adaptación de CRISP-DM que

permite sistematizar las tareas a realizar en un proyecto de explotación de información.

En cuanto al segundo objetivo específico de este TFM y a partir de la revisión de trabajos realizados por distintos autores, se analizaron herramientas y técnicas de minería de datos que se pueden emplear en proyectos de explotación de información.

En la validación del procedimiento de explotación de información, realizada con datos reales, se utilizaron distintas técnicas de minería de datos que permitieron obtener conocimiento de los datos y patrones de comportamiento que ayudaron a detectar y analizar las problemáticas laborales y sus factores de incidencia en el conjunto de datos del relevamiento del Barrio Industrial. A partir de los resultados obtenidos, se considera que el procedimiento diseñado constituye una guía en el desarrollo de un proyecto de explotación de la información ya que posibilita la ejecución ordenada de las actividades y la obtención de las *salidas* en cada tarea, información que se utiliza en las fases sub-siguientes de la metodología.

Además, permite detectar errores y volver en el proceso para solucionarlos, por lo que contribuye a la evaluación y mejora de resultados obtenidos en proyectos de este tipo. Sin embargo, es importante tener en cuenta que el éxito de un proyecto de explotación de la información basado en experimentación siempre esta relacionado con el tipo de problema a resolver y el conjunto de datos utilizado, por lo que es necesario adaptar el procedimiento generado en relacion a los tipos de problemas que se traten en cada estudio y las características del conjunto de datos a explotar.

En relacion al objetivo especifico N° 3 de este TFM, se examinaron estándares para el diseño de un procedimiento, como ser el estándar ISO 10013, que aporta información sobre la estructura del mismo [38]. Asimismo se consideró la Norma ISO 9001:2015, para definiciones de proceso y procedimiento usadas en este trabajo. Si bien los procedimientos no son obligatorios en una organización, es conveniente tenerlos, a fin de facilitar la documentación de tareas y actividades realizadas.

Los procedimientos diseñados en este TFM facilitan la ejecución y documentación de actividades desarrolladas en un proyecto de explotación de la información y se pueden aplicar a otros casos de estudio, previa adaptación de los mismos a las necesidades de cada caso, es decir, a los problemas que presentan y al conjunto de datos disponible.

Asimismo, el estándar BPMN que se eligio como notación para diagramar las tareas a realizar en el proyecto de explotación de la información en las distintas fases de CRISP-DM, ayudó a ejecutar las tareas y mejorar el entendimiento de las mismas, debido a que la notación permite interpretar fácilmente lo que se debe hacer.

En relación al objetivo específico 4, la revisión de la literatura posibilitó el análisis de métricas usadas por los distintos autores en sus estudios, para evaluar los modelos obtenidos.

La validación del procedimiento de explotación de la información diseñado, se realizó en base a datos reales y se obtuvieron varios modelos como resultado de la aplicación de los algoritmos de minería de datos, que fueron evaluados a partir de las métricas obtenidas y con la ayuda de expertos en el dominio.

El análisis y comparación de las métricas obtenidas como resultados de los algoritmos, posibilitó la comprobación de qué algoritmo arrojó mejores resultados (o que modelo mostró un mejor desempeño para el conjunto de datos analizado).

En relación al objetivo de este TFM, se diseñó un procedimiento de explotación de la información que permite la detección de problemáticas laborales y sus factores de incidencia en un conjunto de datos estructurados de población, mediante técnicas de minería de datos. Este procedimiento ayudó a establecer el ordenamiento de las actividades requeridas en un proyecto de explotación de la información, la evaluación de los modelos y la mejora de resultados obtenidos.

5.2. Aportes y ámbitos de aplicación del trabajo realizado

Se espera que el procedimiento de explotación de la información diseñado, como así también la validación realizada y los resultados obtenidos que se documentaron en este TFM, sirvan de antecedente para futuras investigaciones en el campo de la explotación de datos. Con el diseño del procedimiento, en el que se hace una adaptación de la metodología CRISP-DM, se pretende colaborar con la aplicación de mejores prácticas, herramientas y métodos en la explotación de datos, que posibiliten la detección de patrones de comportamiento en conjuntos de datos estructurados.

La explotación de información y en particular el procedimiento de explotación de la información diseñado, son herramientas importantes a ser usadas en el campo de las relaciones laborales. Se considera, que el conocimiento adquirido en este TFM constituye un aporte relevante para estudios similares y futuras investigaciones en este ámbito. Asimismo, el trabajo realizado en este TFM, aporta conocimiento y

sienta las bases para la generación de otras investigaciones en el campo laboral, ya que no se cuenta en la región con trabajos similares. En este contexto, en el ámbito académico, específicamente en el de la Carrera de Relaciones Laborales (UNNE), algunos de los aportes que se consideran importantes, son:

- La información obtenida en este trabajo servirá como antecedente para el cambio de Plan Curricular de la Carrera Licenciatura en Relaciones Laborales, de la Facultad de Ciencias Económicas, de la UNNE.
- Es un aporte para futuras investigaciones que se realicen en la Carrera Licenciatura en Relaciones Laborales, que promuevan la generación de acciones desde el Departamento de Relaciones del Trabajo, de la Facultad de Ciencias Económicas, de la UNNE, que contribuyan al mejoramiento de estas poblaciones.

5.3. Conclusiones y futuras líneas de investigación

Las tecnologías aplicadas en explotación de información y minería de datos están en constante evolución, por lo que se prevé que en el futuro surgirán herramientas y técnicas de relevancia que harán que la explotación de datos sea cada vez más utilizada y se convierta en una herramienta de uso habitual para el análisis de datos.

El procedimiento diseñado facilitó la realización de un proyecto de explotación de información, aplicado a un caso de estudio; permitió establecer una forma ordenada de ejecución este tipo de proyectos y obtener modelos o conocimiento a partir de los datos mediante técnicas de minería de datos. Como futura línea de investigación, se propone estudiar nuevas tecnologías de explotación de datos, que permitan profundizar el estudio en el campo laboral, relacionado a PEA en zonas urbanas. En este contexto, se considera fundamental seguir indagando sobre distintos estudios en los que se hayan probado nuevos modelos o metodologías orientadas al análisis de datos y validar el procedimiento diseñado en otros contextos. Asimismo, podrian surgir nuevas investigaciones que consideren la aplicación combinada de distintos algoritmos de minería de datos o que utilicen tecnologías que permitan el análisis de grandes cantidades de datos, que posibiliten la producción de nuevos conocimientos de forma eficiente y rápida.

Referencias

[1] Norma Internacional ISO 9000, "Sistemas de gestión de la calidad. Fundamentos y vocabulario", ISO 9000:2015, ISO, Suiza, 2015. [En línea]. Disponible: https://uadeo.mx/wp-content/uploads/2020/11/NORMA-ISO-9000-2015.pdf. Accedido: 28-Jun-2024.

[2] Instituto Nacional de Estadística y Censos, INDEC, "Encuesta Permanente de Hogares. Conceptos de Condición de Actividad, Subocupación Horaria y Categoría Ocupacional", 2011. [En línea]. Disponible: https://www.indec.gob.ar/ftp/cuadros/menusuperior/eph/EPH_Conceptos.pdf. Accedido: 28-Jun-2024.

[3] Organización Internacional del Trabajo, "¿Qué es el trabajo decente?", 2004. [En línea]. Disponible: https://www.ilo.org/americas/sala-de-prensa/WCMS_LIM_653_SP/lang--es/index.htm. Accedido: 28-Jun-2024.

[4] P. Britos, "Procesos de explotación de información basados en sistemas inteligentes", Tesis Doctoral, Univ.Nac. de La Plata, Argentina, 2008. [En línea]. Disponible: https://sedici.unlp.edu.ar/bitstream/handle/10915/4142/Documento_completo.pdf?sequence=1&isAllowed=y. Accedido: 28-Jun-2024.

[5] S. Martins, "Modelo de proceso para proyectos de explotación de información", Tesis Doctoral, Univ.Nac. de La Plata, Argentina, 2020. [En línea]. Disponible: http://sedici.unlp.edu.ar/handle/10915/111195. Accedido: 28-Jun-2024.

[6] A. L. Samuel, "Some studies in machine learning using the game of checkers", IBM, 2000. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5389202. Accessed: Jun 22, 2024.

[7] I.H. Witten, E. Frank, M.A. Hall, Ch.J. Pal, "Data Mining: Practical Machine Learning Tools and Techniques", Fourth Edition, Morgan Kaufmann, 2016. [Online]. Available: https://ml.cms.waikato.ac.nz/weka/book.html. Accessed: May 24, 2024.

[8] E. F. Franco y R. J. Ramos, "Aprendizaje de máquina y aprendizaje profundo en biotecnología: aplicaciones, impactos y desafíos", Ciencia, Ambiente y Clima, 2019. [En línea]. Disponible: https://revistas.intec.edu.do/index.php/cienacli/article/view/1579/2173. Accedido: 28-Jun-2024.

[9] W. Maass, & V. C. Storey, "Pairing conceptual modeling with machine learning". Data & Knowledge Engineering, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X21000367. Accessed: May 24, 2024.

[10] I. Goodfellow, Y. Bengio & A. Courville, "Deep Leaning", 2016. [Online]. Available: https://www.deeplearningbook.org/contents/ml.html. Accessed: May 24, 2024.

[11] J. S. Nolasco Valenzuela, "Python Aplicaciones prácticas", Inteligencia Artificial. Data Science, ISBN: 978-84-9964-758-6, España, 2018. [En línea]. Disponible: https://webooks.co/images/team/academicos/ingenieria/informatica/programac

ion/lenguajepython/5.Python_Aplicaciones_practicas_Jorge_Sant.pdf. Accedido: 27-abr-2024.

[12] G. James, D. Witten, T. Hastie & R. Tibshirani, "An introduction to statistical Learning: with Applications in R", Vol. 112, p. 18, New York: Springer, 2013. [Online]. Available: https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf. Accessed: May 24, 2024.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, .. & E. Duchesnay, "Scikit-learn: Machine learning in Python", the Journal of machine Learning research, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Accessed: May 24, 2024.

[14] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases", AI Magazine, 17(3): 37-54, 1996. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230. Accessed: May 24, 2024.

[15] J. M. Moine, "Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo", Tesis de Maestría, Univ. Nac. de La Plata, 2013. [En línea]. Disponible: http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1&isAllowed=y. Accedido: 30-mar-2024.

[16] M. F. Pollo Cattaneo *et al*., "Elementos para una Ingeniería de Explotación de Información", Proyecciones, 10, 2012. [En línea]. Disponible: https://sedici.unlp.edu.ar/bitstream/handle/10915/73853/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y. Accedido: 30-mar-2024.

[17] R. García Martínez *et al*., "Explotación de información geográfica basada en integración de ambientes de trabajo", In XX Congreso Argentino de Ciencias de la Computación. 2014. [En línea]. Disponible: http://sedici.unlp.edu.ar/bitstream/handle/10915/42265/Documento_completo.pdf?sequence=1&isAllowed=y. Accedido: 30-mar-2024.

[18] P. V. Britos y R. García Martínez, "Propuesta de Procesos de Explotación de Información", In XV Congreso Argentino de Ciencias de la Computación, 2009. [En línea]. Disponible: http://sedici.unlp.edu.ar/bitstream/handle/10915/21206/Documento_completo.pdf?sequence=1&isAllowed=y. Accedido: 23-mar-2024.

[19] M.F. Pollo Cattaneo, "Modelo de proceso para elicitación de requerimientos en proyectos de Explotación de Información", Tesis Doctoral, Univ. Nac. de La Plata, 2017, [En línea], Disponible: http://sedici.unlp.edu.ar/handle/10915/96315.

[20] S. Martins, P.M. Pesado y R. García Martínez, "Propuesta de proceso de ingeniería de explotación de información centrado en control y gestión del proyecto", In XX Congreso Argentino de Ciencias de la Computación, Buenos Aires, Argentina, 2014. [En línea], Disponible: https://sedici.unlp.edu.ar/handle/10915/42285. Accedido: 28-Jun-2024.

[21] J. Vanrell, "Un Modelo de Procesos para Proyectos de Explotación de Información", Tesis de Maestría, Univ. Tec. Nacional, Buenos Aires, Argentina, 2011. [En línea]. Disponible: https://www.yumpu.com/es/document/view/36281675/un-modelo-de-procesos-para-proyectos-de-explotacian-de-. Accedido: 23-mar-2024.

[22] H. Kuna, "Procedimientos de Explotación de Información para la Identificación de Datos Faltantes con Ruido e Inconsistentes", Tesis Doctoral, Dep. Leng.y Cs.de la Comp., Univ. de Málaga, 2013. [En línea]. Disponible: https://core.ac.uk/download/pdf/62901022.pdf. Accedido: 30-mar-2024.

[23] K. B. Eckert, "Trabajo Final de Maestría en Tecnologías de la Información. Modelo basado en la toma decisiones con criterios múltiples para la elección de metodologías de data science", Tesis de Maestría, Universidad Nacional de Misiones, 2019. [En línea]. Disponible: https://rid.unam.edu.ar/bitstream/handle/20.500.12219/2186/Eckert_2019_Modelo.pdf?sequence=4&isAllowed=y. Accedido: 30-mar-2024.

[24] J. García, J. Molina, A. Berlanga, M. Patricio, A. Bustamante y W. Padilla, "Ciencia de datos. Técnicas Analíticas y Aprendizaje Estadístico", Bogotá, Colombia, Publicaciones Altaria, 2018. [En linea], Disponible: https://librosfh.mdp.edu.ar/ebooks/index.php/fh/catalog/download/25/19/114-1?inline=1. Accedido: 28-Jun-2024.

[25] W. Maass & V.C. Storey "Pairing conceptual modeling with machine learning", Data & Knowledge Engineering", 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169023X21000367. Accessed: May 24, 2024.

[26] F. Martínez-Plumed *et al*., "CRISP-DM twenty years later: From data mining processes to data science trajectories", IEEE Transactions on Knowledge and Data Engineering, 2019. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8943998. Accessed: May 24, 2024.

[27] SAS Help Center, "Introduction to SEMMA", 2023. [Online], Available: https://documentation.sas.com/doc/en/emref/15.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm. Accessed: May 24, 2024.

[28] SAS Help Center, "What Is SAS Enterprise Miner?". [Online], Available: https://documentation.sas.com/doc/es/emgsj/15.3/n1fio8qutibtuxn1j7hyo6rqw7bj.htm Accessed: May 24, 2024.

[29] IBM SPSS Modeler, "Conceptos básicos de ayuda de CRISP-DM", [En línea]. Disponible: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview. Accedido: 28-Jun-2024.

[30] Orange Datamining, "Correspondence Analysis", [Online], Available: https://orangedatamining.com/widget-catalog/unsupervised/correspondenceanalysis/. Accessed: May 24, 2024.

[31] M. Greenacre, "La práctica del análisis de correspondencias", Fundación BBVA, 2008. [En linea], Disponible: https://www.fbbva.es/wp-content/uploads/2017/05/dat/DE_2008_practica_analisis_correspondencias.pdf Accedido: 28-Jun-2024.

[32] Orange Datamining, "Mosaic Display", [Online], Available: https://orangedatamining.com/widget-catalog/visualize/mosaicdisplay/. Accessed: May 24, 2024.

[33] Orange Widget Base, "Getting Started", [Online], Available: https://orange-widget-base.readthedocs.io/en/latest/tutorial.html. Accessed: May 24, 2024.

[34] Orange Datamining, "Interactive Data Visualization", [Online], Available: https://orangedatamining.com/home/interactive-data-visualization/. Accessed: May 24, 2024.

[35] International Organization for Standardization, October 2021, [Online], Available: https://www.iso.org/home.html. Accessed: Jul 19, 2024.

[36] Estándar Internacional ISO/IEC 17799, "Tecnología de la Información-Técnicas de seguridad-Código para la práctica de la gestión de la seguridad de la información", [En linea], Disponible: https://mmujica.files.wordpress.com/2007/07/iso-17799-2005-castellano.pdf. Accedido: 19-Jul-2024.

[37] Norma Internacional ISO 9001:2015, "Sistemas de gestión de la calidad-Requisitos", [En linea], Disponible: https://repositorio.buap.mx/rcontraloria/public/inf_public/2019/0/NOM_ISO_9001-2015.pdf. Accedido: 19-Jul-2024.

[38] AulaFacil, "Enfoque de la Norma ISO 10013: Directrices para la documentación de sistemas de gestión de la calidad", [En línea], Disponible: https://www.aulafacil.com/cursos/administracion/sistema-gestioncalidad-iso-9001-enfoque-por-procesos-elaboracion-de-manuales-iso-10013-ydirectrices-para-auditoria/enfoque-de-la-norma-iso-10013-directrices-para-adocumentacion-de-sistemas-de-gestion-de-la-calidad-l36576. Accedido: 19-Jul-2024.

[39] ISO/IEC 19501:2005, "Information technology. Open Distributed Processing. Unified Modeling Language (UML)", Version 1.4.2., [Online], Available: https://www.iso.org/standard/32620.html. Accessed: Jul 19, 2024.

[40] UML. "What is UML. Introduction to Omg's Unified Modeling Language", [Online], Available: https://www.uml.org/what-is-uml.htm. Accessed: Jul 19, 2024.

[41] OMG, "Object Management Group. Standards Development Organization", [Online], Available: https://www.omg.org/about/index.htm. Accessed: Jun 15, 2024.

[42] ISO/IEC 19510:2013, "Information technology. Object Management Group Business Process Model and Notation", [Online], Available: https://www.iso.org/standard/62652.html. Accessed: Jun 15, 2024.

[43] J. F. Gómez Estupiñán, "Análisis de BPMN como herramienta integral para el Modelado de Procesos de Negocio", Vent. Inform., n.o 30, pp. 75-77, 2014, [En línea], Disponible: https://revistasum.umanizales.edu.co/ojs/index.php/ventanainformatica/article/view/274/397. Accedido: 12-Jul-2024.

[44] BPMN Quick Guide, "Business Process Model and Notation (BPMN)", Version 2.0. OMG, [Online], Available:

https://www.bpmnquickguide.com/view-bpmn-quick-guide/. Accessed: Jun 15, 2024.

[45] IBM Analytics, "Metodología Fundamental para la Ciencia de Datos", [En línea], Disponible: https://www.ibm.com/downloads/cas/WKK9DX51. Accedido: 12-Jul-2024.

[46] Microsoft, "¿Qué es el Proceso de ciencia de datos en equipo (TDSP)?", [En línea], Disponible: https://learn.microsoft.com/es-es/azure/architecture/data-science-process/overview. Accedido: 12-Jul-2024.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Anexos

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

Anexo A: Mercado de trabajo. Tasas e indicadores socioeconómicos

Según datos del INDEC, en Argentina, la tasa de desocupación (personas que no tienen ocupación, están disponibles para trabajar y buscan empleo activamente), como proporción de la PEA se ubicó en **7,6%,** en el segundo trimestre de 2024, considerando el total de aglomerados urbanos (Fuente: https://www.indec.gob.ar/uploads/informesdeprensa/mercado_trabajo_eph_2trim2404BDC_5E521.pdf).

Por otra parte, según datos publicados por el Instituto Provincial de Estadística y Ciencia de Datos, organismo público rector de la actividad estadística en la provincia de Corrientes. Perteneciente al Ministerio de Hacienda y Finanzas de la provincia, la desocupación en el segundo trimestre de 2024 fue del **8.9 %** (Fuente: https://estadistica.corrientes.gob.ar/).

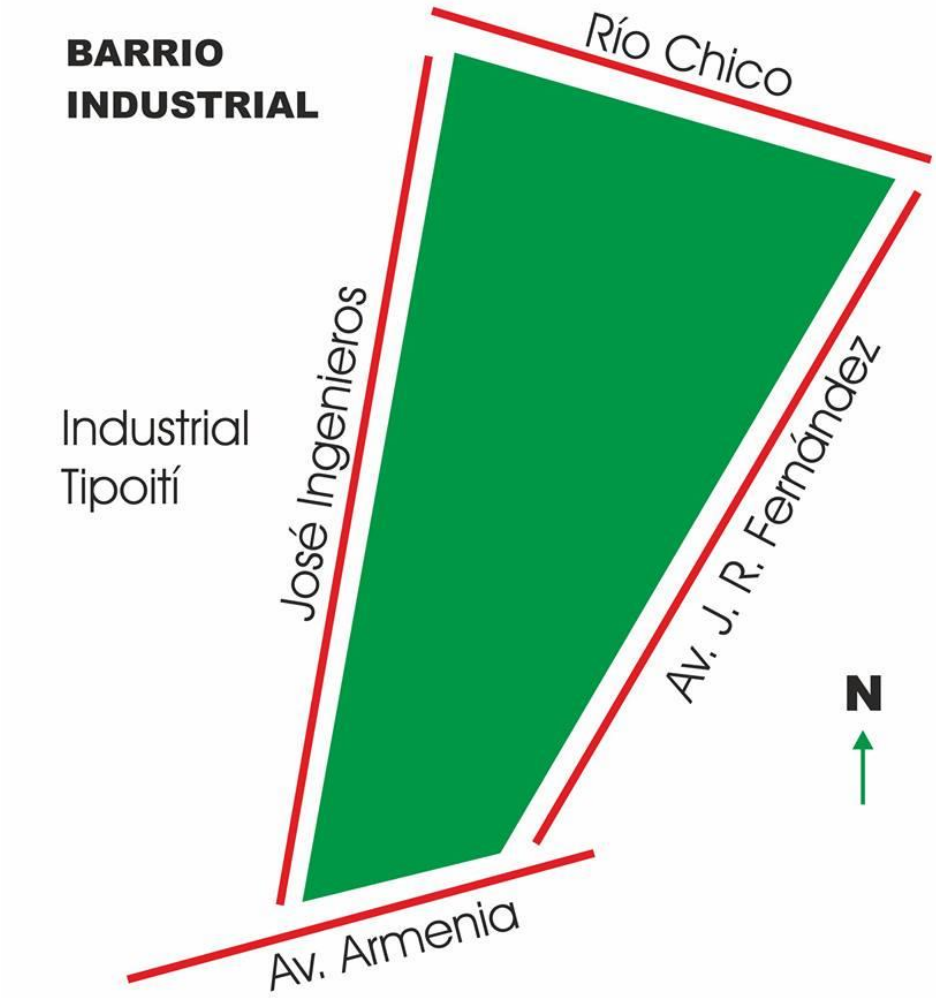
Anexo B: Información sobre Relevamiento Sociodemográfico "Barrio Industrial"

a)- Ubicación del Barrio Industrial.

Relevamiento Sociodemográfico - Barrio Industrial

Ciudad de Corrientes

(Año 2016)



b)- Bloques correspondientes al cuestionario.

Bloque I- Viviendas y Hogar – Características Habitacionales

Bloque II- Población

Bloque III – Traslado de Personas fuera del Barrio

Bloque IV – Atención Médica

Bloque V – Anexo 1- Estrategias laborales de Supervivencia Económica

Bloque VI - Anexo 2- Expectativas laborales de jóvenes de 15 a 19 años que asistan a colegio secundario

Bloque VII - Anexo 3- Planes Sociales – Inserción Laboral

Bloque VIII – Anexo 4-Empleabilidad

A continuación se presenta el cuadro de distribución de la muestra por segmentos con el total de viviendas listadas.

R:15	
Segmentos	**Conteo Viviendas**
1	12
3	14
5	12
7	16
9	14
11	15
12	16
13	15
16	16
17	14
18	14
19	15
20	10
21	10
22	14
23	15

R:22	
Segmentos	**Conteo Viviendas**
2	16
4	15
5	12
7	10
9	14
10	14
14	11
16	15
18	14
20	17
23	14
26	16
29	14

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

c)- Formulario del Relevamiento

RELEVAMIENTO SOCIO – DEMOGRAFICO DEL BARRIO INDUSTRIAL

(MUNICIPIO CAPITAL)

CARACTER ESTRICTAMENTE CONFIDENCIAL Y RESERVADO - Ley N° 17.622

Año 2016

Carátula del Segmento

1 - UBICACION GEOGRAFICA DEL SEGMENTO

Código Localidad / Paraje:		**Localidad / Paraje:**			
Fracción Nº:		**Radio Nº:**		**Segmento Nº:**	

2 - ENTORNO E INFRAESTRUCTURA DEL SEGMENTO

Instrucción: formule estas preguntas en el primer hogar que cense		SI	NO
¿ Se registraron inundaciones en los últimos cinco años ?	1	1	1
¿ Hay algún basural permanente a menos de 300 metros (tres cuadras) ?	2	1	1
¿ Hay servicio de cloacas ?	3	1	1
¿ Hay servicio de agua de red (agua corriente) ?	4	1	1
¿ Hay servicio de energía eléctrica por red domiciliaria ?	5	1	1
¿ Hay servicio regular de recolección de residuos (al menos dos veces por semana) ?	6	1	1
¿ Hay transporte público a menos de 300 metros (tres cuadras) ?	7	1	1
¿ Hay teléfono público, semipúblico o locutorio a menos de 300 metros (tres cuadras) ?	8	1	1
Instrucción: marque por observación y considere la situación predominante del segmento			
Ubicado en villa (de emergencia) a asentamiento	9	1	1
Ubicado en barrio, plan o monoblock	10	1	1
Ubicado en un country o barrio cerrado	11	1	1
Existencia de al menos una cuadra pavimentada	12	1	1
Existencia de al menos una boca de tormenta o alcantarilla	13	1	1
Existencia de alumbrado público	14	1	1

3 - RESUMEN DEL SEGMENTO

Total de Población:		Total de Varones:		Total de Mujeres:	
Ultimo número de vivienda del segmento:				Total de cuestionarios entregados:	

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

d)- Bloque I- Viviendas y Hogar – Características Habitacionales

RELEVAMIENTO SOCIO – DEMOGRAFICO DEL BARRIO INDUSTRIAL
(MUNICIPIO CAPITAL)
CARACTER ESTRICTAMENTE CONFIDENCIAL Y RESERVADO - Ley N° 17.622
Año 2016

Cuestionario de VIVIENDAS PARTICULARES

UBICACION GEOGRAFICA

Cód. Calle: | **Calle:** | **Puerta Nº** | **Piso Nº** | **Dpto./Pieza Nº** | **Fracción Nº:** | **Radio Nº:** | **Segmento Nº:** | **Manzana Nº:** | **Vivienda Nº:**

VIVIENDA Y HOGAR

1. Tipo de vivienda particular *(Anote por observación)*

1 Casa
2 Rancho
3 Casilla
4 Departamento
5 Pieza en inquilinato
6 Pieza en hotel familiar o pensión
7 Local no construído para habitación
8 Vivienda móvil
9 Persona/s viviendo en la calle → *pase a Preg. 4*

2. La vivienda está ...

habitada
1 con personas presentes
2 con todas las personas temporalmente ausentes

deshabitada
3 en alquiler o venta
4 en construcción
5 se usa como comercio, oficina o consultorio
6 se usa para vacaciones, fin de semana u otro uso temporal
7 por otra razón

→ *Fin de la entrevista*

ATENCION: Recuerde que, un hogar es la persona o grupo de personas que comparten los gastos de comida y/o alquiler, impuestos, servicios, etc. y vive/n bajo el mismo techo

3. Deteccion de viviendas y hogares

	Sí	No
1 Existen otras viviendas en la misma dirección?	1	1
Atención hay mas de una vivienda		
2 Todas las personas que habitan en esta vivienda comparten gastos de alimentos y/o alquiler, impuestos, servicios, etc.?	2	1 → *Atención hay mas de un hogar*
3 ¿Habitan en ésta vivienda personas del servicio doméstico, cama adentro?	3	1
4 ¿Habitan pensionistas en ésta vivienda?	4	1
Atención hay mas de un hogar		

. Abra un cuestionario para cada hogar detectado en la vivienda.
. En caso de que detecte más de un hogar en la vivienda, abra un nuevo cuestionario y transcriba los mismos datos de **Ubicación Geográfica** incluyendo el mismo número de Vivienda en el nuevo cuestionario. Continúe en la pregunta **4** →
. Numere correlativamente los hogares en el casillero **Hogar Nº.**

5 Cantidad de hogares en la vivienda (Hogares particulares mas servicio doméstico y pensionistas)
6 Hogar Nº

4. ¿Quiénes son las personas que habitan este hogar, incluyendo los bebés, niños y ancianos?
LISTA DE PERSONAS QUE RESIDEN EN EL HOGAR
. Anote los nombres comenzando por el jefe o jefa del hogar.
. Siga completando la lista tratando de continuar en el siguiente orden: cónyuge o pareja, hijo(a) / hijastro(a), yerno / nuera, nieto(a), padre / madre / suegro(a), otros familiares, otros no familiares, servicio doméstico y sus familiares.

Persona Nº	Nombre	Relación o parentesco con el Jefe /a
0 1		JEFE / JEFA

Si en este hogar hay más de seis personas:
· Abra un nuevo cuestionario y transcriba los mismos datos de **Ubicación Geográfica**, incluyendo **Vivienda Nº.**
· Anote el mismo número de hogar en **Hogar Nº** (pregunta 3).
Pase a la pregunta **4** y continúe la entrevista, dejando en blanco el primer renglón.

1

CARACTERISTICAS HABITACIONALES DEL HOGAR

5. ¿Cuál es el material predominante de los pisos ...
1 ☐ Cerámica, baldosa, mosaico, mármol, madera o alfombrado?
2 ☐ cemento o ladrillo fijo?
3 ☐ tierra o ladrillo suelto?
4 ☐ otro

6. ¿Cuál es el material predominante de la cubierta exterior del techo ...
1 ☐ Cubierta asfáltica o membrana?
2 ☐ baldosa o losa *(sin cubierta)*?
3 ☐ pizarra o teja?
4 ☐ chapa de metal *(sin cubierta)*?
5 ☐ chapa de fibrocemento o plástico?
6 ☐ chapa de cartón?
7 ☐ caña, palma, tabla o paja, con o sin barro?
8 ☐ otro?

7.¿El techo tiene revestimiento interior ó cielorraso?
1 ☐ Si
2 ☐ No

8. Tiene agua...
1 ☐ por cañería dentro de la vivienda?
2 ☐ fuera de la vivienda pero dentro del terreno?
3 ☐ fuera del terreno?

9. El agua que usa, ¿proviene de ...
1 ☐ red pública?
2 ☐ perforación con bomba a motor?
3 ☐ perforación con bomba manual?
4 ☐ pozo?
5 ☐ transporte por cisterna?
6 ☐ agua de lluvia, río, canal, arroyo o acequia?

10. Este hogar tiene, ¿baño / letrina?
1 ☐ Sí
2 ☐ No → *pase a Preg. 14*

11. En el baño, ¿tiene botón, cadena, mochila para limpieza del inodoro?
1 ☐ Sí
2 ☐ No

12. El desague del inodoro, ¿es...
1 ☐ a red pública *(cloaca)*?
2 ☐ a cámara séptica y pozo ciego?
3 ☐ solo a pozo ciego?
4 ☐ a hoyo, excavación en la tierra,etc.?

13. El baño / letrina, ¿es...
1 ☐ usado sólo por este hogar?
2 ☐ compartido con otros hogares?

14. Para cocinar, ¿utiliza principalmente...
1 ☐ gas a granel *(zeppelin)*?
2 ☐ gas en tubo 45 kg.?
3 ☐ gas en garrafa?
4 ☐ electricidad?
5 ☐ leña o carbón?
6 ☐ Otros (*especificar*):.....
.....

15. ¿Cuántas habitaciones o piezas para dormir tiene este hogar?
Cantidad de habitaciones o piezas para dormir ☐☐

16. Y en total, ¿cuántas habitaciones ó piezas tiene este hogar? *(sin contar baños, cocinas, etc., ver Glosario pag. 88)*
Cantidad de habitaciones o piezas en total ☐☐

17. La vivienda que ocupa este hogar, es...
1 ☐ propia?
2 ☐ alquilada?
3 ☐ prestada?
4 ☐ cedida por trabajo?
5 ☐ otra situación
→ *pase a Preg. 19*

18. ¿El terreno es propio?
1 ☐ Sí
2 ☐ No

19. Este hogar, ¿tiene... *(Respuesta Múltiple)*

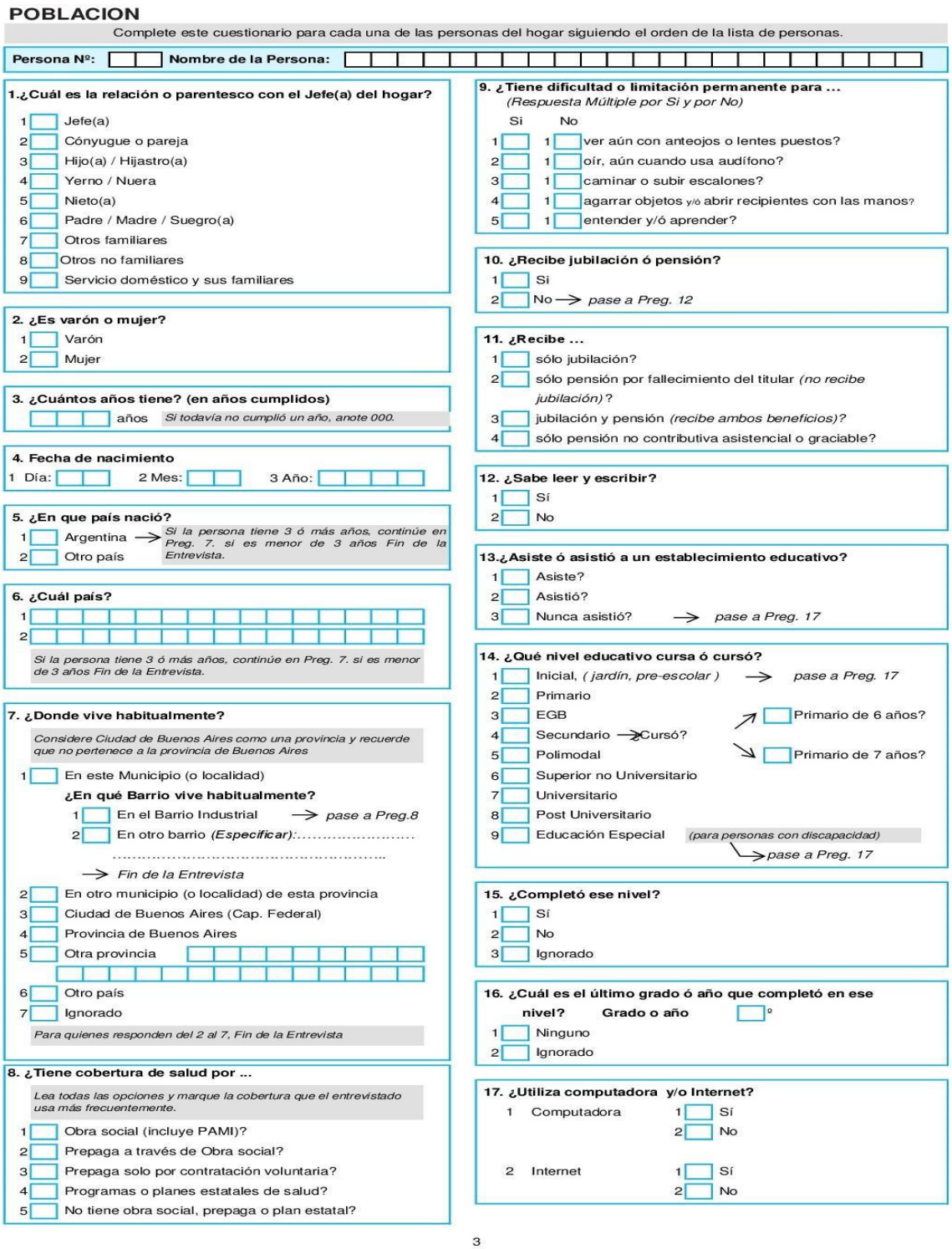
	Sí	No
heladera?	1 ☐	1 ☐
computadora?	2 ☐	1 ☐
teléfono celular?	3 ☐	1 ☐
teléfono de línea?	4 ☐	1 ☐
acceso a Internet?	5 ☐	1 ☐

pase a Población

2

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

e)- Bloque II- Población



POBLACION

Complete este cuestionario para cada una de las personas del hogar siguiendo el orden de la lista de personas.

Persona Nº: ☐☐ **Nombre de la Persona:** ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

1.¿Cuál es la relación o parentesco con el Jefe(a) del hogar?

1 ☐ Jefe(a)
2 ☐ Cónyugue o pareja
3 ☐ Hijo(a) / Hijastro(a)
4 ☐ Yerno / Nuera
5 ☐ Nieto(a)
6 ☐ Padre / Madre / Suegro(a)
7 ☐ Otros familiares
8 ☐ Otros no familiares
9 ☐ Servicio doméstico y sus familiares

2. ¿Es varón o mujer?

1 ☐ Varón
2 ☐ Mujer

3. ¿Cuántos años tiene? (en años cumplidos)

☐☐☐ años *Si todavía no cumplió un año, anote 000.*

4. Fecha de nacimiento

1 Día: ☐☐ 2 Mes: ☐☐ 3 Año: ☐☐☐☐

5. ¿En que país nació?

1 ☐ Argentina → *Si la persona tiene 3 ó más años, continúe en Preg. 7. si es menor de 3 años Fin de la Entrevista.*
2 ☐ Otro país

6. ¿Cuál país?

1 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐
2 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Si la persona tiene 3 ó más años, continúe en Preg. 7. si es menor de 3 años Fin de la Entrevista.

7. ¿Donde vive habitualmente?

Considere Ciudad de Buenos Aires como una provincia y recuerde que no pertenece a la provincia de Buenos Aires

1 ☐ En este Municipio (o localidad)
¿En qué Barrio vive habitualmente?
1 ☐ En el Barrio Industrial → *pase a Preg.8*
2 ☐ En otro barrio *(Especificar):*.....
.....
→ *Fin de la Entrevista*
2 ☐ En otro municipio (o localidad) de esta provincia
3 ☐ Ciudad de Buenos Aires (Cap. Federal)
4 ☐ Provincia de Buenos Aires
5 ☐ Otra provincia ☐☐☐☐☐☐☐☐☐☐
☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐
6 ☐ Otro país
7 ☐ Ignorado

Para quienes responden del 2 al 7, Fin de la Entrevista

8. ¿Tiene cobertura de salud por ...

Lea todas las opciones y marque la cobertura que el entrevistado usa más frecuentemente.

1 ☐ Obra social (incluye PAMI)?
2 ☐ Prepaga a través de Obra social?
3 ☐ Prepaga solo por contratación voluntaria?
4 ☐ Programas o planes estatales de salud?
5 ☐ No tiene obra social, prepaga o plan estatal?

9. ¿Tiene dificultad o limitación permanente para ...
(Respuesta Múltiple por Sí y por No)

	Sí	No	
1	☐	1 ☐	ver aún con anteojos o lentes puestos?
2	☐	1 ☐	oír, aún cuando usa audífono?
3	☐	1 ☐	caminar o subir escalones?
4	☐	1 ☐	agarrar objetos y/ó abrir recipientes con las manos?
5	☐	1 ☐	entender y/ó aprender?

10. ¿Recibe jubilación ó pensión?

1 ☐ Sí
2 ☐ No → *pase a Preg. 12*

11. ¿Recibe ...

1 ☐ sólo jubilación?
2 ☐ sólo pensión por fallecimiento del titular *(no recibe jubilación)*?
3 ☐ jubilación y pensión *(recibe ambos beneficios)*?
4 ☐ sólo pensión no contributiva asistencial o graciable?

12. ¿Sabe leer y escribir?

1 ☐ Sí
2 ☐ No

13.¿Asiste ó asistió a un establecimiento educativo?

1 ☐ Asiste?
2 ☐ Asistió?
3 ☐ Nunca asistió? → *pase a Preg. 17*

14. ¿Qué nivel educativo cursa ó cursó?

1 ☐ Inicial, *(jardín, pre-escolar)* → *pase a Preg. 17*
2 ☐ Primario
3 ☐ EGB
4 ☐ Secundario → Cursó? ↗ ☐ Primario de 6 años? ↘ ☐ Primario de 7 años?
5 ☐ Polimodal
6 ☐ Superior no Universitario
7 ☐ Universitario
8 ☐ Post Universitario
9 ☐ Educación Especial *(para personas con discapacidad)* → *pase a Preg. 17*

15. ¿Completó ese nivel?

1 ☐ Sí
2 ☐ No
3 ☐ Ignorado

16. ¿Cuál es el último grado ó año que completó en ese nivel? Grado o año ☐º

1 ☐ Ninguno
2 ☐ Ignorado

17. ¿Utiliza computadora y/o Internet?

1	Computadora	1 ☐	Sí
		2 ☐	No
2	Internet	1 ☐	Sí
		2 ☐	No

3

A partir de aquí contestan todas las personas de 14 años o más

18. ¿Convive en pareja o matrimonio?
1 ☐ Sí
2 ☐ No

19. Durante la semana pasada, ¿trabajó por lo menos una hora ? *(sin contar las tareas de su hogar)*
1 ☐ Sí → *pase a Preg. 23*
2 ☐ No

20. En esa semana, ¿hizo alguna changa, algo para vender fuera o ayudó a un familiar/amigo en una chacra o negocio?
1 ☐ Sí → *pase a Preg. 23*
2 ☐ No

21. En esa semana, ¿tenía trabajo, pero estuvo de licencia por vacaciones, enfermedad, suspensión, conflicto laboral, etc.?
1 ☐ Sí → *pase a Preg. 23*
2 ☐ No

22. En las últimas 4 semanas, ¿estuvo buscando trabajo: contesto avisos, consultó amigos/parientes, puso carteles, hizo algo para ponerse por su cuenta ?
1 ☐ Sí
2 ☐ No → *Mujer de 14 años ó más continúa en Preg. 30, Varón continúa en Preg. 34*

Las preguntas 23 a 29 refieren al trabajo donde pasa más horas

23. ¿A qué se dedica o qué servicio presta la empresa o el lugar en que trabaja más horas?
1 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐
2 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐
3 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

24. ¿Cuál es el nombre de la ocupación?
Si el nombre del cargo de una persona explica claramente el tipo de trabajo, indique este nombre (por ejemplo "cocinero" ó "maestro") de lo contrario describa el tipo de trabajo que realiza durante la semana.
1 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐
2 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐
3 ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

25. ¿Cuantas personas hay en la empresa o lugar donde trabaja?
1 ☐ Hasta 5 personas
2 ☐ De 6 a 25 personas
3 ☐ De 26 a 100 personas
4 ☐ Más de 100 personas

26. ¿En ese trabajo es...
1 ☐ Obrero(a) empleado(a)?
2 ☐ Patrón(a)?
3 ☐ Trabajador(a) por cuenta propia? → *pase a Preg. 29*
4 ☐ Trabajador(a) familiar? → *pase a Preg. 28*

27. ¿Trabaja en el sector...
1 ☐ público nacional?
2 ☐ público provincial?
3 ☐ público municipal?
4 ☐ privado?

28. En ese trabajo, ¿le descuentan para la jubilación?
1 ☐ Sí → *Mujer de 14 años ó más continúa en Preg. 30, Varón continúa en Preg. 34*
2 ☐ No
3 ☐ Ignorado

29. En ese trabajo,¿aporta por sí mismo para la jubilación?
1 ☐ Sí
2 ☐ No
3 ☐ Ignorado
→ *Mujer de 14 años ó más continúa en Preg. 30, Varón continúa en Preg. 34*

A partir de aquí contestan todas las mujeres de 14 años o más

Recuerde preguntar a las más jóvenes y a las solteras

30. ¿Tuvo hijas o hijos nacidos vivos?
1 ☐ Sí
2 ☐ No → *pase a Preg. 34*

31. ¿Cuántas hijas e hijos nacidos vivos tuvo en total?
Cantidad de hijos é hijas nacidos vivos ☐☐

32. ¿Cuántas hijas e hijos están vivos actualmente?
Cantidad de hijos é hijas
que están vivos actualmente ☐☐

33. ¿Cuál es la fecha de nacimiento de su último hijo o hija nacido(a) vivo(a)?
Mes: ☐☐ Año: ☐☐☐☐

4

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

f)- Bloque III – Traslado de Personas fuera del Barrio

g)- Bloque IV – Atención Médica

TRASLADO DE LAS PERSONAS FUERA DEL BARRIO - ATENCION MEDICA

Complete este cuestionario para cada una de las personas de la lista.

Persona Nº: ☐☐ **Nombre de la Persona:** ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

TRASLADO DE LAS PERSONAS FUERA DEL BARRIO

34. ¿ Con qué frecuencia se traslada fuera del barrio?

1 ☐ Diariamente
2 ☐ Una vez por semana
3 ☐ Más de una vez por semana
4 ☐ Una vez por mes
5 ☐ Más de una vez por mes
6 ☐ Nunca? ⟶ *pase a Preg. 38*
88 ☐ Otra frecuencia *(especificar)*:.....
.....

35. Indique el motivo por el cual se traslada principalmente fuera del Barrio:

1 ☐ Trabajo
2 ☐ Educación
4 ☐ Salud *(para control ó asistencia médica)*
88 ☐ Otros *(especificar)*: .....
.....
.....

36. ¿ Qué distancia máxima recorre habitualmente *-en forma aproximada-* desde su vivienda, por motivos de:...

(Respuesta Múltiple)	Distancia en cuadras	No Sabe	No Corresponde
1 ☐ Trabajo?	☐☐☐	☐	☐
2 ☐ Educación?	☐☐☐	☐	☐
4 ☐ Salud *(para control ó asistencia médica)?*	☐☐☐	☐	☐
88 ☐ Otros (especificar):	☐☐☐	☐	☐

37. ¿Por qué medios se traslada principalmente?

1 ☐ a pié
2 ☐ colectivo
3 ☐ carro
4 ☐ bicicleta
5 ☐ moto
6 ☐ auto
7 ☐ remis o taxi
88 ☐ Otros medios que utiliza para trasladarse *(especificar)*:.....
.....

ATENCION MEDICA

38. ¿ Este año sintió algún malestar ó tuvo algún problema de salud ?

1 ☐ SI
2 ☐ NO ⟶ *pase a Preg. 41*

39. ¿ Este año consultó a un ... *(Respuesta Múltiple)* *Lea cada opción por separado y marque SI o NO en cada una*

SI	NO	
1 ☐	1 ☐	médico ? *(clínico y/o especialista)*
2 ☐	1 ☐	dentista ?
3 ☐	1 ☐	sicólogo ?
4 ☐	1 ☐	psiquiatra ?
5 ☐	1 ☐	kinesiólogo ?

88 ☐ Otros *(especificar)*: .....

Si contestó SI a alguna de las opciones pase a Preg. 41

40. ¿ Porqué no consultó?

1 ☐ No tenía tiempo
2 ☐ No tenía dinero
3 ☐ No le pareció importante
4 ☐ No fue necesario, me sentía bien de salud
88 ☐ Otra razón *(especificar)*: .....
.....
.....

41. ¿ Dónde atiende habitualmente su salud? *(Respuesta Múltiple)*

1 ☐ Hospital "Ángela Iglesia de Llano"
2 ☐ Hospital "Dr. Jose Ramon Vidal"
3 ☐ Hospital Escuela "Gral. José Francisco de San Martín"
4 ☐ Hospital Pediátrico "Juan Pablo II"
5 ☐ Hospital de Salud Mental "San Francisco de Asis"
6 ☐ **Instituto de Cardiología de Corrientes "Juana Francisca Cabral"**
7 ☐ Instituto Dermatológico "H. Cáceres de Blaquier"
8 ☐ C.A.P.S.
9 ☐ S.A.P.S.
10 ☐ D.A.P.S.
11 ☐ Clínicas Privadas
12 ☐ Consultorios Privados
88 ☐ En otra localidad *(especificar)*: .....
.....
.....

5

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

h)- Bloque V – Anexo 1- Estrategias laborales de Supervivencia Económica

ANEXO 1: ESTRATEGIAS LABORALES DE SUPERVIVENCIA ECONOMICA

Este anexo es respondido por el Jefe/Jefa de Hogar.

Persona Nº: ☐☐ **Nombre de la Persona:** ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

1, ¿Realiza Ud. o algún miembro del hogar, alguna actividad de manera informal para obtener dinero, y asi poder cubrir sus necesidades básicas de alimento y vivienda?

1 ☐ Si → *Pase a la Preg. 2*
2 ☐ No → *Fin del anexo 1.*

2. ¿Qué actividades realiza Ud o algun miembro del hogar para obtener dinero suficiente que le permita cubrir sus necesidades de alimento y vivienda? (Respuesta múltiple)

1 ☐ Albañil
2 ☐ Carrero/a
3 ☐ Cartonero/a
4 ☐ Costurero/a
5 ☐ Carrero/a
6 ☐ Cartonero/a
7 ☐ Cuidado de enfermos
8 ☐ Cuidado de niños
9 ☐ Elaboración de alimentos
10 ☐ Elaboración de artesanias
11 ☐ Electricista
12 ☐ Empleada doméstica por hora
13 ☐ Maestra particular
14 ☐ Mallonero
15 ☐ Motomandado
16 ☐ Remisero
17 ☐ Técnico
18 ☐ Trapito
88 ☐ Otros (*especificar*): ……

3. ¿Qué motivos dieron inicio a estas actividades? *(Respuesta múltiple)*

1 ☐ Crisis económica
2 ☐ Nacimiento de un hijo
3 ☐ Crecimiento del grupo familiar y/o del hogar
4 ☐ Pérdida del puesto de trabajo
88 ☐ Otros (*especificar*): ……

4. ¿Hace cuánto tiempo realizan estas actividades? En caso de que marquen más de una opcion contestar en base a la actividad que llevan más tiempo desarrollando.

1 ☐ Menos de 1 año
2 ☐ Entre 1 y 5 años
3 ☐ Entre 6 y 10 años
4 ☐ Más de 10 años

5. De los integrantes del hogar, ¿Quiénes conforman el equipo de trabajo?

1 ☐ Solo adultos del grupo familiar
2 ☐ Adultos y niños del grupo familiar
3 ☐ Más de un grupo familiar
4 ☐ Sólo el jefe/jefa de hogar

6. De ese equipo de trabajo, ¿Cuánto tiempo en el dia dedican a realizar estas actividades?

1 ☐ Entre 1 y 2 horas
2 ☐ Entre 3 y 4 horas
3 ☐ Entre 5 y 6 horas
4 ☐ Entre 7 y 8 horas
5 ☐ Más de 8 horas

7. El dinero que obtienen por estas actividades, ¿Les alcanza para cubrir sus necesidades básicas de alimento y vivienda?

1 ☐ Alcanza
2 ☐ Alcanza parcialmente
3 ☐ Alcanza con un plan social
4 ☐ Alcanza con otra actividad
5 ☐ No alcanza ni cuenta con un plan social

8. Algunos miembros del equipo de trabajo, ¿Asiste a algún tipo de taller de oficio para aprender la actividad elegida?

1 ☐ Si
2 ☐ No

9. ¿A través de qué medios financiaron el inicio de estas actividades?

1 ☐ Dinero ahorrado
2 ☐ Dinero de préstamo de algún familiar
3 ☐ Dinero de otra actividad que realiza
4 ☐ Subsidio del Estado
88 ☐ Otros *(especificar)*: ……

10. ¿En qué lugar desarrollan estas actividades?

1 ☐ Desde el hogar
2 ☐ Fuera del hogar
3 ☐ Una parte desde el hogar y otra fuera del hogar

Fin del anexo 1.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

i)- Bloque VI - Anexo 2- Expectativas laborales de jóvenes de 15 a 19 años que asistan a colegio secundario

ANEXO 2: EXPECTATIVAS LABORALES DE JOVENES DE 15 A 19 AÑOS QUE ESTEN ASISTIENDO AL COLEGIO SECUNDARIO

Persona Nº: ☐☐ **Nombre de la Persona:** ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

1. ¿ Que te gustaría hacer al terminar el colegio secundario...

1 ☐ Estudiar? ☐ → *pase a Preg. 7*
2 ☐ Trabajar?
3 ☐ Estudiar y Trabajar?
4 ☐ No Sabe? ☐ → *Pase a Preg. 8*

2. Con la formación recibida hasta el momento en el colegio...¿creés que te será fácil encontrar trabajo?

1 ☐ Si
2 ☐ No
3 ☐ No Sabe

3. ¿Cuáles es la razón principal que te impulsa a trabajar...

1 ☐ Lograr independencia económica?
2 ☐ Aprender cosas nuevas?
3 ☐ Relacionarte con personas?
4 ☐ Sentirte útil?
5 ☐ Ocupar el tiempo?
6 ☐ Porque no te gusta estudiar?
7 ☐ Ayudar a tu familia económicamente?
8 ☐ Porque formaste tu propia familia?
9 ☐ Para contribuir a la sociedad?
10 ☐ Para sentirte parte de algún grupo?
11 ☐ Por considerarlo una obligación?
12 ☐ Otros *(especificar)* : .....

4. ¿ Que tipo de trabajo buscarías...

1 ☐ **En relación de dependencia en el sector privado?**
- 1 ☐ Cadetería (Ej: Motomandado)
- 2 ☐ Empresa Mediana o Grande/Multinacional
- 3 ☐ En una Oficina (Ej: Estudio Jurídico)
- 4 ☐ Fábrica/Taller
- 5 ☐ Negocio/Local (Pyme)

2 ☐ **Comenzar un emprendimiento personal?**
- 1 ☐ Crear una Pyme
- 2 ☐ Iniciar un emprendimiento en el hogar
- 3 ☐ Realización de un servicio (Ej: Plomería,Doméstica,etc
- 4 ☐ Ventas vía internet
- 5 ☐ Venta ambulante en la vía pública
- 88 ☐ Otros (especificar): .....

3 ☐ **Continuar o adherirse a un emprendimiento familiar?**
- 1 ☐ Iniciar un emprendimiento en el hogar
- 2 ☐ Fabrica/Taller
- 3 ☐ Pyme familiar
- 4 ☐ Realización de un servicio (Plomería,Domestica,etc)
- 5 ☐ Ventas vía internet

4 ☐ **En un Organismo Público? ¿Por qué?**
- 1 ☐ Por considerar la posibilidad concreta de ascenso dentro del Organismo
- 2 ☐ Por una cuestión horaria
- 3 ☐ Por facilidad en el ingreso al tener algún contacto?
- 4 ☐ Para lograr estabilidad laboral
- 5 ☐ Otro motivo (*especificar*):.....

5 ☐ **Ingresar a algúna Fuerza Armada y/o de Seguridad?**
- 1 ☐ Ejercito
- 2 ☐ Fuerza Aerea
- 3 ☐ Gerdamería
- 4 ☐ Marina
- 5 ☐ Penitenciario
- 6 ☐ Policía
- 7 ☐ Seguridad Privada

5. ¿Considerás la posibilidad de irte a otra Provincia/País a buscar trabajo?

1 ☐ Si
2 ☐ No
¿Por que razón? .....
.....

6. ¿Tenés alguna experiencia laboral en alguna pasantía?

1 ☐ Si
2 ☐ No
¿ Dónde? .....

Si optó en la pregunta n° 1 por la respuesta n° 2 (Trabajar), pase a la Preg. 8

7. ¿ Que posibilidades creés tener de realizar estos estudios...

1 ☐ Mucha?
2 ☐ Poco?
3 ☐ Ningúna?

8. Tu familia te aconseja que al terminar el colegio...

1 ☐ Continues estudiando?
2 ☐ Que comiences a trabajar?
3 ☐ Que Estudias y Trabajes?
4 ☐ No es un tema tratado hasta el momento?

9. ¿Tenés una formación/estudio paralelo al que te brinda el colegio secundario?

1 ☐ Si
2 ☐ No
¿Cuál? .....
.....

10. ¿Tenés acceso a Internet en forma fluida?

1 ☐ Si
2 ☐ No
3 ☐ No tengo acceso a Internet

11. ¿Creés que el acceso a Internet te facilitará el ingreso a un trabajo?

1 ☐ Si
2 ☐ No
¿Por qué? .....
.....

Fin del anexo 2.

j)- Bloque VII - Anexo 3- Planes Sociales – Inserción Laboral

ANEXO 3: PLANES SOCIALES - INSERCIÓN LABORAL. (RESPONDE LA POBLACIÓN DE 18 A 65 AÑOS.)

Persona N° ☐☐ **Nombre de la Persona:** ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

1. ¿Está de acuerdo con que los Planes Sociales ayudan conseguir trabajo?
(Respuesta Múltiple)

Motivo. ¿Por qué?

1 ☐ Sí
- 1 ☐ Proveen herramientas para conseguir empleo.
- 2 ☐ Forman en un oficio.
- 3 ☐ Acceden al primer empleo.
- 88 ☐ Otros (*especificar*): ....

2 ☐ No
- 1 ☐ Fomentan el ocio.
- 2 ☐ No promueven la búsqueda de empleo.
- 3 ☐ No incentivan la cultura del trabajo.
- 88 ☐ Otros (*especificar*):..

2. ¿Es o fue beneficiario de un Plan Social?
(Respuesta Múltiple)

1 ☐ Es
2 ☐ Fue → pase a Preg. 3
3 ☐ Nunca fue → *pase a Preg. 6*

Responde el que ES o FUE Titular y/o Beneficiario de un Plan Social.

3. ¿Percibe o percibió algún Plan Social?. ¿Cuál?
(Respuesta Múltiple).

Percibe	Percibió	
1 ☐	1 ☐	Argentina Enseña, Trabaja y Aprende.
2 ☐	2 ☐	Compremos lo nuestro.
3 ☐	3 ☐	Ellas Hacen.
4 ☐	4 ☐	Emprendedores de nuestra tierra.
5 ☐	5 ☐	Entrenamiento para el trabajo.
6 ☐	6 ☐	Ingreso social con Trabajo.
7 ☐	7 ☐	Jóvenes con Mas y Mejor Trabajo.
8 ☐	8 ☐	Marca colectiva.
9 ☐	9 ☐	Microcréditos.
10 ☐	10 ☐	Plan Jefes y Jefas de Hogar.
11 ☐	11 ☐	PROG.R.ES.AR
12 ☐	12 ☐	Programa de empleo independiente.
13 ☐	13 ☐	Programa de Inserción Laboral.
14 ☐	14 ☐	Promoción del empleo.
15 ☐	15 ☐	Proyecto Manos a la Obra.
	16 ☐	Seguro de capacitación y empleo.
88 ☐	88 ☐	Otros (*especificar*)..

4. ¿Qué contraprestación realiza o realizaba?
(Respuesta Múltiple)

- 1 ☐ Asistir cursos de capacitación y entrenamiento laboral.
- 2 ☐ Terminar escuela primaria y/o secundaria.
- 3 ☐ Trabajar en una empresa.
- 4 ☐ Trabajar en una cooperativa.
- 5 ☐ Elaborar proyecto de microemprendimiento.
- 6 ☐ Trabajar en un organismo público.
- 7 ☐ Sin contraprestación.
- 88 ☐ Otros (*especificar*): ....

Responde solo el que PERCIBIÓ un Plan Social.

5. ¿Dejó de percibir el Plan social porque consiguió trabajo?

1 ☐ Si → *pase a Preg. 5.1*
2 ☐ No → *Fin del anexo 3*

5.1. ¿El Plan Social lo ayudó a conseguirlo?

1 ☐ Si
2 ☐ No

Fin del anexo 3.

Responde el que NUNCA FUE Titular y/o Beneficiario de un plan social.

6. ¿Necesita un Plan Social?

1 ☐ Si
2 ☐ No ☐ → *Fin del anexo3*

7. ¿Lo tramitó?

1 ☐ Si
2 ☐ No ☐ → *pase a Preg. 9*

8. ¿Por qué aún no tiene el Plan? Porque...

- 1 ☐ Completó los tramites y no consultó?
- 2 ☐ No completó los requisitos que piden?
- 3 ☐ Aún no le adjudicaron el Plan?
- 88 ☐ Otros (*especificar*): ....

9. ¿Por qué no lo tramitó? Porque...
(Respuesta Múltiple)

- 1 ☐ Desconoce los organismos que los otorgan.
- 2 ☐ Desconoce como acceder a Planes Sociales.
- 3 ☐ No intentó.
- 88 ☐ Otros (*especificar*): ....

Fin del anexo 3.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

k)- Bloque VIII – Anexo 4-Empleabilidad

ANEXO 4: EMPLEABILIDAD (Trabajador formal entre 18 y 65 años)

Complete este formulario para todas las personas del hogar que actualmente tienen un trabajo formal y están en un rango de edad de 18 y 65 años.
Nota: Una persona tiene un trabajo Formal, si en el mismo le realizan descuentos jubilatorios. Incluir al trabajador que realice aporte jubilatorios por su cuenta.

Persona Nº: ☐☐ **Nombre de la Persona:** ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

1. ¿A qué edad tuvo su primer trabajo formal?
Edad.....años cumplidos.

2. ¿Continua en ese trabajo formal?
1 ☐ Si → *pase a Preg. 4* 2 ☐ No

3. ¿A qué edad consiguio su trabajo actual?
Edad.....años cumplidos.

4. ¿Tuvo alguna dificultad para conseguirlo?
1 ☐ Si 2 ☐ No ——→ *pase a Preg 6*

5. Indique los motivos de la dificultad *(Respuesta múltiple)*
1 ☐ Falta de formación
2 ☐ Falta de experiencia
3 ☐ Mucha edad
4 ☐ Tener familia a cargo
5 ☐ No conocer idiomas
6 ☐ No manejar herramientas informáticas
7 ☐ Sobrecalificación
8 ☐ Muchos candidatos
88 ☐ Otros (especificar): .....
99 ☐ Ns / Nc .....
.....
.....

6. ¿Para conseguir ese trabajo usted... *(Respuesta múltiple)*
1 ☐ Envío CV por correo postal?
2 ☐ Realizó anuncios en la via pública ?
3 ☐ Fue llamado por un contratista?
4 ☐ Realizó un concurso / inscripción?
5 ☐ Le avisó a amigos/conocidos que buscaba trabajo?
6 ☐ Se dirigió a Agencias de empleo/bolsas de trabajo?
7 ☐ Respondió avisos de diarios/radio?
8 ☐ Ofrecio trabajo por cuenta propia?
9 ☐ Se presentó personalmente?
10 ☐ Envío CV por Email?
11 ☐ Realizó avisos por Red social?
12 ☐ Se inscribío en Internet a una Bolsa de trabajo?
13 ☐ Se inscribío en portal o web de la empresa?

7. ¿Tenía experiencia laboral antes de conseguirlo?
1 ☐ Si 2 ☐ No → *pase a Preg 8*
7.1 ¿De que tipo?
1 ☐ Administrativas
2 ☐ Oficios
88 ☐ Otras (*especificar*):.....
.....
7.2 ¿Su trabajo está relacionado con la experiencia laboral previa?
1 ☐ Si 2 ☐ No
7.3¿Se valoró su experiencia para ingresar a su trabajo?
1 ☐ Si 2 ☐ No
7.4¿Es reconocida su experiencia laboral en su trabajo?
1 ☐ Si 2 ☐ No → *pase a Preg. 7.5*
7.4.1 ¿De que manera?
1 ☐ Ascenso
2 ☐ Reconocimiento monetario
3 ☐ Formar parte de equipos de trabajo
88 ☐ Otras (especificar):.....
7.5 ¿Esa experiencia está relacionada con sus funciones actuales?
1 ☐ Si 2 ☐ No

8.¿Realizó cursos de capacitación previo al ingreso a su trabajo?
1 ☐ Si 2 ☐ No → *pase a Preg 9*
8.1 ¿De que tipos?
1 ☐ Administrativos
2 ☐ Oficios
88 ☐ Otros (*especificar*):.....
.....
.....
.....
8.2 ¿Su trabajo está relacionado con la capacitación laboral?
1 ☐ Si 2 ☐ No
8.3 ¿Fue valorada la capacitación para el ingreso en su trabajo?
1 ☐ Si 2 ☐ No
8.4 ¿Es reconocida esa capacitación en su trabajo?
1 ☐ Si 2 ☐ No → *pase a Preg 8.5*
8.4.1 ¿Cómo?
1 ☐ Ascenso
2 ☐ Reconocimiento monetario
3 ☐ Formar parte de equipos de trabajo
88 ☐ Otras (especificar):.....
.....
8.5 ¿Esa capacitación está relacionada con sus funciones actuales?
1 ☐ Si 2 ☐ No

9. ¿Donde Ud. trabaja Ofrecen oportunidad de capacitarse?
1 ☐ Si 2 ☐ No

10. ¿Actualmente Ud. realiza Curso/s de capacitación...
1 ☐ Ofrecidos por la empresa?
2 ☐ En forma particular ?
3 ☐ No realiza?
10.1 ¿Referidos a que? *(Respuesta multiple)*
1 ☐ Su profesión
2 ☐ Mejorar la forma de hacer tareas o actividades
3 ☐ Intereses personales
4 ☐ Trabajo actual
88 ☐ Otros (*especificar*):.....
.....
.....
.....
.....

11. Donde Ud, trabaja. ¿lo evaluan de alguna forma?
1 ☐ Si 2 ☐ No
11.1 ¿En que consiste?
1 ☐ Autoevaluación
2 ☐ Evaluación a través de jefe inmediato superior
3 ☐ Examen múltiple
98 ☐ Ns / Nc
88 ☐ Otros (*especificar*):.....
.....
.....

12. En su futuro laboral ¿Qué le gustaria hacer?
1 ☐ Tener otro trabajo en relación de dependencia
2 ☐ Trabajar por cuenta propia
3 ☐ Mantener el trabajo actual

Fin del anexo 4

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

l)- Instrumento de Recolección de Datos Específico:

Anexo 2- Expectativas laborales de jóvenes de 15 a 19 años que asistan a colegio secundario

ANEXO 2: EXPECTATIVAS LABORALES DE JOVENES DE 15 A 19 AÑOS QUE ESTEN ASISTIENDO AL COLEGIO SECUNDARIO

Persona Nº: ☐☐ **Nombre de la Persona:** ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

1. ¿ Que te gustaría hacer al terminar el colegio secundario...

1 ☐ Estudiar? ☐ → *pase a Preg. 7*
2 ☐ Trabajar?
3 ☐ Estudiar y Trabajar?
4 ☐ No Sabe? ☐ → *Pase a Preg. 8*

2. Con la formación recibida hasta el momento en el colegio...¿creés que te será fácil encontrar trabajo?

1 ☐ Si
2 ☐ No
3 ☐ No Sabe

3. ¿Cuáles es la razón principal que te impulsa a trabajar...

1 ☐ Lograr independencia económica?
2 ☐ Aprender cosas nuevas?
3 ☐ Relacionarte con personas?
4 ☐ Sentirte útil?
5 ☐ Ocupar el tiempo?
6 ☐ Porque no te gusta estudiar?
7 ☐ Ayudar a tu familia económicamente?
8 ☐ Porque formaste tu propia familia?
9 ☐ Para contribuir a la sociedad?
10 ☐ Para sentirte parte de algún grupo?
11 ☐ Por considerarlo una obligación?
12 ☐ Otros *(especificar)* : .....

4. ¿ Que tipo de trabajo buscarías...

1 ☐ **En relación de dependencia en el sector privado?**
- 1 ☐ Cadetería (Ej: Motomandado)
- 2 ☐ Empresa Mediana o Grande/Multinacional
- 3 ☐ En una Oficina (Ej: Estudio Jurídico)
- 4 ☐ Fábrica/Taller
- 5 ☐ Negocio/Local (Pyme)

2 ☐ **Comenzar un emprendimiento personal?**
- 1 ☐ Crear una Pyme
- 2 ☐ Iniciar un emprendimiento en el hogar
- 3 ☐ Realización de un servicio (Ej: Plomería,Doméstica,etc
- 4 ☐ Ventas vía internet
- 5 ☐ Venta ambulante en la vía pública
- 88 ☐ Otros (especificar): .....

3 ☐ **Continuar o adherirse a un emprendimiento familiar?**
- 1 ☐ Iniciar un emprendimiento en el hogar
- 2 ☐ Fabrica/Taller
- 3 ☐ Pyme familiar
- 4 ☐ Realización de un servicio (Plomería,Domestica,etc)
- 5 ☐ Ventas vía internet

4 ☐ **En un Organismo Público? ¿Por qué?**
- 1 ☐ Por considerar la posibilidad concreta de ascenso dentro del Organismo
- 2 ☐ Por una cuestión horaria
- 3 ☐ Por facilidad en el ingreso al tener algún contacto?
- 4 ☐ Para lograr estabilidad laboral
- 5 ☐ Otro motivo (*especificar*):.....

5 ☐ **Ingresar a algúna Fuerza Armada y/o de Seguridad?**
- 1 ☐ Ejercito
- 2 ☐ Fuerza Aerea
- 3 ☐ Gerdamería
- 4 ☐ Marina
- 5 ☐ Penitenciario
- 6 ☐ Policía
- 7 ☐ Seguridad Privada

5. ¿Considerás la posibilidad de irte a otra Provincia/País a buscar trabajo?

1 ☐ Si
2 ☐ No
¿Por que razón? .....
.....

6. ¿Tenés alguna experiencia laboral en alguna pasantía?

1 ☐ Si
2 ☐ No
¿ Dónde? .....

Si optó en la pregunta nº 1 por la respuesta nº 2 (Trabajar), pase a la Preg. 8

7. ¿ Que posibilidades creés tener de realizar estos estudios...

1 ☐ Mucha?
2 ☐ Poco?
3 ☐ Ningúna?

8. Tu familia te aconseja que al terminar el colegio...

1 ☐ Continues estudiando?
2 ☐ Que comiences a trabajar?
3 ☐ Que Estudias y Trabajes?
4 ☐ No es un tema tratado hasta el momento?

9. ¿Tenés una formación/estudio paralelo al que te brinda el colegio secundario?

1 ☐ Si
2 ☐ No
¿Cuál? .....
.....

10. ¿Tenés acceso a Internet en forma fluida?

1 ☐ Si
2 ☐ No
3 ☐ No tengo acceso a Internet

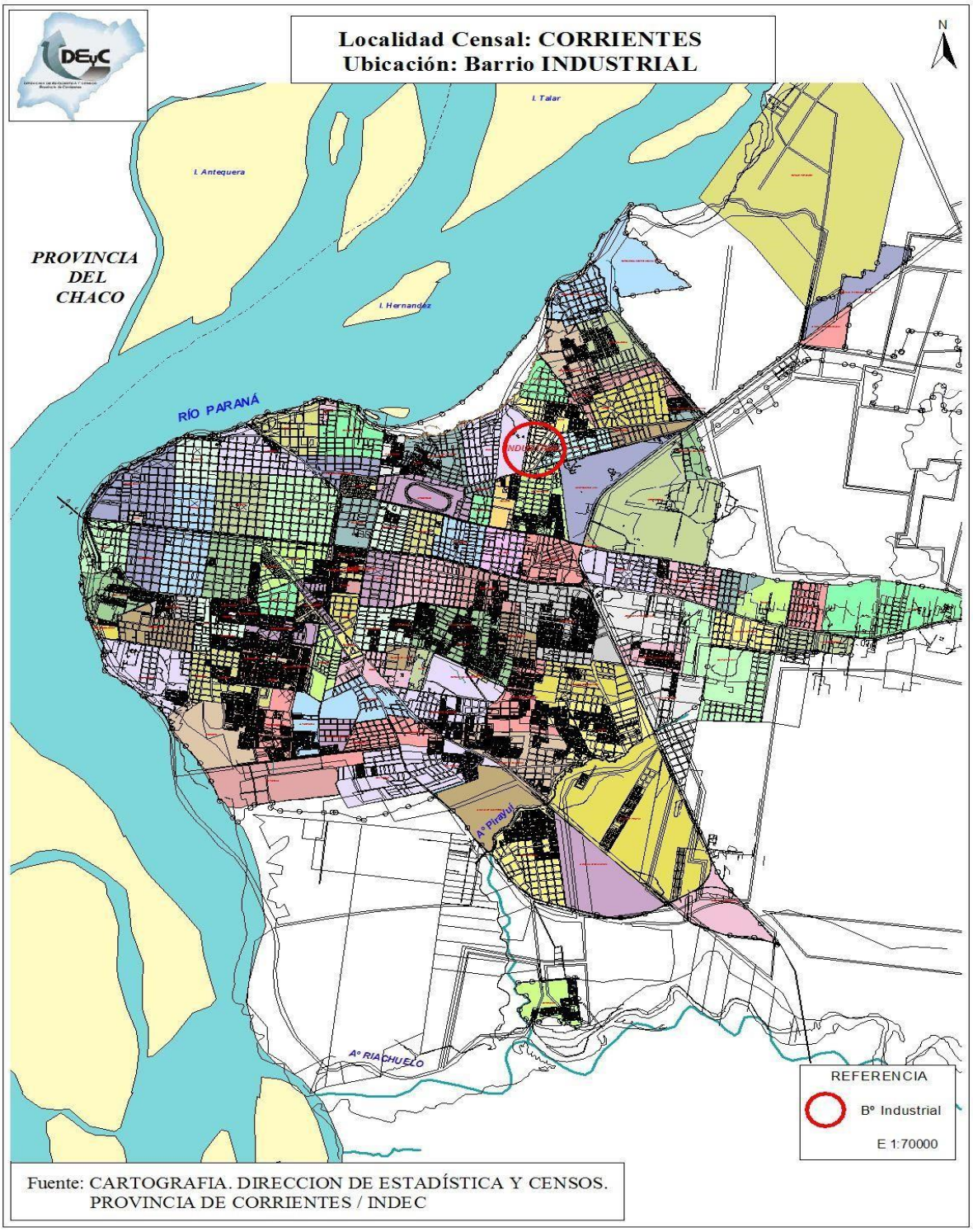
11. ¿Creés que el acceso a Internet te facilitará el ingreso a un trabajo?

1 ☐ Si
2 ☐ No
¿Por qué? .....
.....

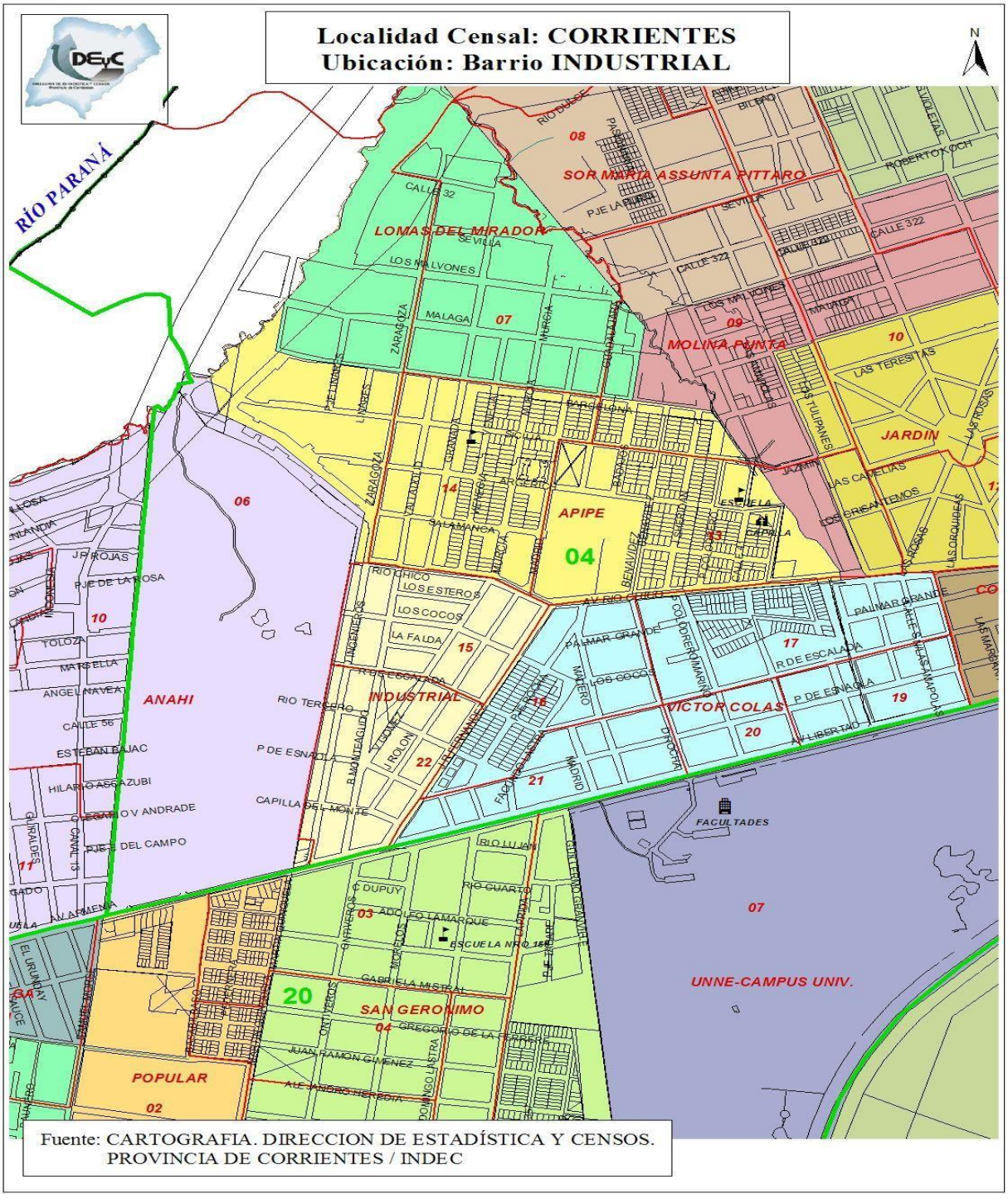
Fin del anexo 2.

Procedimiento de explotación de la información para detectar problemáticas laborales y sus factores de incidencia, basado en normas internacionales.

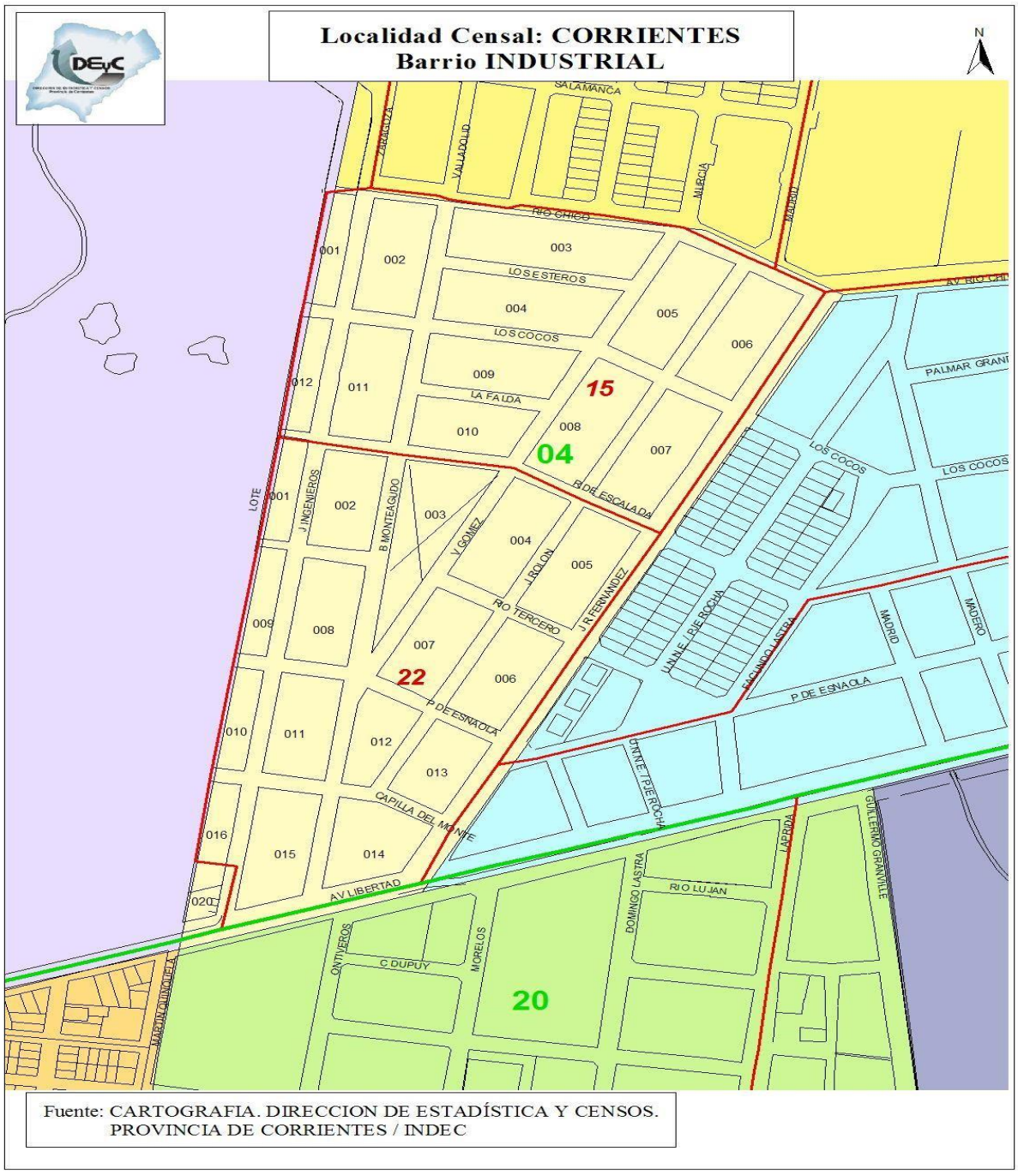
m)- Apartado II: Mapas del Barrio Industrial Mapa 1: Ubicación del Barrio Industrial en la Localidad de Corrientes Capital



n)- Apartado II: Mapas del Barrio Industrial Mapa 2: Ubicación del Barrio Industrial en la Zona Norte de la Localidad de Corrientes Capital



o)- Apartado II: Mapas del Barrio Industrial Mapa 3: Detalle de fracción censal 4, con y radios censales 15 y 22 del Barrio Industrial en la Localidad de Corrientes Capital



Anexo C: Terminología a utilizadas en este TFM

Como parte de la comprensión de esta investigación, se agrega y describe a continuación un *glosario* que establece un lenguaje común usado en este trabajo, a fin de que las personas involucradas conozcan el las terminologías relevantes utilizadas en este TFM:

- **Datos de entrenamiento:** Se refiere al conjunto de datos que se usa para entrenar un modelo de minería de datos
- **Datos de prueba:** Se refiere al conjunto de datos que se usa para probar la calidad de un modelo de minería de datos
- **Desocupados**: Personas sin ocupación que buscan trabajo activamente y están disponibles para trabajar.
- **Inactivos**: Personas que no tienen trabajo ni lo buscan activamente. Incluye el trabajo no remunerado y de cuidado relevada por EPH como *Amas de casa* y quienes no trabajan por incapacidad, los menores de 10 años, estudiantes, pensionados/as y jubilados/as.
- **Métodos descriptivos**: Permiten examinar las propiedades de los datos y proveen información sobre las relaciones que existen entre los mismos.
- **Métodos predictivos**: Permiten estimar el valor futuro de una variable *objetivo* o *explicada* (dependiente), a partir de las variables *explicativas* (independientes).
- **Ocupados**: Personas que tienen al menos una ocupación (trabajaron al menos una hora). Pueden ser, *Con descuento jubilatorio* (Trabajo Formal) o *Sin descuento jubilatorio* (Trabajo Informal).
- **PEA (Población Económicamente Activa):** Está integrada por personas que tienen una ocupación o que sin tenerla la están buscando activamente. Está compuesta por la población *ocupada* más la población *desocupada*.
- **Población no económicamente activa**: Personas no incluidas en la población económicamente activa. Por ejemplo, jubilados y estudiantes que no buscan ocupación.
- **Población urbana:** Población que reside en localidades de 2.000 o más habitantes.

Algunos de los términos usados en este TFM relacionados con explotación de información y minería de datos son [8]:

- **Técnicas de minería de datos:** Son herramientas que se usan para obtener conocimiento o encontrar patrones en un conjunto de datos. Pueden ser *predictivas descriptivas*.
- **Vista minable**: Conjunto de datos con formato tabular, en el que las filas representan las *observaciones* (ejemplos) y las columnas las variables en estudio (características). Estos datos se usan como entrada para la aplicación de algoritmos de minería de datos)
- **Widgets:** Los widgets de Orange son componentes básicos de los *flujos de trabajo* de análisis de datos que se ensamblan en el entorno de programación visual del Software Orange para minería de datos.

Anexo D: Otros enfoques utilizados en explotación de información

Además de los enfoques explicados en el capítulo2, se exponen los siguientes [5]:

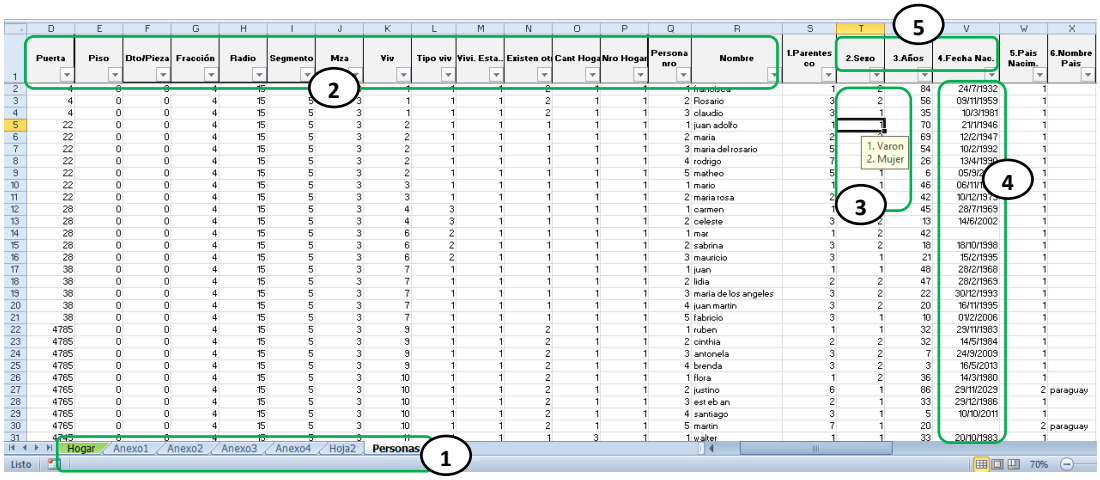
- **IKDDM** (*Integrated Knowledge Discovery and Data Mining* o *Proceso Integrado de Descubrimiento de Conocimiento y Minería de Datos*): La definición de esta metodología fue motivada por las limitaciones que presenta CRISP-DM. IKDDM es orientada al usuario, proporciona una visión detallada de las tareas y las dependencias del proceso y provee de técnicas y procedimientos que guían al usuario en el desarrollo de las actividades.
- **ASD-DM** (*Adaptive Software Development – Data Mining* o *Modelo de proceso de Desarrollo de Software Adaptativo para Minería de Datos):* Se presenta a partir de los métodos de gestión de proyectos ágiles que se utilizan en ingeniería de software.
- **ASD-BI** (*Adaptive Software Development –Business Intelligence* o *Desarrollo de Software Adaptativo para Inteligencia de Negocios*): Esta propuesta es una mejora del modelo ASD-DM, fue definida en 2.010**,** se basa en el método ágil Desarrollo de Software Adaptativo y presenta un nuevo marco de desarrollo de procesos (o ciclo de vida) alternativo a la tradicional estructura secuencial.
- **Agile KDD:** Este modelo fue propuesto en 2012 y es un proceso basado en CRISP-DM y KDD, que se estructura a partir del método ágil *Proceso Unificado Abierto* (OpenUP, Open Unified Process), que define una estructura de ciclo de vida iterativa e incremental.
- **FMDS** (*Foundational Methodology for Data Science* o *Metodología Fundamental para la Ciencia de Datos):* Fue propuesta en el año 2015 por IBM; se basa en KDD y CRISP-DM e incorpora nuevas prácticas de procesamiento de grandes volúmenes de datos, analítica de texto e imagen, inteligencia artificial, Aprendizaje Profundo y procesamiento de lenguajes [45].
- **TDSP** (*Team Data Science Process* o *Proceso de Ciencia de Datos en Equipo*), es una metodología de ciencia de datos ágil e iterativa definida por Microsoft en 2016. Proporciona soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente [46]. Se basa en un ciclo de vida ágil, permite mejorar el aprendizaje y la colaboración del equipo de trabajo y ayuda a implementar de forma correcta proyectos de ciencia de datos.

- **ADS** (*Agil Data Science* o *Ciencia de Datos Ágil*. Esta propuesta fué presentada en 2017; se basa en los modelos de ciclo de vida ágiles, promueve el desarrollo iterativo e incremental y define para cada fase del proceso, los lineamientos a seguir para el desarrollo del proyecto.

Anexo E: Datos del Relevamiento del Barrio Industrial

a)- Preparación de los Datos

Como se puede ver en la siguiente imagen, el conjunto de datos del relevamiento realizado en el Barrio Industrial se encuentra en una planilla de cálculo con varias hojas:



En esta etapa de preparación de los datos, los datos fueron limpiados a fin de poder usarlos para el estudio realizado. Algunas de las modificaciones realizadas fueron:

1)- Los datos del relevamiento estaban distribuidos en varias hojas de cálculo, por lo que debieron unificarse en una sola hoja, para esto se seleccionaron los datos más relevantes para el estudio realizado en este TFM.

2)- Muchas de las variables del conjunto de datos no aportaban información al estudio realizado (como *puerta*, *piso*, *dto/pieza*, *fracción*, *radio*, *segmento*, *manzana*, *vivienda*, etc.), estas fueron eliminadas del conjunto de datos.

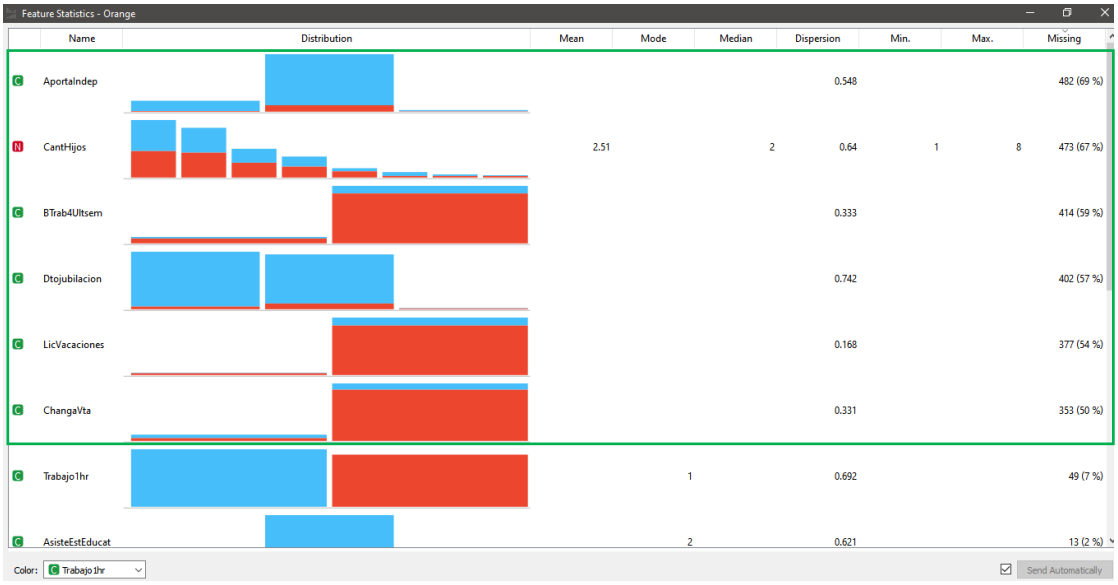
3)- La mayoría de las columnas tenían validaciones de datos que debieron ser eliminadas, como es el caso de la variable *Sexo* (mostrada en el punto 3 señalado en la imagen), a fin de que no afectara el análisis de datos.

4)- El campo *FechaNac* (con formato dd/mm/aaaa) fue analizado en conjunto con el campo *Edad* y se decidió conservar ésta ultima variable en el conjunto de datos, por considerarse más relevante para el estudio realizado.

5)- El nombre de las variables fueron arreglados de modo que no generen problemas al ser utilizados en el software Orange, como ser, *"2.Sexo"* se modificó por *"Sexo"* y *"3.Años"* se modificó por *"Años"*.

b)- Distribución de variables

La siguiente imagen muestra la distribución de variables del conjunto de datos del Relevamiento del Barrio Industrial, en relación con la variable *Trabajo1hr* (1= "Si"; 2= "No"), que indica si la persona *trabajó* una hora la semana anterior al relevamiento (en color celeste) o si la persona *No trabajó* una hora la semana anterior al relevamiento (en color rojo). Como se puede visualizar en la imagen, hay varias variables con un alto porcentaje de valores ausentes (*AportaIndep, CantHijos, BTrab4Ultsem, Dtojubilacion, LicVacaciones, ChangaVta*):



La gran cantidad de valores ausentes se presenta en la mayoría de los casos cuando *no corresponde que la persona conteste* esa pregunta. Por ejemplo:

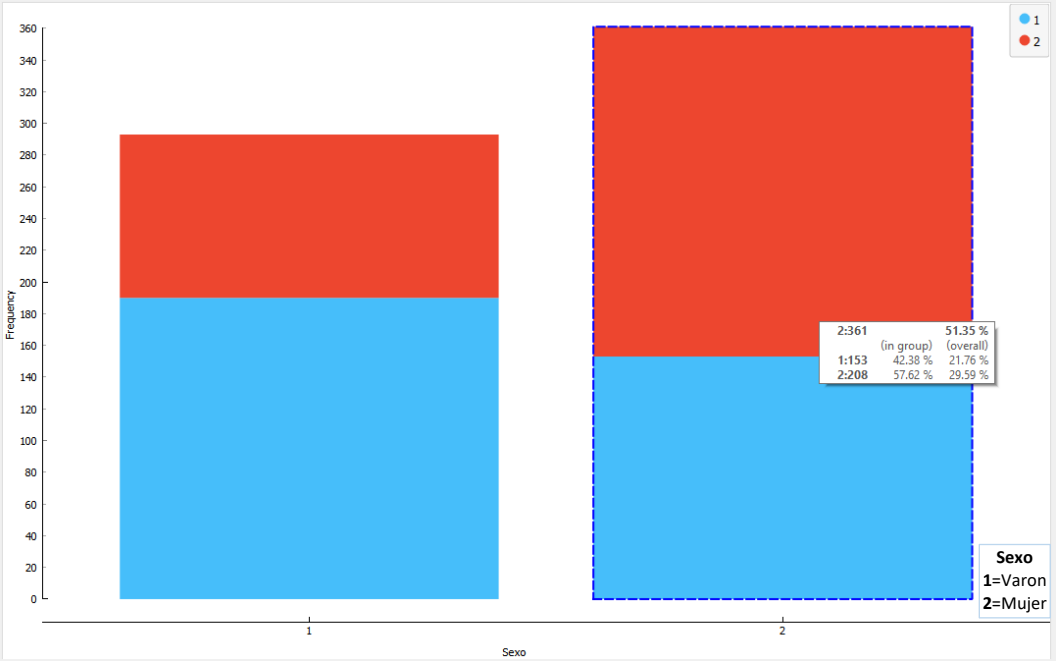
- En el caso de la primera variable *AportaIndep***,** que indica si la persona *aporta por si mismo para la jubilación, en el trabajo que realiza*, hay muchos valores ausentes (69%). En muchos casos, *no corresponde que la persona conteste* esa pregunta, por ejemplo, en el caso que sea estudiante, ama de casa o una persona Jubilada.

Variable	Descripción	Valores
AportaIndep	En ese trabajo, ¿aporta por si mismo para la jubilación?	**1= "Si"** **2= "No"** **3=Ignorado**

- En el caso de la variable *CantHijos***,** que indica *cuantos hijos vivos* tiene la persona, hay muchos valores ausentes (67%). En muchos casos, *no corresponde que la persona conteste esa pregunta*, por ejemplo, en el caso de que sean niños, estudiantes, o personas que no tengan hijos.

Variable	Descripción	Valores
CantHijos	Cantidad de hijos vivos	**1-9**

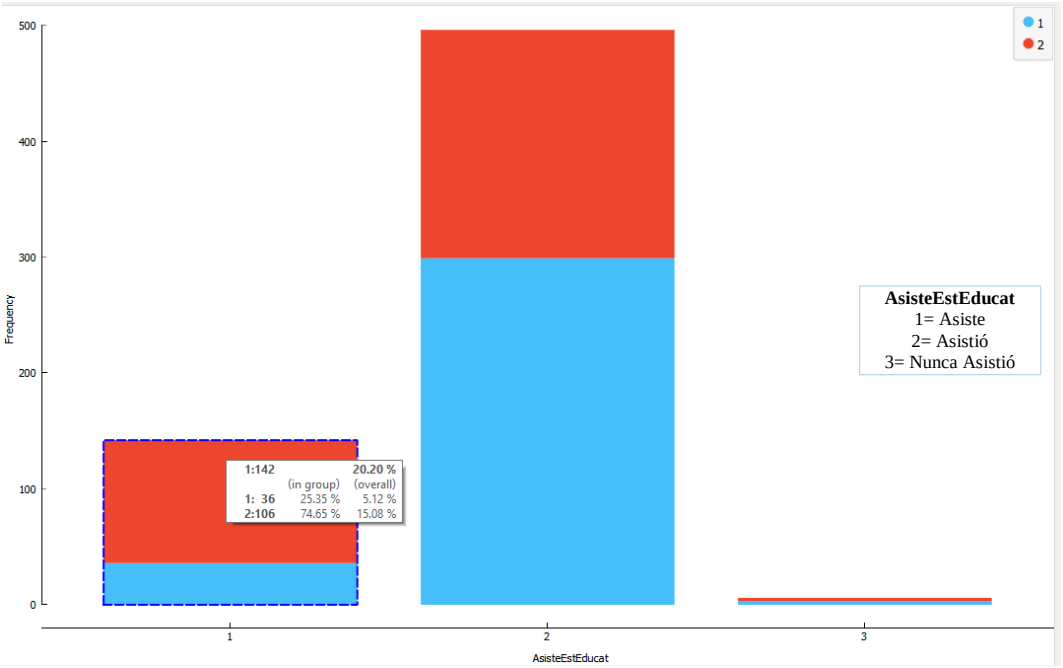
En la siguiente imagen, se muestra la distribución de la variable *Sexo*, respecto a la variable *Trabajo1hr* (1= "Si"; 2= "No"), que indica si la persona *trabajó o No trabajó* una hora la semana anterior al relevamiento (en color celeste y rojo respectivamente), se puede notar que hay una mayor cantidad de *mujeres* en comparación con los *varones* en el conjunto de datos:



Asimismo, en el grupo de *varones* se puede ver que aproximadamente el 65% de los mismos están *ocupados*. Sin embargo, en el grupo de *mujeres*, el porcentaje de *mujeres* que están *ocupadas* es menor al 50%.

En la distribución de la variable *AsisteEstEducat***,** respecto a la variable *Trabajo1hr* (1= "Si"; 2= "No"; que indica si la persona *trabajó o no trabajó* una hora la semana anterior al relevamiento, en color celeste y rojo respectivamente), se muestra que de las personas que asisten a *establecimientos educativos,* el 25% indica que está *ocupada (o trabajó una hora la semana anterior al relevamiento*).

Variable	Descripcion	Valores
AsisteEstEducat	*¿Asiste o Asistió a un establecimiento educativo?*	**1= Asiste** **2= Asistió** **3= Nunca Asistió**



Ademas, se puede ver que entre las personas que asistieron a establecimientos educativos, aproximadamente el 60 % esta *ocupado*. Asimismo, se puede ver que existe un pequeño porcentaje de personas que *nunca asistieron* a un establecimiento educativo. En este grupo, el 50 % esta *ocupado*.