Universidad Nacional del Nordeste

Facultad de Ciencias Exactas y Naturales y Agrimensura

Maestría en Tecnologías de la Información

Trabajo Final

Modelo predictivo para la detección temprana de alumnos en riesgo de abandono de la carrera de Profesorado en Ciencias de la Educación, Facultad de Humanidades de la UNNE.

Autora: E.E. y C. Viviana Elizabeth Moschner

Directora: Dra. Paola Verónica Britos

AÑO: 2021

A mi madre, por haberme inculcado, con su ejemplo, el valor del esfuerzo y la confianza en Dios.

Resumen

El abandono o deserción en la educación superior constituye una preocupación creciente para las autoridades de la Universidad Nacional del Nordeste, como también para los responsables de la gestión de cada Facultad. Para contribuir al aporte de soluciones para esta problemática, la aplicación de técnicas y herramientas utilizadas en Ciencia de Datos en el ámbito educativo tiene resultados positivos, en tanto permite predecir factores a partir de los cuales es posible implementar acciones correctivas o mitigadoras de las situaciones observadas. Este Trabajo Final de Maestría desarrolla un modelo predictivo utilizando la metodología MoProPEI orientado al descubrimiento de factores comunes en la población estudiantil del Profesorado en Ciencias de la Educación de la Facultad de Humanidades de la Universidad Nacional del Nordeste que hayan abandonado la carrera o bien presentado marcado rezago.

La fuente de datos utilizada incluye atributos personales y académicos de los alumnos del Profesorado en Ciencias de la Educación, cohortes 2010 a 2018, se utilizaron técnicas de clasificación e inducción logrando como resultado identificar factores comunes en los diferentes grupos clasificados, los que contribuirán a la elaboración de nuevas estrategias que permitan aumentar la retención estudiantil.

Palabras claves: *Deserción, Ciencia de Datos, MoProPEI*

Abstract

The dropout or desertion in higher education constitutes a growing concern for the authorities of the National University of the Northeast, as well as for those responsible for the management of each Faculty.

In order to contribute to provide solutions for this problem, the application of techniques and tools used in Data Science have shown positive results in the educational field. These techniques allow to predict factors that make possible to implement corrective or mitigating actions that would improve the observed situations.

The current Final Project develops a predictive model, using the MoProPEI methodology. This model leads to the discovery of common factors of the student population of the educational sciences program of the Faculty of Humanities of the National University of the Northeast that has dropped out or present a marked lag.

The data was taken from the cohorts 2010 to 2018 of students of the educational sciences program and includes their personal and academic characteristics. To achieve the result of identifying common factors in the different student group's classification and induction techniques were used, these results will contribute to the development of new strategies to increase student retention.

Keywords: *Dropout, Data Science, MoProPEI*

Agradecimientos

A mi Directora, Dra. Paola V. Britos, por su guía y su paciente acompañamiento en este proceso.

A los docentes y compañeros de la Maestría de la Facultad de Ciencias Exactas, Naturales y Agrimensura de la Universidad Nacional del Nordeste.

A las Profesoras Soledad Almirón y María Guadalupe Portillo, por su valiosa colaboración.

A mi familia, por la paciencia, el apoyo y la comprensión.

1 INTRODUCCIÓN 2

1.1 Contexto 2

1.1. Objetivos del Trabajo Final de Maestría 2

1.1.1 Objetivo general 2

1.1.2 Objetivos específicos 3

1.1.3 Estructura del Trabajo 3

2 MARCO TEÓRICO 5

2.1 Ciencia de Datos 5

2.2 Explotación de Información 5

2.3 Procesos de descubrimiento de patrones 6

2.3.1 Descubrimiento de Reglas de Comportamiento 6

2.3.2 Descubrimiento de Grupos 7

2.3.3 Ponderación de Interdependencia de atributos 8

2.3.4 Descubrimiento de Reglas de Pertenencia a Grupos 9

2.3.5 Ponderación de Reglas de Comportamiento 10

2.4 Proceso de Derivación de Modelos 12

2.5 Educción de requisitos 17

2.6 Herramientas de Minería de Datos disponibles 17

2.7 Deserción 18

2.8 Desgranamiento o Rezago 19

2.9 Investigaciones de Deserción y Rezago en la Educación Superior, utilizando herramientas de Explotación de la Información 19

3 METODOLOGÍAS DE EXPLOTACIÓN DE INFORMACIÓN 23

3.1 KDD 23

3.2 CRISP-DM 24

3.3 P^{3}TQ 27

3.4 SEMMA 28

3.5 MoProPEI 30

3.6 ASUM-DM 34

3.7 TSDP 34

3.8 Fases comunes de las metodologías 35

3.9 Metodología a utilizar en el TFM 36

4 ANÁLISIS DEL CASO DE ESTUDIO 38

4.1 Descripción del problema 38

4.2 Recopilación de datos 41

4.3 Características de la población en estudio 43

4.4 Políticas universitarias implementadas para aumentar la retención 45

5 DESARROLLO APLICANDO LA METODOLOGÍA MOPROPEI 48

5.1 Subproceso de gestión 48

5.1.1 Fase: Iniciación del proyecto 48

5.1.1.1 Definición de la comunicación 48

Definir Protocolo de comunicación Externa ***..... 48***

Definir Protocolo de documentación Interna ***..... 50***

Definir Protocolo de comunicación interna ***..... 50***

5.1.1.2 Exploración de conceptos iniciales 51

Planificar la adquisición de conocimientos ***..... 51***

Implementar técnicas de adquisición de conocimientos ***..... 53***

Generar el reporte de exploración inicial ***..... 55***

5.1.1.3 Evaluación de la situación 55

Identificación de recursos externos ***..... 55***

Identificación de recursos internos ***..... 57***

Identificación de las suposiciones del proyecto ***..... 57***

Identificación de riesgos del proyecto ***..... 58***

Definición del plan de contingencia ***..... 58***

Determinación de la viabilidad del proyecto ... ***59***

5.1.1.4 Definición del ciclo de vida ... 61

Selección del ciclo de vida ... ***61***

5.1.2 Fase: Planificación del proyecto ... **62**

5.1.2.1 Planificación de las actividades ... 62

Definir las actividades asociadas al proyecto ... ***62***

Identificar las métricas a realizar ... ***63***

Estimación del proyecto ... ***64***

5.1.2.2 Planificación de los recursos ... 68

Planificar la necesidad de recursos ... ***68***

Planificar la capacitación de RRHH ... ***68***

5.1.2.3 Estimaciones y responsabilidades ... 69

Estimar el tiempo de desarrollo del proyecto ... ***69***

Definir las responsabilidades de las partes ... ***69***

5.1.3 Fase: Soporte ... **70**

5.1.3.1 Gestión del ciclo de vida ... 70

Formalizar el inicio del ciclo ... ***70***

5.1.4 Fase: Gestión de control y calidad ... **70**

5.1.4.1 Control de los recursos ... 70

Controlar la capacitación de RRHH ... ***70***

5.1.5 Fase: Gestión de la entrega ... **71**

5.1.5.1 Formalización del cierre del proyecto ... 71

Verificación y validación del proyecto ... ***71***

5.2 Subproceso de desarrollo ... **71**

5.2.1 Fase: Entendimiento del dominio ... **71**

5.2.1.1 Análisis del dominio ... 72

Descripción de la terminología ... ***72***

Identificación de objetivos ... ***73***

5.2.1.2 Comprensión del problema de negocio ... 74

Identificación de los problemas de negocio ... ***74***

Identificación de los expertos en el problema ... ***75***

Identificación de los criterios de éxito del problema de negocio ... ***75***

5.2.2 Fase: Entendimiento de los datos ... **76**

5.2.2.1 Análisis de los datos ... 76

Descripción de las tablas ... ***77***

Identificar campos asociados al problema de negocio ... ***78***

5.2.2.2 Exploración de los datos ... 81

Integrar los datos en un medio digital ... ***81***

Explorar los datos ... ***81***

5.2.2.3 Evaluación de los datos ... 85

Verificación de la calidad de los datos ... ***85***

Identificación de campos riesgosos ... ***86***

5.2.3 Fase: Modelado ... **87**

5.2.3.1 Modelado del problema ... 88

Definir el problema de explotación de la información ... ***88***

Modelar el problema de explotación de la información ... ***88***

5.2.3.2 Configuración del modelo ... 89

Identificación de las herramientas alternativas ... ***89***

Identificación de los algoritmos de Minería de Datos ... ***92***

Selección de los algoritmos de Minería de Datos ... ***93***

5.2.4 Fase: Preparación de los datos ... **94**

5.2.4.1 Construcción de la fuente temporaria de datos ... 94

Definir la fuente temporaria de datos ... ***97***

Generar la fuente temporaria de datos ... ***99***

5.2.4.2 Adecuación de la fuente temporaria de datos ... 102

Limpiar los datos ... ***102***

Formatear los datos ... ***102***

5.2.5 Fase: Implementación ... **103**

5.2.5.1 Configuración de la implementación 103

Entrenar algoritmos de Minería de Datos ***104***

6 RESULTADOS 118

6.1 Modelo Predictivo para identificar factores potenciales de riesgo de abandono 118

7 CONCLUSIONES Y FUTURAS LÍNEAS DE INVESTIGACIÓN 123

BIBLIOGRAFÍA 125

ANEXO I - PROCESO DE DERIVACIÓN DE MODELOS 128

Técnica Tabla Término-Categoría-Definición del Dominio **128**

Técnica Tabla Concepto-Atributo-Relación-Valor del Dominio **130**

Técnica Tabla Concepto-Relación del Dominio **132**

Técnica Red Semántica del Modelo de Negocio **133**

Técnica Tabla Término-Categoría-Definición del Problema de Explotación de 134

Información 134

Técnica Tabla Concepto-Relación del Problema de Explotación de Información **139**

ANEXO II – REGLAS DEL MODELO 141

Reglas C5.0 ..... **141**

Reglas Rule Induction **143**

Índice de figuras

Figura 2-1: PEI-Descubrimiento de Reglas de Comportamiento [6] *7*
Figura 2-2: PEI-Descubrimiento de Grupos [6] *8*
Figura 2-3: PEI-Descubrimiento de Atributos Significativos [6] *9*
Figura 2-4: PEI-Descubrimiento de Reglas de Pertenencia a Grupos [6] *10*
Figura 2-5: PEI-Ponderación de Reglas de Comportamiento o de Pertenencia [6] *12*
Figura 2-6: Representación Gráfica de la Inserción del proceso de Derivación de Modelos [8] *13*
Figura 3-1: Etapas del Proceso de Extracción del Conocimiento [27] *24*
Figura 3-2: Fases del Modelo CRISP DM [28] *25*
Figura 3-3: Fases Componentes de la Metodología CRISP DM [28] *26*
Figura 3-4: Niveles de abstracción de procesos de CRISP-DM [5] *26*
Figura 3-5: P3TQ [29] *28*
Figura 3-6: Fases Metodología SEMMA [28] *29*
Figura 3-7: Estructura General – MoProPEI *31*
Figura 5-1: Selección de características *106*
Figura 5-2: Selección de características (Modeler) *106*
Figura 5-3: Auditoría de datos *107*
Figura 5-4: Supernodo de valores perdidos *108*
Figura 5-5: Precisión de la clasificación *109*
Figura 5-6: Matriz de confusión (%) *110*
Figura 5-7: Matriz de confusión (recuento de casillas) *110*
Figura 5-8: Informe de precisión *112*
Figura 5-9: Flujo del proceso en IBM SPSS Modeler *112*
Figura 5-10: Flujo del proceso con RapidMiner *115*

Índice de gráficos

Gráfico 4-1: Alumnos por género *43*
Gráfico 4-2: Alumnos según edad al ingresar a la carrera *44*
Gráfico 4-3: Alumnos según procedencia *44*
Gráfico 4-4: Alumnos según localidad (chaco) *44*
Gráfico 4-5: Alumnos según título secundario *45*
Gráfico 6-1: Clasificación de grupos *119*

Índice de tablas

Tabla 2-1: Carga de Trabajo por Fase del Modelo [7] *12*
Tabla 3-1: Fases en común-Elaboración propia *35*
Tabla 4-1: Análisis por cohorte según el número de asignaturas aprobadas-PCE-Elaboración propia *39*
Tabla 4-2: Ingresos / Egresos-PCE Elaboración propia *40*
Tabla 4-3: Ingresos/Egresos por cohorte-PCE-Elaboración propia *40*
Tabla 4-4: Índice de abandono por cohorte *41*
Tabla 4-5: Descripción de tablas *42*
Tabla 4-6: Tipo de datos extraídos *42*
Tabla 5-1: Discurso del Cliente - Acta Reunión Nº 1 *48*
Tabla 5-2: Protocolo de Comunicación Externa *49*
Tabla 5-3: Protocolo de Documentación Interna *50*
Tabla 5-4: Protocolo de Comunicación Interna *50*
Tabla 5-5: Conceptos Teóricos Asociados al Dominio *51*
Tabla 5-6: Estudio de la Organización *52*
Tabla 5-7: Plan de Adquisición de Conocimientos *53*
Tabla 5-8: Conocimiento Adquirido *54*
Tabla 5-9: Reporte de Exploración Inicial *55*
Tabla 5-10: Análisis de Recursos Existentes *56*
Tabla 5-11: Reporte de Recursos Externos *56*
Tabla 5-12: Reporte de Recursos Internos *57*
Tabla 5-13: Suposiciones del Proyecto *57*
Tabla 5-14: Reporte de Riesgos del Proyecto *58*
Tabla 5-15: Plan de Contingencia del Proyecto *59*
Tabla 5-16: Preguntas asociadas a la caracterización *59*
Tabla 5-17: Reporte de Viabilidad *60*
Tabla 5-18: Alternativas de Ciclo de Vida *61*
Tabla 5-19: Modelo de Ciclo de Vida *61*
Tabla 5-20: Mapa y Calendario de Actividades *62*
Tabla 5-21: Listado de Métricas *63*
Tabla 5-22: Valores del Factor de Costo OBTY *64*

Tabla 5-23: Valores del Factor de Costo LECO *65*
Tabla 5-24: Valores de Factor de Costo AREP *65*
Tabla 5-25: Valores del Factor de Costo QTUM *65*
Tabla 5-26: Valores del Factor de Costo QTUA *66*
Tabla 5-27: valores del Factor de Costo KLDS *66*
Tabla 5-28: Valores del Factor de Costo KEXT *67*
Tabla 5-29: Valores del Factor de Costo TOOL *67*
Tabla 5-30: Estimación del Proyecto *67*
Tabla 5-31: Reporte de Recursos Requeridos *68*
Tabla 5-32: Plan de Capacitación de RRHH *69*
Tabla 5-33: Contrato del Proyecto *69*
Tabla 5-34: Reporte de Inicio Formal del Ciclo *70*
Tabla 5-35: Control de capacitación de RRHH *70*
Tabla 5-36: Reporte de Aceptación *71*
Tabla 5-37: Glosario de Términos – Definiciones *72*
Tabla 5-38: Objetivos del Negocio *74*
Tabla 5-39: Problemas del Negocio *74*
Tabla 5-40: Expertos en el Problema de Negocio *75*
Tabla 5-41: Criterios de Éxito del Problema de Negocio *76*
Tabla 5-42: Repositorio de Datos *77*
Tabla 5-43: Descripción de Tablas *77*
Tabla 5-44: Campos Asociados al Negocio *78*
Tabla 5-45: Reporte de Datos Explorados *81*
Tabla 5-46: Reporte de Calidad de los Datos *85*
Tabla 5-47: Reporte de Tipos Campos Riesgosos *86*
Tabla 5-48: Reportes de Campos Riesgosos *86*
Tabla 5-49: Problemas de Explotación de la Información *88*
Tabla 5-50: Identificación de la Solución *88*
Tabla 5-51: Reporte de Herramientas Alternativas *89*
Tabla 5-52: Evaluación de Herramientas *90*
Tabla 5-53: Herramienta seleccionada *92*
Tabla 5-54: Algoritmos de MD Soportados *92*

Tabla 5-55: Algoritmos de MD Seleccionados *93*
Tabla 5-56: Reporte de Datos Seleccionados *94*
Tabla 5-57: Descripción de la Fuente Temporaria de Datos *98*
Tabla 5-58: Reporte de Transformación de Datos *98*
Tabla 5-59: Generación del Atributo Target *99*
Tabla 5-60: Reporte de Fuente Temporaria de Datos *100*
Tabla 5-61: Reporte de Limpieza de Datos *102*
Tabla 5-62: Reporte de Datos transformados *103*
Tabla 5-63: Reporte de Configuración de Algoritmos Modeler *103*
Tabla 5-64: Reporte de Configuración de algoritmos RapidMiner *103*
Tabla 5-65: Reporte de Entrenamiento del Algoritmo Selección de Características *104*
Tabla 5-66: Reporte de Entrenamiento del algoritmo Redes Neuronales *108*
Tabla 5-67: Reporte de Entrenamiento del Algoritmo C5.0 - Modeler *111*
Tabla 5-68: Regla para Egresados con IBM SPSS-Modeler *112*
Tabla 5-69: Regla para Posible Desertor con IBM SPSS Modeler *113*
Tabla 5-70: Regla para Rezagados con SPSS-Modeler *113*
Tabla 5-71: Regla para Egresado SPSS Modeler (c/datos académicos) *114*
Tabla 5-72: Regla para Posible Desertor SPSS Modeler (c/datos académicos) *114*
Tabla 5-73: Regla para Rezagado SPSS Modeler (c/datos académicos) *114*
Tabla 5-74: Reporte de Entrenamiento del Algoritmo Rule Induction-RapidMiner *115*
Tabla 5-75: Regla para Egresado con RapidMiner *115*
Tabla 5-76: Regla para Posible Desertor con RapidMiner *116*

NOMENCLATURA

ASUM-DM	Analytics Solutions Unified Method for Data Mining
CARVD	Concepto-Atributo-Relación-Valor del Dominio
CARVEPEI	Concepto-Atributo-Relación-Valor Extendido del PEI
CAV	Concepto-Atributo-Valor
CCD	Concepto-Categoría-Definición
CRD	Concepto-Relación del Dominio
CRISP-DM	Cross Industry Standard Process for Data Mining
CRPEI	Concepto-Relación del Problema de Explotación de la Información
CUE	Código Único de Establecimiento
DM	Data Mining (Minería de Datos)
GR	Atributo Clase de un Grupo
HUM	Facultad de Humanidades
KDD	Knowledge Discovery in Databases (Extracción de Conocimiento)
KPI	Key Performance Indicator (Indicador clave o medidor de desempeño)
MoProPEI	Modelo de Procesos para una Ingeniería de Explotación de la Información
NCR	National Cash Register
PEI	Problema de Explotación de la Información
PCE	Profesorado en Ciencias de la Educación
P^3TQ	Product, Place, Price, Time, Quantity
RN	Redes Neuronales
RSMN	Red Semántica del Modelo del Negocio
RSPEI	Red Semántica del Problema de Explotación de la Información
SAS	Fabricante de Software de Inteligencia Empresarial
SEMMA	Sample, Explore, Modify, Model, Assess

SIU	Sistema de Información Universitaria
SOM	Self-Organizing Maps (Mapas Auto Organizados)
TSDT	Team Data Science Process
SQL	Structured Query Language (Lenguaje de consulta estructurada)
SP	Stored Procedure (Procedimiento almacenado)
TCDD	Término-Categoría-Definición del Dominio
TCDPEI	Término-Categoría-Definición del PEI
TDITT	Top-Down Induction of Decision Trees (Inducción de Arboles de Decisión)
TFM	Trabajo Final de Maestría
UA	Unidad Académica
UNNE	Universidad Nacional del Nordeste

Capítulo 1

Introducción

1 Introducción

1.1 Contexto

La baja proporción de egresos anuales es una situación que preocupa a las autoridades de la Universidad Nacional del Nordeste, como también a los responsables de la gestión de cada Facultad. La mayoría de la Unidades Académicas que integran dicha institución incrementaron el número de inscriptos desde 1983, año en el que se eliminaron los ingresos por cupo. Sin embargo, el ingreso directo a las facultades solo posterga el fracaso, con costos elevados para la Institución y para la sociedad en general. El crecimiento de la matrícula no se vio reflejado en el número de egresados, por el contrario, la brecha existente entre ingresos y egresos anuales aumentó [1].

En la República Argentina, las universidades cuentan desde hace 25 años con el sistema de información universitaria, SIU GUARANI, desarrollado oportunamente por el consorcio SIU, dependiente hoy del CIN (Consejo Interuniversitario Nacional). El mismo registra una abundante y amplia gama de información de la población estudiantil, desde el ingreso y hasta el egreso de los alumnos.

Este Trabajo Final de Maestría (TFM) se basa en la disciplina en auge que se conoce como Ciencia de Datos. La misma tiene como objetivo la búsqueda de patrones interesantes presentes en grandes masas de información, sin necesidad de tener planteada una hipótesis previa. En primer lugar, hay que identificar y familiarizarse con la información con la que se trabajará, como así también con el modelo que se quiere obtener, ya que existe entre ellos una relación directamente proporcional.

1.1. Objetivos del Trabajo Final de Maestría

1.1.1 Objetivo general

El principal objetivo de este trabajo es desarrollar un modelo predictivo que permita detectar situaciones potenciales de deserción o desgranamiento en la población estudiantil de la carrera Profesorado en Ciencias de la Educación de la Facultad de Humanidades de la UNNE, con el fin de diseñar políticas educacionales estratégicas que contribuyan con el incremento de la retención estudiantil.

1.1.2 Objetivos específicos

- Relevar conceptos, técnicas y herramientas vinculadas con la Explotación de Información.
- Analizar las metodologías de Explotación de Información disponibles, seleccionando la más adecuada para el caso de estudio.
- Describir el problema, contextualizando las distintas normativas institucionales vigentes en lo referente a la permanencia del estudiante.
- Desarrollar y validar un modelo predictivo, utilizando la metodología MoProPEI, para identificar factores potenciales de riesgo de abandono de los estudiantes del Profesorado en Ciencias de la Educación.

1.1.3 Estructura del Trabajo

El trabajo consta de 7 capítulos. En el Capítulo 1, además de una breve introducción, se presentan los objetivos del Trabajo Final de Maestría. En el Capítulo 2, se introducen conceptos de Ciencia de Datos, Explotación de Información, Procesos de Descubrimiento de Patrones, Derivación del proceso Explotación de Información, Educción de Requisitos. Se presentan definiciones de deserción, rezago o desgranamiento, conceptos utilizados en el TFM. Además se describen investigaciones relacionadas realizadas en distintas Universidades de América Latina. En el capítulo 3, se presentan las Metodologías de Explotación de Información disponibles, se enuncian las fases comunes de las mismas y se selecciona la metodología a utilizar en el presente TFM. En el capítulo 4, se describe el problema, se presentan las normativas referidas a la permanencia del alumno y se mencionan algunas de las políticas implementadas para paliar la problemática en estudio. En el capítulo 5, se aplican los pasos de la metodología MoProPEI, se desarrolla el modelo predictivo y se realiza la validación del caso de estudio. En el capítulo 6, se exponen los resultados obtenidos. Finalmente, el capítulo 7, presenta las conclusiones del TFM y futuros trabajos comprendidos en esta línea de investigación.

Capítulo 2

Marco Teórico

2 Marco Teórico

En este capítulo se presenta una síntesis del estado del arte relacionado con los temas utilizados para el desarrollo del Trabajo final de Maestría, se introducen conceptos de Ciencia de Datos (sección 2.1), Explotación de Información (sección 2.2), Procesos de Descubrimiento de Patrones (sección 2.3), Derivación del proceso Explotación de Información (sección 2.4), Educción de requisitos (sección 2.5), herramientas de Minería de Datos disponibles (sección 2.6). También se introducen definiciones de deserción (sección 2.7), rezago o desgranamiento (sección 2.8), conceptos utilizados en el TFM, y en la sección 2.9, se describen investigaciones relacionadas, realizadas en distintas Universidades de América Latina.

2.1 Ciencia de Datos

La Ciencia de Datos es considerada actualmente como la principal herramienta para la explotación de datos y la generación de conocimiento. Tiene como objetivo la búsqueda de modelos que describan patrones y comportamientos a partir de los datos con el fin de tomar decisiones o hacer predicciones [2]. Es un área que experimentó un importante crecimiento al extenderse el acceso a grandes volúmenes de datos e incluso su tratamiento en tiempo real, requiriendo de técnicas sofisticadas que puedan tratar con los problemas prácticos como escalabilidad, robustez ante errores, adaptabilidad con modelos dinámicos. Abarca varios grupos de investigación de diferentes áreas, como computación, estadística, matemáticas, ingeniería, que trabajan en la elaboración de nuevos algoritmos, técnicas de computación e infraestructuras para la captura, almacenamiento y procesado de grandes masas de datos [3].

2.2 Explotación de Información

La Explotación de Información es una sub-disciplina de los sistemas de información que brinda a la Inteligencia de Negocios, las herramientas para transformar la información en conocimiento [4]. La misma se define como la búsqueda de patrones interesantes y de regularidades importantes en grandes masas de información.

Cada proceso de Explotación de Información aplica un conjunto de técnicas de Minería de Datos, la mayoría provenientes del campo del aprendizaje [5]. Por ello se concluye que los términos Minería de Datos y Explotación de Información no deben utilizarse para referirse al mismo cuerpo de conocimientos [4], ya que la Minería de Datos se relaciona con los

algoritmos necesarios para transformar los datos en conocimiento mientras que la Explotación de Información lo hace con los procesos y las metodologías propias de la ingeniería que son necesarias para lograr este objetivo. Es por esto que la Minería de Datos se aproxima a la operatoria propia de la programación y la Explotación de Información se acerca más a los procesos de la Ingeniería de Software.

Un proceso de Explotación de Información se define como un grupo de tareas relacionadas lógicamente [5] que, partiendo de un conjunto de información válido para la organización, se ejecuta para lograr otro, con un grado de mayor valor que el inicial. Estas tienen como objetivo la extracción de conocimiento partiendo de la información disponible y dependen en gran medida de la calidad de los datos y de la preparación de los mismos.

Los patrones hallados pueden ser utilizados para predecir situaciones futuras o explicar observaciones pasadas, capacidades fundamentales para mejorar el comportamiento en relación a un fenómeno como es el de la deserción y el desgranamiento universitario.

2.3 Procesos de descubrimiento de patrones

Los procesos de Explotación de Información definen las técnicas o algoritmos a utilizar basándose en las características del problema de explotación. Britos [6] presenta los siguientes procesos de Explotación de Información: descubrimiento de reglas de comportamiento, descubrimiento de grupos, descubrimiento de atributos significativos, descubrimiento de reglas de pertenencia a grupos y ponderación de reglas de comportamiento o de pertenencia.

2.3.1 Descubrimiento de Reglas de Comportamiento

En [7] se define que el proceso de descubrimiento de reglas de comportamiento aplica cuando se requiere identificar cuáles son las condiciones para obtener determinado resultado en el dominio del problema. Para el descubrimiento de reglas de comportamiento definidos a partir de atributos clases en un dominio de problema que representa la masa de información disponible, se propone la utilización de algoritmos de inducción TDIDT. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 2-1 . Como resultado de la aplicación del algoritmo de inducción TDIDT al atributo clase se obtiene un conjunto de reglas que definen el comportamiento de dicha clase.

En primer lugar, se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras) y se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se aplican mapas auto-organizados (SOM). Como resultado de la aplicación de SOM se obtiene una partición del conjunto de registros en distintos grupos a los que se llamará grupos identificados. Para cada grupo identificado se generará el archivo correspondiente.

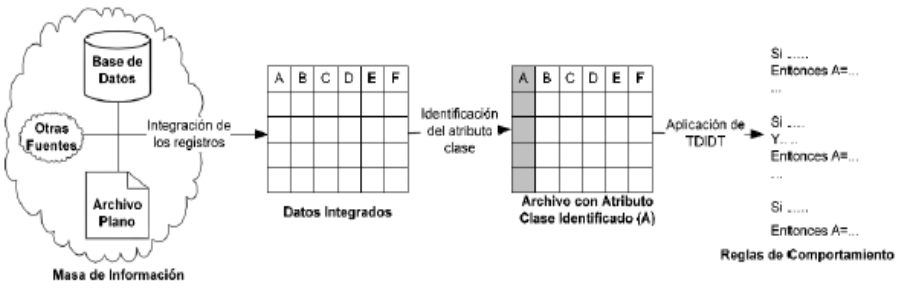


Figura 2-1: PEI-Descubrimiento de Reglas de Comportamiento [7]

2.3.2 Descubrimiento de Grupos

En [7] se define que el proceso de descubrimiento de grupos aplica cuando se requiere identificar una partición en la masa de información disponible sobre el dominio de problema. Para el descubrimiento de grupos a partir de masas de información del dominio de problema sobre las que no se dispone ningún criterio de agrupamiento ‘a priori’ se propone la utilización de Mapas Auto Organizados de Kohonen o SOM por su sigla en inglés. El uso de esta tecnología busca descubrir si existen grupos que permitan una partición representativa del dominio de problema que la masa de información disponible representa. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 2-2.

En primer lugar, se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras) y se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se aplican mapas auto-organizados (SOM). Como resultado de la aplicación de SOM se obtiene una partición del conjunto de registros en distintos grupos a los que se llamará grupos identificados. Para cada grupo identificado se generará el archivo correspondiente.

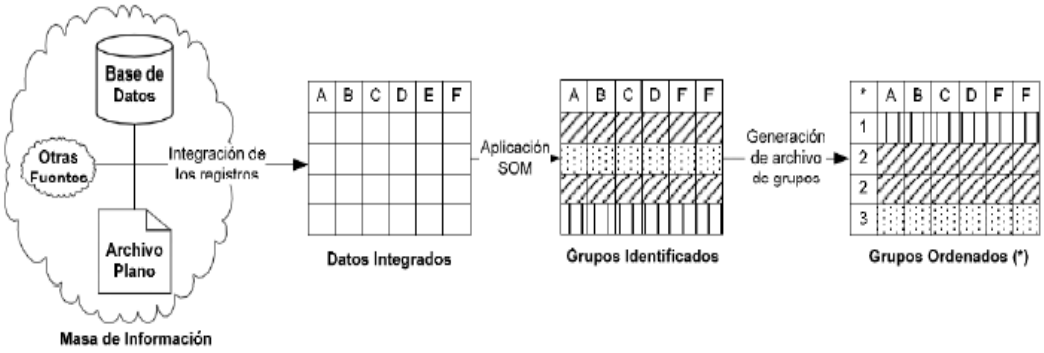


Figura 2-2: PEI-Descubrimiento de Grupos [7]

2.3.3 Ponderación de Interdependencia de atributos

En [7] se define que el proceso de ponderación de interdependencia de atributos aplica cuando se requiere identificar cuáles son los factores con mayor incidencia (o frecuencia de ocurrencia) sobre un determinado resultado del problema. Para ponderar en qué medida la variación de los valores de un atributo incide sobre la variación del valor de un atributo clase se propone la utilización de Redes Bayesianas. El uso de esta tecnología busca identificar si existe algún grado de interdependencia entre los atributos que modelan el dominio de problema que la masa de información disponible representa. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 2-3. En primer lugar, se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras) y se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se selecciona el atributo clase (atributo A en la Figura 2-3).

Como resultado de la aplicación del aprendizaje estructural de las Redes Bayesianas al archivo con atributo clase identificado se obtiene el árbol de aprendizaje. A este se le aplica el aprendizaje predictivo Redes Bayesianas y se obtiene el árbol de ponderación de interdependencias que tiene como raíz al atributo clase y como nodos hojas a los otros atributos con la frecuencia (incidencia) sobre el atributo clase.

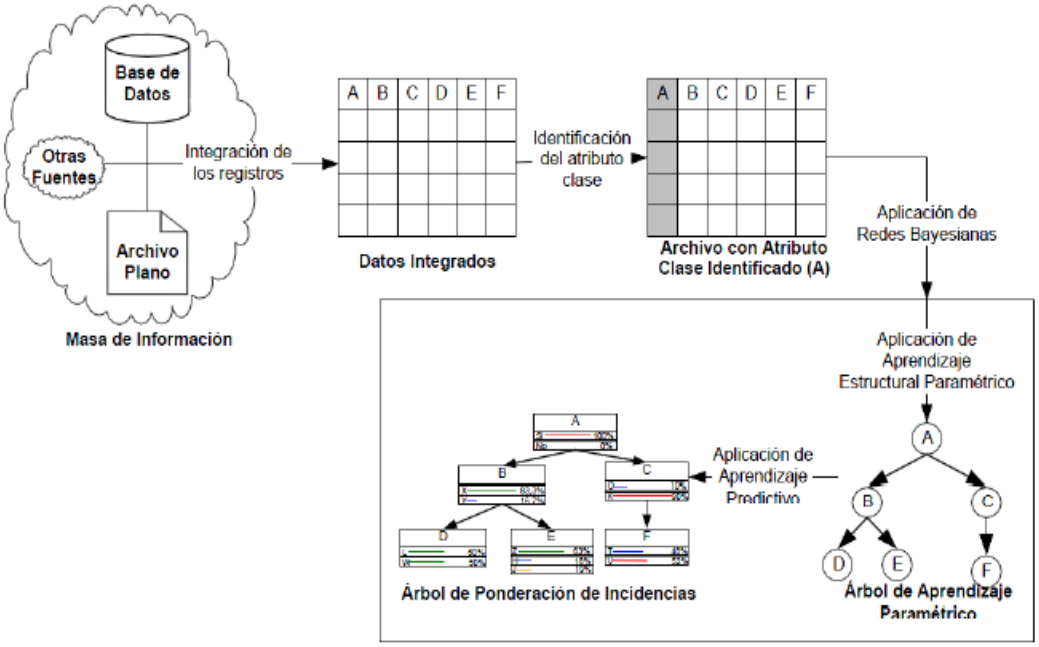


Figura 2-3: PEI-Descubrimiento de Atributos Significativos [7]

2.3.4 Descubrimiento de Reglas de Pertenencia a Grupos

En [7] se define que el proceso de descubrimiento de reglas de pertenencia a grupos aplica cuando se requiere identificar cuáles son las condiciones de pertenencia a cada una de las clases en una partición desconocida 'a priori', pero presente en la masa de información disponible sobre el dominio de problema. Para el descubrimiento de reglas de pertenencia a grupos se propone la utilización de mapas auto-organizados (SOM) para el hallazgo de los mismos y, una vez identificados los grupos, la utilización de algoritmos de inducción (TDIDT) para establecer las reglas de pertenencia a cada uno. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 2-4.

En primer lugar, se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras) y se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se aplican mapas auto-organizados (SOM). Como resultado de la aplicación de SOM se obtiene una partición del conjunto de registros en distintos grupos a los que se llama grupos identificados. Se generan los archivos asociados a cada grupo identificado. A este conjunto de archivos se lo llama grupos ordenados. El atributo grupo de cada grupo ordenado se identifica como el atributo clase de dicho grupo, constituyéndose este en un archivo con atributo clase identificado

(GR). Se aplica el algoritmo de inducción TDIDT al atributo clase de cada grupo GR y se obtiene un conjunto de reglas que definen el comportamiento de cada grupo.

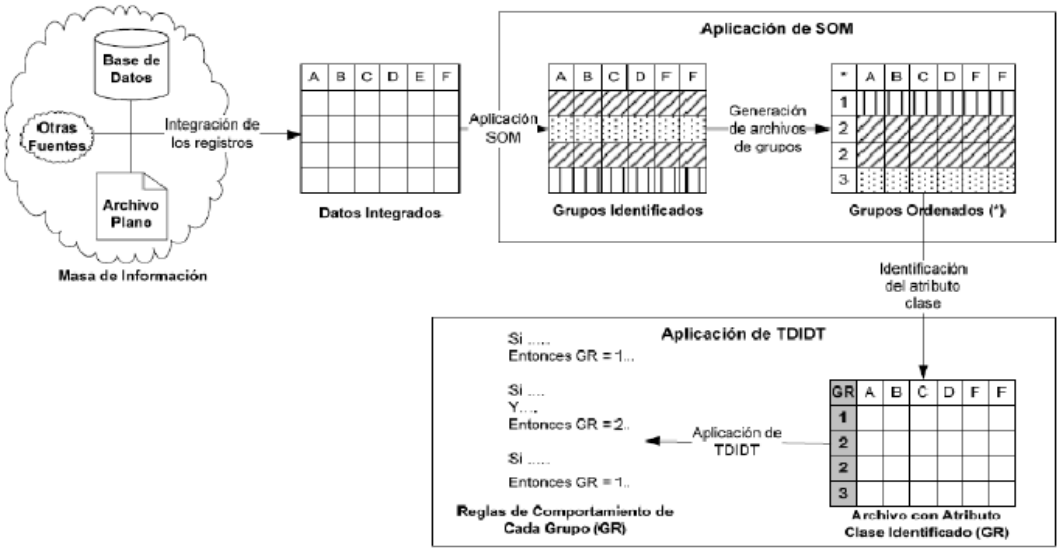


Figura 2-4: PEI-Descubrimiento de Reglas de Pertenencia a Grupos [7]

2.3.5 Ponderación de Reglas de Comportamiento

En [7] se define que el proceso de ponderación de reglas de comportamiento o de pertenencia a grupos aplica cuando se requiere identificar cuáles son las condiciones con mayor incidencia (o frecuencia de ocurrencia) sobre la obtención de un determinado resultado en el dominio del problema, sean estas las que en mayor medida inciden sobre un comportamiento o las que mejor definen la pertenencia a un grupo. Para la ponderación de reglas de comportamiento o de pertenencia a grupos se propone la utilización de Redes Bayesianas. Esto puede hacerse a partir de dos procedimientos dependiendo de las características del problema a resolver: a) cuando no hay clases/grupos identificados, b) cuando hay clases/grupos identificados. El procedimiento a aplicar cuando hay clases/grupos identificados consiste en la utilización de algoritmos de inducción TDIDT para descubrir las reglas de comportamiento de cada atributo clase y posteriormente se utilizan Redes Bayesianas para descubrir cuál de los atributos establecidos como antecedentes de las reglas tiene mayor incidencia sobre el atributo establecido como consecuente. Este proceso y sus subproductos pueden ser visualizados gráficamente en la Figura 2-5.

En primer lugar, se identifican todas las fuentes de información (bases de datos, archivos planos, entre otras) y se integran entre sí formando una sola fuente de información a la que se llamará datos integrados. Con base en los datos integrados se selecciona el atributo clase

(atributo A en la Figura 2-5). Como resultado de la aplicación del algoritmo de inducción TDIDT al atributo clase se obtiene un conjunto de reglas que definen el comportamiento de dicha clase. Seguidamente, se construye un archivo con los atributos antecedentes y consecuentes identificados por la aplicación del algoritmo TDIDT.

Como resultado de la aplicación del aprendizaje estructural de las Redes Bayesianas al archivo con atributo clase obtenido por la utilización del algoritmo TDIDT (en la Figura 2-5), se obtiene el árbol de aprendizaje; a este se le aplica aprendizaje predictivo y se obtiene el árbol de ponderación de interdependencias que tiene como raíz al atributo clase (en este caso el atributo consecuente) y como nodos hojas a los atributos antecedentes con la frecuencia (incidencia) sobre el atributo consecuente.

El procedimiento cuando no hay clases/grupos identificados consiste en aplicar mapas auto-organizados (SOM). Como resultado de la aplicación de SOM se obtiene una partición del conjunto de registros en distintos grupos a los que se llamará grupos identificados. Para cada grupo identificado se generará el archivo correspondiente. A este conjunto de archivos se lo llama grupos ordenados. El atributo grupo de cada grupo ordenado se identifica como el atributo clase de dicho grupo, constituyéndose este en un archivo con atributo clase identificado (GR). Como resultado de la aplicación del aprendizaje estructural se obtiene el árbol de aprendizaje; a este se le aplica el aprendizaje predictivo y se obtiene el árbol de ponderación de interdependencias que tiene como raíz al atributo grupo y como nodos hojas a los otros atributos con la frecuencia (incidencia) sobre el atributo grupo.

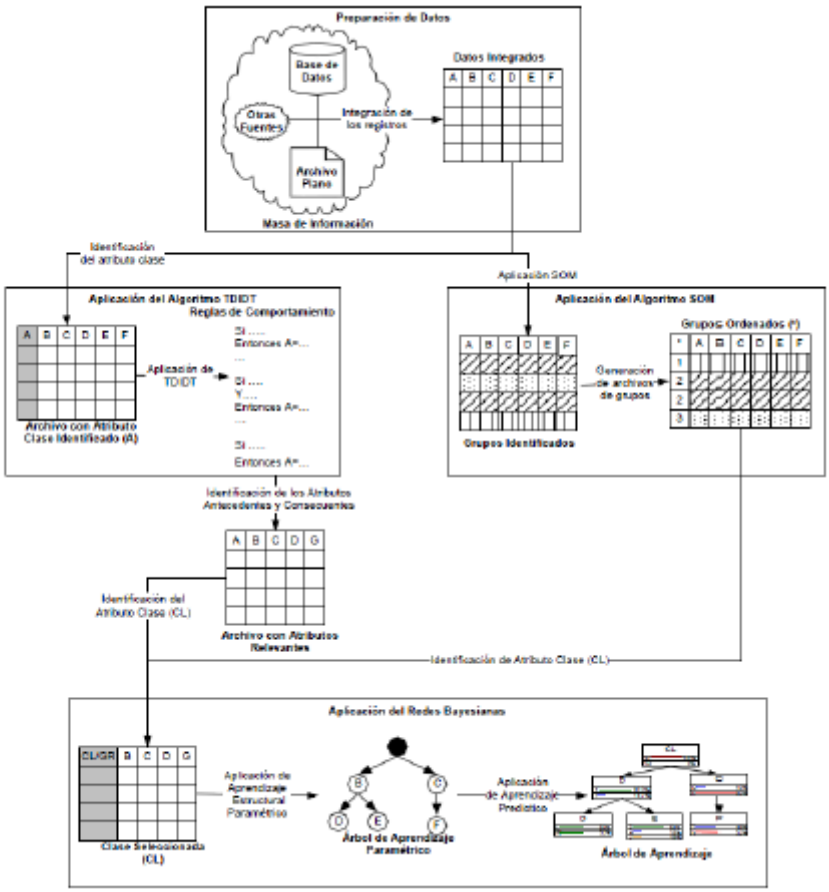


Figura 2-5: PEI-Ponderación de Reglas de Comportamiento o de Pertenencia [7]

2.4 Proceso de Derivación de Modelos

En [8] se describen los tiempos requeridos por cada fase de un proyecto y se detallan las fases y sub-fases que requieren más tiempo. Cabe destacar que entre las fases de Comprensión del Negocio y de Modelado se invierte más del 50% de la duración del proyecto. Además, las sub-fases Determinar los Objetivos de Negocio y Evaluar la Situación, que integran la fase de Comprensión del Negocio, utilizan más del 70% del tiempo pautado. Por otro lado, en la fase de Modelado, la sub-fase Construcción del Modelo requiere el 62,97% del plazo concedido a la totalidad de la ejecución. En la Tabla 2-1, se presenta el porcentaje de tiempo que cada fase insume:

Tabla 2-1: Carga de Trabajo por Fase del Modelo [8]

Fase	% Tiempo
Comprensión del negocio	20,70
Entendimiento de los datos	10,90
Preparación de los datos	15,61
Modelado	34,41

Evaluación	7,45
Implantación	10,93

Martins en [9] destaca que ninguna de las metodologías para proyectos de Explotación de Información disponibles ofrece un método que permita definir de manera estándar el proceso de Explotación de la Información a aplicar en base al dominio del negocio y al problema identificado. Por consiguiente, cada ingeniero debe representar los conocimientos de la forma que considere oportuna.

Por ello, propone y desarrolla un modelo denominado **Proceso de derivación de modelos**, cuya función es la de actuar como enlace entre las fases de Comprensión del Negocio y determinación del Problema de Explotación de la Información (PEI) y la fase de definición del Proceso de Explotación a utilizar. La Figura 2-6 ilustra este concepto.

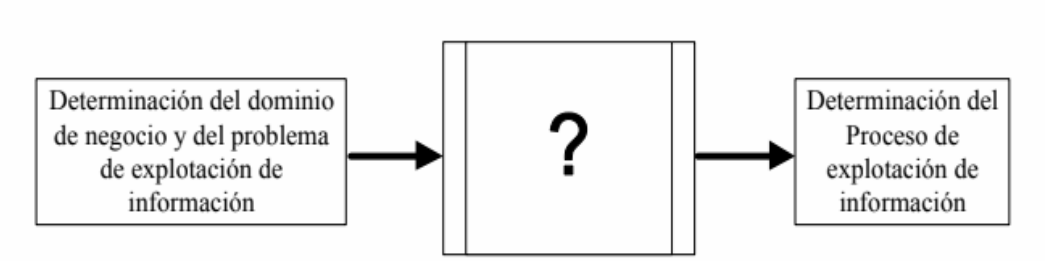


Figura 2-6: Representación Gráfica de la Inserción del proceso de Derivación de Modelos [9]

El proceso de derivación de modelos toma como elementos de partida la descripción del dominio del negocio, dentro de la que se encuentran definidos los datos del negocio, y la descripción del problema de Explotación de Información junto con la descripción de sus datos. Proporciona como salida el proceso de Explotación de Información a aplicar para dar solución al problema de negocio representado mediante el problema de Explotación de Información. El mismo está conformado por tres fases, de las cuales las primeras dos poseen tres etapas y la última está integrada por una solo etapa.

Las tres etapas que conforman la primera fase están orientadas al modelado del dominio del negocio cuyo resultado final es la Red Semántica del Modelo de Negocio, que brinda una representación de los elementos que componen el dominio del negocio y la interacción entre los mismos.

La segunda fase, orientada al modelado del problema de Explotación de Información, está compuesta por tres actividades cuyo resultado final es la Red Semántica del Problema de

Explotación de Información, que brinda una representación de los elementos que componen el dominio del negocio y la interacción entre los mismos. La tercera fase, orientada a determinar el proceso de Explotación de Información, está conformada por una actividad que se compone de ocho sub-pasos, y su salida es el proceso de Explotación de Información a aplicar, es decir, el resultado final esperado por el proceso de derivación de modelos.

La distribución de las fases no es arbitraria, sino que las actividades que las conforman definen el orden en el que se deben realizar. Cada paso o etapa del proceso tiene elementos de entrada, los cuales se procesan generando elementos de salida, siendo estos a su vez elementos de entrada de actividades posteriores.

Las fases del proceso son:

1. **Fase orientada al análisis del dominio del negocio**: su objetivo es el análisis, la comprensión y la conceptualización del modelo de negocio de la organización. Esta fase se realizará por única vez para los distintos problemas de negocio que la organización desea enfrentar. La misma está compuesta por los siguientes pasos:

    1.1. **Identificación de los elementos y estructura del dominio:** tiene como elementos de entrada la descripción del dominio del negocio y los datos del negocio y produce como salida la tabla Concepto-Atributo-Relación-Valor del Dominio (CARVD) y la tabla Término-Categoría-Definición del Dominio (TCDD). Este paso tiene como objetivo generar, a partir de la descripción del dominio de negocio y de los datos del negocio, una estructura que defina los elementos que integran el modelo de negocio, en la que se identifiquen los conceptos que forman parte del mismo junto con las características que lo definen. En base a lo establecido previamente, es primordial definir los siguientes términos: ‘conceptos’ son aquellos elementos que poseen características que permiten distinguirse de otros, que pueden interactuar con otros elementos, o pueden ser agrupados o generalizados. Las características de los conceptos se denominan ‘atributos’, los que pueden tener distintos ‘valores’, entendiéndose por estos el estado en el que se encuentra un atributo, y los que definen los distintos estados posibles que un atributo puede tener. Las ‘relaciones’ son las interacciones entre los distintos

elementos. Las técnicas utilizadas son: tabla Término-Categoría-Definición del Dominio (TCDD) y la tabla Concepto-Atributo-Relación-Valor del Dominio (CARVD) (Anexo).

1.2. El segundo paso denominado **Identificación de relaciones entre conceptos del dominio**: tiene como entrada las tablas de salida del paso anterior, (TCDD - CARVD) y como elemento resultado a la tabla Concepto-Relación del Dominio (CRD). El objetivo de este paso es detectar, a partir de los conceptos identificados en el paso anterior, las relaciones existentes en el modelo del negocio estableciendo los vínculos y modos de interacción que tienen los distintos elementos que componen al dominio del negocio. La técnica utilizada en este paso es la tabla Concepto-Relación del Dominio, (CRD) (Anexo).

1.3. El tercer paso se denomina **Conceptualización del Dominio:** utiliza como entrada los elementos de salida del paso anterior, (CARVD-CRD), y se obtiene un modelo que representa gráficamente toda la información relevante del dominio del negocio enfocado en las relaciones entre los elementos que componen al negocio. Se implementa la Técnica Red Semántica del Modelo de Negocio, (RSMN) (Anexo).

2. **Fase orientada al análisis del problema de Explotación de Información:** tiene como objetivo el análisis, la comprensión y el modelado del problema de Explotación de Información. Deberá realizarse por cada problema de negocio que se quiera afrontar. Los pasos que conforman esta fase son:

2.1. **Identificación de los elementos y estructura del Problema de Explotación de Información:** su objetivo es detectar, a partir de la descripción del problema de Explotación de Información, los distintos elementos que lo componen. A partir de los elementos detectados, junto con la descripción de los datos del problema de Explotación de Información, se define la estructura de los conceptos identificados, se determinan las relaciones existentes entre conceptos y se define el flujo

del problema de Explotación de Información mediante los elementos de entrada y salida del mismo. Los elementos de entrada de dicho paso son la descripción del problema de Explotación de Información y los datos del problema de Explotación de Información. Se aplican las técnicas, tabla Término-Categoría-Definición del Problema de Explotación de Información (TCDPEI) y la tabla Concepto-Atributo-Relación-Valor Extendida del Problema de Explotación de Información (CARVEPEI) (Anexo).

2.2. **Identificación de relaciones entre conceptos del Problema de Explotación de Información:** el objetivo es detectar, a partir de los conceptos identificados en el paso anterior, las relaciones existentes en el problema de Explotación de Información entre dichos conceptos estableciendo los vínculos y modos de interacción que tienen los distintos elementos que componen el problema. Los elementos de entrada de dicho paso son las tablas TCDPEI y CARVEPEI generadas en el paso anterior. Para la aplicación de este paso, se utiliza la técnica tabla Concepto-Relación del Problema del Explotación de Información, (CRPEI) (Anexo).

2.3. El tercer paso se denomina **Conceptualización del Problema de Explotación de Información:** como resultado se obtiene un modelo que representa gráficamente toda la información relevante del problema de Explotación de Información, enfocado en las relaciones entre los elementos que lo componen. Los elementos de entrada de este paso son las tablas CARVEPEI y CRPEI, esta última obtenida en el paso anterior. Para su aplicación se implementa la Técnica Red Semántica del Problema de Explotación de Información, (RSPEI) (Anexo).

3. **Fase orientada a determinar el proceso de Explotación de Información:** su objetivo es identificar el proceso de Explotación de Información a aplicar de acuerdo al dominio del negocio y al problema de Explotación de Información, expresados mediante las representaciones intermedias obtenidas en las fases

previas. Esta fase deberá realizarse por cada problema de Explotación de Información a analizar y está compuesta por el siguiente paso:

2.1. **Derivación del proceso de Explotación de Información:** su finalidad es brindar un mecanismo que permita deducir el proceso de Explotación de Información a aplicar para dar solución al problema definido. Los elementos de entrada son los resultados finales de cada fase previa la RSMN y la RSPEI. Para su aplicación se implementa la Técnica Algoritmo de Derivación del Proceso de Explotación (Anexo).

2.5 Educción de requisitos

En [6] se señala la necesidad de adaptar el proceso tradicional de especificación de requerimientos de sistemas software para proyectos de Explotación de Información y en [10] se ofrece la solución a dicho problema, con el desarrollo de un proceso de educción de requisitos que contempla la elaboración de un conjunto de plantillas que son utilizadas para documentar los conceptos educidos durante las fases de comprensión del negocio y del dominio de proyecto. Cada plantilla se asocia a un concepto educido y contiene una descripción detallada del mismo, lo cual posibilita la evolución del concepto a lo largo del proyecto. De esta forma, a partir de la información contenida en las plantillas, el profesional de Ciencia de Datos puede realizar las actividades del proyecto, tal como identificar los repositorios de datos, determinar el modelo y seleccionar las herramientas a utilizar dentro del proyecto.

2.6 Herramientas de Minería de Datos disponibles

Se presentan a continuación algunas de las herramientas de Minería de Datos disponibles:

WEKA, desarrollado en java. Se ejecuta bajo la Licencia Pública General de GNU y contiene herramientas para la aplicación de diferentes técnicas de Minería de Datos como la clasificación, la regresión, el agrupamiento entre otras [11].

KNIME, software de código abierto para crear aplicaciones y servicios de Ciencia de Datos. Es intuitivo, abierto e integra continuamente nuevos desarrollos. Esto hace que la comprensión de datos y el diseño de flujos de trabajo de Ciencia de Datos y componentes reutilizables sean accesibles para todos [12].

RapidMiner, software de código abierto, desarrollado inicialmente por la Universidad de Dortmund [13].

IBM SPSS Modeler, software de Minería de Datos de IBM, originalmente Clementine, desarrollado por Integral Solutions Limitad. Se utiliza para construir modelos predictivos y realizar otras tareas analíticas. Tiene una interfaz visual que permite a los usuarios aprovechar los algoritmos estadísticos y de Minería de Datos sin programación [14].

2.7 Deserción

Son varias las definiciones de indicadores de deserción, Tinto [15] plantea : *"El estudio de la deserción de la educación superior es extremadamente complejo, pues implica no solo una variedad de perspectivas sino también una amplia gama de diferentes tipos de abandono. Probablemente ninguna definición pueda captar en su totalidad la complejidad de este fenómeno universitario."* También indica que existe una gran variedad de comportamientos denominados con el rótulo común de deserción, mas no debe definirse con este término a todos los abandonos de estudios ni todos ellos merecen intervención institucional. Añade que *"solo algunos de los abandonos de la educación superior son producidos por bajo desempeño académico pues la mayor parte de las deserciones son voluntarias. Los estudiantes que abandonan la universidad a menudo tienen niveles de rendimiento superiores a los que persisten."* Tinto [15] considera 'desertor' al estudiante que no presenta actividad académica durante tres semestres académicos consecutivos.

Himmel, sin embargo, en [16] define la deserción como *"el abandono prematuro de un programa antes de alcanzar el título o grado y comprende un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante se reincorpore"* y señala al igual que Berger [17] que los enfoques del análisis de la deserción y retención pueden ser agrupados en cinco grandes categorías: psicológicos, sociológicos, económicos, organizacionales y de interacción.

Ante esta pluralidad conceptual, para iniciar cualquier análisis predictivo que impacte en la definición de políticas institucionales de retención resultaría conveniente que las universidades definan cuáles son los tipos de abandono que consideran deserciones y cuáles podrían intervenirse para aumentar la retención.

2.8 Desgranamiento o Rezago

En [18] se define el desgranamiento o rezago como la apreciación de la pérdida de matrícula que ocurre en el transcurso de una cohorte. En la actualidad se habla de persistencia o prolongación de los estudios. Una de las causas de rezago en el sistema educativo es la repitencia, la que se entiende como la acción de cursar reiterativamente una actividad académica, sea por mal rendimiento del estudiante o por causas ajenas al ámbito académico [19].

En el caso de las instituciones universitarias, la repitencia se expresa como el recursado de asignaturas o reprobación de finales, mientras que en los procesos de evaluación para la mejora de las carreras resulta un buen indicador de rezago la acumulación de finales sin rendir una vez aprobada la cursada. Este indicador resulta valioso en la identificación de estudiantes en riesgo académico y/o de deserción, lo cual posibilita la toma de medidas de intervención preventiva.

González Fiegehen en [19] manifiesta que el alumno que deserta se siente abandonado por la institución, se inicia así una ruptura espaciotemporal dentro del aula y la relación con los compañeros y docentes se hace cada vez más distante. El alumno que se desgrana de su cohorte es un potencial desertor y se puede considerar como un indicador de rendimiento que permite medir la eficiencia de la institución.

2.9 Investigaciones de Deserción y Rezago en la Educación Superior, utilizando herramientas de Explotación de la Información

En la Argentina, a partir del año 1918, se presentan una serie de transformaciones de mucha importancia para el sistema de Educación Superior, se introducen valores como la libertad, periodicidad de las cátedras, concurso para profesores, autarquía, cogobierno, entre otros. Más adelante, en 1949, se incorpora la gratuidad universitaria.

En gobiernos democráticos de la década del 80 se modifican las restricciones de ingreso, se eliminan los cupos, se incorpora el ingreso al sistema de personas mayores de 25 años que no culminaron los estudios secundarios, se incorpora la asistencia económica a los alumnos carenciados a través de becas. Estas políticas universitarias llevaron a que la matrícula de las instituciones públicas de Educación Superior aumentara considerablemente, pero estas altas tasas de cobertura del sistema universitario no se reflejan en las tasas de graduación, los índices de deserción continúan siendo elevados durante el primer año de las carreras [20].

Alrededor de un 40 % de los estudiantes abandonan su carrera en primer año; un porcentaje menor pero todavía importante lo hacen en el segundo año. Algunos de estos estudiantes cambian de carrera, pero la mayoría de ellos abandona los estudios. También se demostró que la deserción impacta mayoritariamente en el 40% de jóvenes con menor ingreso per cápita familiar [20].

En el año 2017, según estadísticas de SPU, se registraron 1.584.392 alumnos matriculados en universidades del sector público y 86.174 graduados [21]. En la UNNE, según datos suministrados por el Departamento de Estadísticas de la misma, se registraron 51.629 alumnos matriculados y 3.154 egresados, es decir que de cada 100 alumnos matriculados solo se gradúan 6 [1].

El problema de la deserción preocupa a la Universidades Argentinas y latinoamericanas entre otras. Así es que se vienen realizando una serie de estudios empíricos con el objetivo de encontrar patrones de comportamiento en forma automática utilizando las bases de datos de los sistemas de gestión académica de las Universidades [22]. La Universidad Nacional de Misiones-UNAM-Argentina, utilizando la información recabada por el Sistema de Información Universitaria provisto por el consorcio SIU y valiéndose del uso de algoritmos TDIDT, ha realizado investigaciones con el fin de identificar variables que inciden en la deserción [23]. En el ámbito de la Universidad Nacional de Río Negro (UNRN), y en particular en la Sede Atlántica desde la Licenciatura en Sistemas, se desarrolló el trabajo en el que se describe el proceso de identificación de las características más relevantes del problema a través de las cuales, utilizando técnicas de Minería de Datos, puede obtenerse un modelo de la deserción universitaria [24]. La UNNOBA elaboró un trabajo muy interesante, en el cual se aplicaron técnicas de Explotación de la Información utilizando datos del SIU para detectar alumnos en riesgo de abandono. Además, desarrolló un tablero de control que permite a los docentes que no dominan las técnicas de DM visualizar datos de la situación de las distintas cohortes y de cada alumno en detalle [22].

En el año 2015 en la Universidad Gastón Dachary, utilizando algoritmos de clasificación, se desarrolla el trabajo denominado Análisis de deserción-permanencia de estudiantes universitarios, a fin de identificar atributos que caracterizan a los casos de deserción [25].

En la Universidad Nacional de Catamarca se elabora el proyecto, Minería de datos para un sistema de alerta temprana de deserción en carreras de Ingeniería, en el que aplicando

técnicas de agrupamiento, arboles de decisión, reglas de asociación y clasificación, y máquinas de vectores soporte, se desarrolla un modelo con el que se caracteriza la trayectoria académica del alumno a fin de detectar patrones compatibles con situaciones de dificultades en el aprendizaje [26].

La Universidad Nacional de La Matanza (UNLaM), presenta en el año 2010, los resultados preliminares del trabajo denominado: Abandono y egresos en las carreras de Ingeniería. El mismo tiene como objetivo identificar, mediante el uso de técnicas de minería de datos, factores que inciden en la deserción de los alumnos de las carreras de Ingeniería de la UNLaM [27].

En otros países latinoamericanos también se han desarrollado varias investigaciones para entender el problema de la deserción en la educación superior. En Chile en el año 2018 se elaboró un estudio con el objetivo de presentar una clasificación basada en árboles de decisión con parámetros optimizados para predecir la deserción de los estudiantes universitarios [28]. En la Universidad Simón Bolívar, Barranquilla, Colombia, se realizó un trabajo, en el que también se optó por la inducción de árboles de decisión porque, además de ser la más común dentro de las técnicas de clasificación de datos, representa una gran ventaja con respecto a las demás técnicas de clasificación debido a que se puede representar el conocimiento extraído en un conjunto de reglas de decisión de fácil entendimiento [29].

Sobre esta base de importantes conceptualizaciones que sustentan teóricamente el TFM, a continuación de describen las metodologías de Explotación de Información disponibles.

Capítulo 3

Metodologías de Explotación de Información

3 Metodologías de Explotación de Información

Las metodologías de Explotación de Información son herramientas que permiten llevar a cabo el proceso de Minería de Datos en forma sistemática y no trivial. Estas definen las fases del proceso y además describen las tareas a realizarse y el modo de llevarlas a cabo. Entre las metodologías disponibles que se pueden aplicar en la ejecución de los proyectos de Ciencia de Datos se encuentran: KDD (3.1), CRISP-DM (3.2), P^3TQ (3.3), SEMMA (3.4), MoProPEI (3.5), ASUM-DM (3.6), TDSP (3.7).

3.1 KDD

En el año 1996, el modelo KDD fue el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de Explotación de Información [30]. Esta metodología está integrada por cinco fases Figura 3-1Figura **3-1**.

Integración y recopilación: consiste en establecer un entendimiento del dominio de la aplicación y de los conocimientos previos relevantes. En esta fase se determina la selección de un conjunto de datos que pueden ser obtenidos de diferentes fuentes, sobre los cuales se realiza el descubrimiento.

Selección, limpieza y transformación: en esta etapa se seleccionan y preparan los datos que se van a utilizar. Sin embargo, existen factores como el ruido o valores atípicos que afectan la calidad de los datos, ante lo cual la limpieza es una de las tareas más importantes, ya que permite la selección de la técnica que más se ajuste al problema a resolver.

Minería de datos: es la fase más representativa, en la que se determina qué tipo de proceso es el más apropiado, ya sea agrupamiento, reglas de asociación, correlación, clasificación, regresión, entre otros. Los resultados obtenidos dependen de las fases anteriores, por lo que existe la posibilidad de regresar a los pasos previos para obtener nuevos datos o para redefinir la solución al problema planteado.

Evaluación e interpretación: los patrones descubiertos deben ser precisos, comprensibles e interesantes. En esta fase se evalúan e interpretan los patrones obtenidos. Algunas validaciones pueden ser a través de índices de evaluación, validación cruzada, matrices de confusión, entre otras.

Difusión y uso: como última fase, el conocimiento descubierto debe de ser incorporado en algún sistema o simplemente documentarlo para su difusión a las partes interesadas. Este

proceso incluye también la revisión y resolución de posibles conflictos con los conocimientos previos.

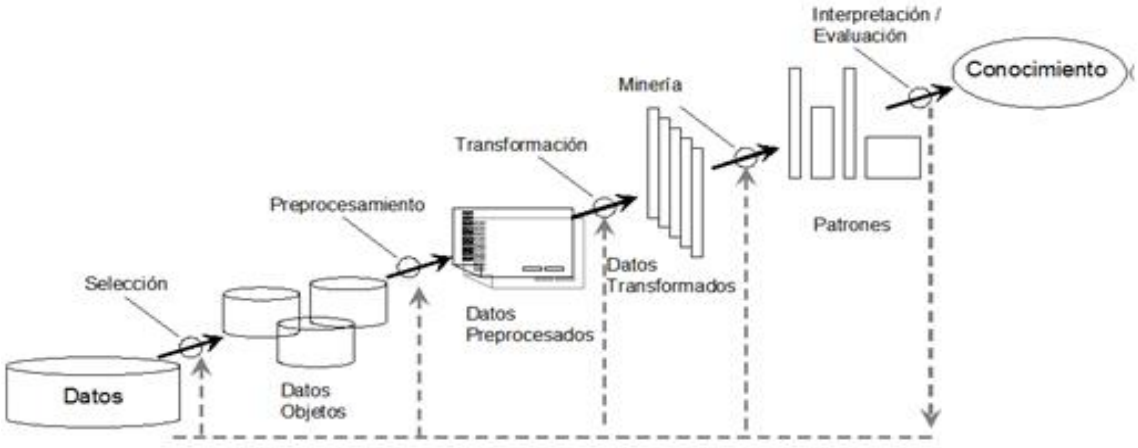


Figura 3-1: Etapas del Proceso de Extracción del Conocimiento [30]

3.2 CRISP-DM

Los orígenes de CRISP-DM se remontan al año 1999 cuando un importante consorcio de empresas integrado por SPSS, NCR y Daimler Chrysler entre otras, propone, a partir de diferentes versiones de KDD, el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM. Esta tiene como objetivo el desarrollo de proyectos a partir de un proceso estandarizado (independiente de la industria y la herramienta) que permite minimizar los costos que implica un proyecto de este tipo en una organización [31].

La metodología está definida en base a un modelo jerárquico de procesos. Se mantendrá el foco en los procesos del nivel superior que son lo suficientemente genéricos como para cubrir todas las posibles aplicaciones de Explotación de Información. La misma define un ciclo de vida de los proyectos de Explotación de Información, establece las principales fases de un proyecto de este tipo junto con sus relaciones, como puede observarse en la Figura 3-2. Estas relaciones son las más comunes, aunque pueden establecerse entre cualquiera de las etapas. Estas fases difieren de las definidas para un proyecto de desarrollo de software clásico (inicio, requerimientos, análisis y diseño, construcción, integración y pruebas y cierre).

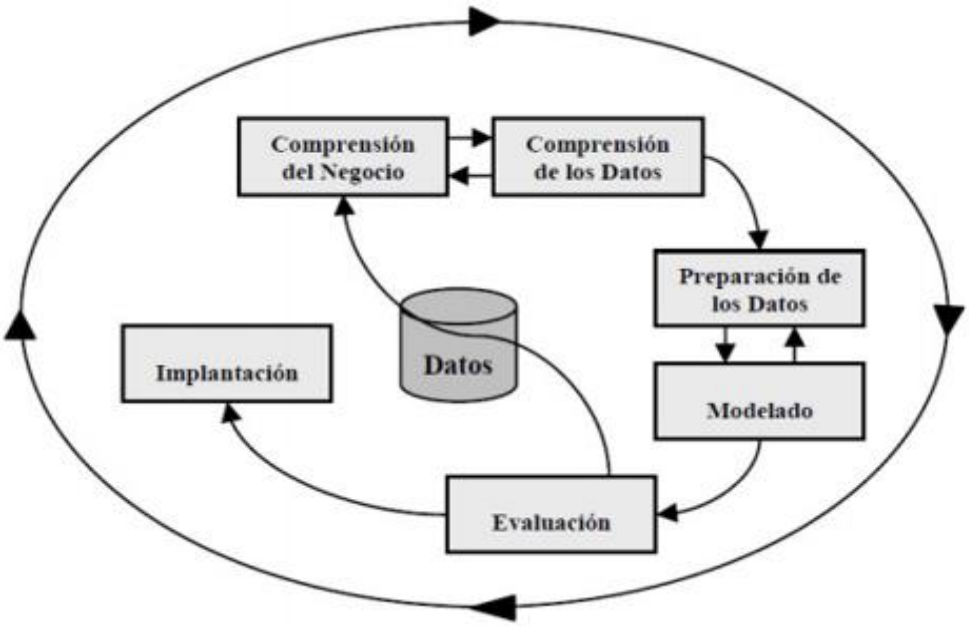


Figura 3-2: Fases del Modelo CRISP DM [31]

Las fases que comprende CRISP DM son:

Comprensión del Negocio: en esta de deben entender los objetivos del proyecto y los requerimientos desde una perspectiva del negocio y luego convertir este conocimiento en una definición de un problema de Explotación de Información y diseñar un plan preliminar para lograr dichos objetivos.

Comprensión de los Datos: comienza con la recolección inicial de datos y procede con las acciones para familiarizarse con ellos, identificar problemas de calidad, identificar primeras pautas en los datos o detectar subconjuntos interesantes de las hipótesis de información oculta.

Preparación de los Datos: cubre todas las actividades para construir el conjunto de datos final desde los datos iniciales, las tareas de esta fase pueden ser realizadas muchas veces y sin un orden reestablecido. Incluye tanto la selección de tablas, registros y atributos como transformación y limpieza de datos para herramientas de modelado.

Modelado: comprende la selección de técnicas de modelado y la calibración de sus parámetros a los valores óptimos. Existen distintas técnicas para un mismo problema de Explotación de Información y cada una de ellas suele tener ciertos requisitos sobre los datos. Con frecuencia es necesario volver a la fase de preparación de los datos.

Evaluación: durante esta fase se construyen varios modelos buscando aquellos que contribuyan a una mayor calidad de análisis. A fines de asegurar el logro de los objetivos de

negocio, se requiere la evaluación de cada modelo y la revisión de los pasos ejecutados para la construcción del mismo.

Despliegue: puede ser tan simple como generar un reporte o tan compleja como implementar un proceso de Explotación de Información repetible a través de toda la empresa.

En la Figura 3-3, se observan en el orden secuencial y natural las distintas fases que componen la metodología CRISP-DM.

Figura 3-3: Fases Componentes de la Metodología CRISP DM [31]

Cada una de estas se divide en varias fases de nivel inferior que indican tareas generales a realizar dentro de la misma. A su vez, estas tareas de segundo nivel son divididas en tareas específicas en la que se describen las acciones que deben ser desarrolladas en situaciones concretas. En un cuarto nivel se recogen acciones, decisiones y resultados sobre el proyecto de Explotación de Información [6]. Esta abstracción de procesos puede verse gráficamente en la Figura 3-4.

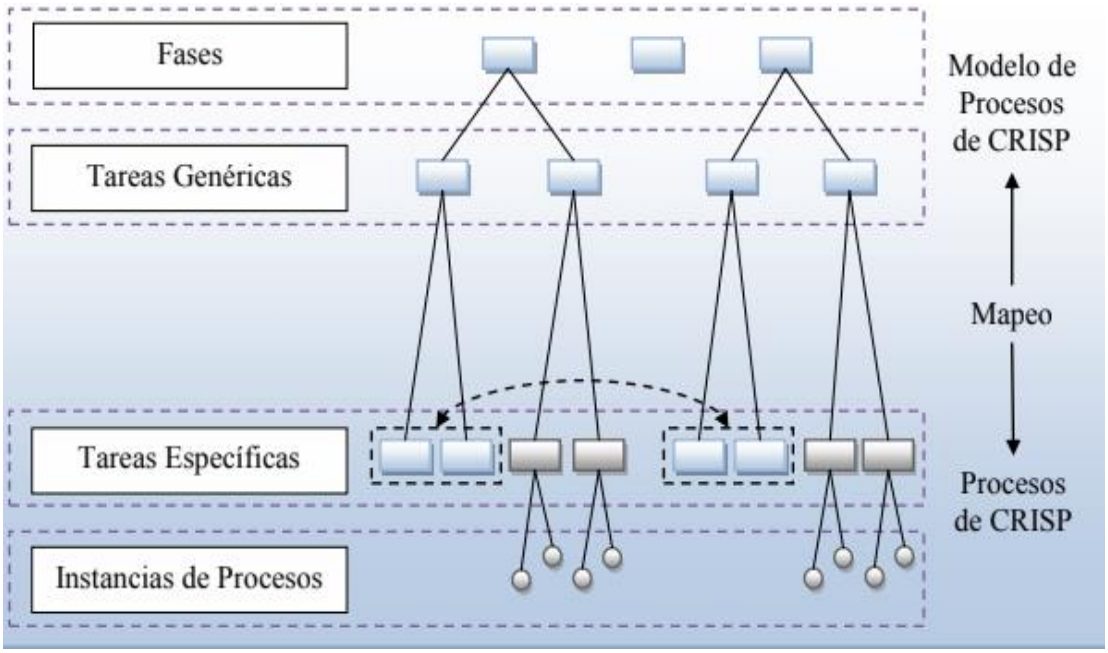


Figura 3-4: Niveles de abstracción de procesos de CRISP-DM [6]

3.3 P^{3}TQ

La metodología Catalyst, conocida como P^{3}TQ surge en el año 2003. Esta plantea la formulación de dos modelos: el Modelo de Negocio (MII) y el Modelo de Explotación de Información (MIII). El Modelo de Negocio proporciona una serie de pasos para identificar un problema y los requerimientos reales de la organización, Figura 3-5. Para proyectos donde el problema no está definido, se recomienda comenzar analizando las relaciones P^{3}TQ que existen en la cadena de valor organizacional, es decir, aquellas relaciones precio/lugar/producto /tiempo/cantidad que son importantes para la empresa. El Modelo de Explotación de Información proporciona una guía de pasos para la construcción y ejecución de modelos de Minería de Datos a partir del Modelo de Negocio. El foco que propone la metodología Catalyst en su Modelo de Negocio sobre la cadena de valor organizacional hizo que sea difundida en la comunidad científica como metodología P^{3}TQ, aunque esta no sea su denominación original [31]. La metodología Catalyst, en sus dos modelos, está compuesta por una serie de pasos denominados boxes/cajas. El concepto es que, luego de llevar a cabo una acción, se deben evaluar los resultados y determinar cuál es el próximo paso (box/caja) a seguir. La secuencia y la interacción entre los distintos pasos permiten una flexibilidad muy grande y una amplia variedad de caminos posibles.

El Modelo de Negocio plantea cinco escenarios diferentes de acuerdo con las circunstancias del negocio [6]:

1. **Dato:** el proyecto comienza con un conjunto de datos a ser explorados para encontrar patrones de interés.
2. **Oportunidad:** el proyecto inicia como un problema u oportunidad de negocio que debe ser explorado.
3. **Prospectiva:** el objetivo del proyecto es descubrir dónde la Minería de Datos puede ofrecer un valor a la organización.
4. **Definido:** el proyecto comienza con la premisa de crear la especificación del modelo de Minería de Datos con un propósito determinado.
5. **Estratégico:** el proyecto comienza con una estrategia de análisis para dar soporte a un escenario planificado por la organización.

Por su parte, el Modelo de Explotación de Información proporciona una guía de referencia para la Explotación de Información mediante una serie de pasos [32]:

1. **Preparación de los datos:** incluye una serie de actividades que permiten comprobar la calidad de los datos a utilizar. Se revisan las características de las variables, así como el tamaño de los datos, entre otros.
2. **Selección de herramientas y modelado inicial:** permite seleccionar la herramienta y el modelo con base al análisis del problema, por ejemplo, si se van a predecir los datos es necesario conocer el tipo de tarea predictiva que más se ajuste.
3. **Refinar el modelo seleccionado**.
4. **Implementar el modelo**.
5. **Comunicación de resultados:** presentar el resultado obtenido al público interesado y a los responsables de la toma de decisiones.

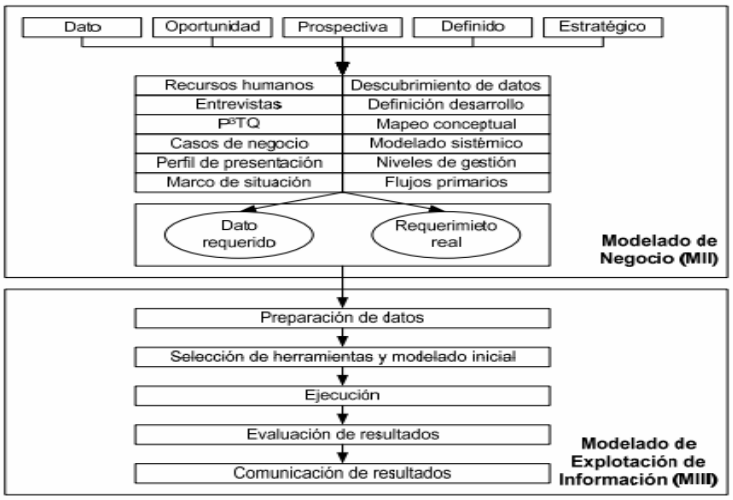


Figura 3-5: P3TQ [32]

3.4 SEMMA

SEMMA, desarrollada por el SAS Institute y propuesta para trabajar con el software de dicha compañía, se define como el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos. El nombre de esta metodología es el acrónimo correspondiente a las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado), Assess

(Valoración) [31] . Esta metodología organiza sus herramientas (llamadas 'nodos') en base a las distintas fases que la componen. Es decir, el software proporciona un conjunto de herramientas especiales para la etapa de muestreo, otras para la etapa de exploración, y así sucesivamente. Sin embargo, el usuario podría hacer uso del mismo siguiendo cualquier otra metodología (CRISP-DM por ejemplo).

Figura 3-6: Fases Metodología SEMMA [31]

El proceso se inicia con la extracción de la muestra sobre la que se aplicará el análisis. El objetivo de esta fase consiste en seleccionar un conjunto de datos representativo del problema en estudio. La representatividad de dichos datos es indispensable ya que de no cumplirse invalida todo el modelo y los resultados dejan de ser admisibles [6]. La forma más común de obtener la muestra es a través del muestreo simple o aleatorio, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido.

La metodología SEMMA establece que se debe asociar un nivel de confianza a cada muestra considerada para el análisis del proceso. Una vez determinada la muestra o conjunto de muestras representativas de la población en estudio, la metodología SEMMA indica que se debe proceder a una exploración de la información disponible con el fin de simplificar en lo posible el problema para optimizar la eficiencia del modelo. Para lograr este objetivo se propone la utilización de herramientas de visualización o de técnicas estadísticas que ayuden a poner de manifiesto relaciones entre variables. De esta forma se pretende determinar cuáles son las variables explicativas que serán útiles como entradas al modelo [6].

La tercera fase de la metodología consiste en la manipulación de los datos. En base a la exploración realizada, deben obtenerse datos que tengan el formato adecuado para ser introducidos en el modelo. Una vez que se han definido las entradas del modelo, con el formato correcto para la aplicación de la técnica de modelado, se procede al análisis y modelado de los datos. El objetivo de esta fase consiste en establecer una relación entre las variables explicativas y las variables objeto del estudio que posibilite inferir el valor de las mismas con un nivel de confianza determinado.

Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales (análisis discriminante, métodos de agrupamiento, análisis de regresión) y también técnicas específicas (redes neuronales, técnicas adaptativas, lógica difusa, árboles de decisión, reglas de asociación y computación evolutiva).

Finalmente, la última fase del proceso consiste en la valoración de los resultados mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas muestras poblaciones.

3.5 MoProPEI

En el año 2015, como consecuencia de identificar en las distintas metodologías de Explotación de Información una serie de deficiencias y carencias para la ejecución exitosa de un proyecto, se presenta un modelo de procesos que guía el desarrollo del proyecto, considerando tanto los aspectos de gestión, como los técnicos, con la finalidad de generar conocimiento que pueda ser transmitido al cliente para que el mismo sea utilizado como soporte en la toma de decisiones [33]. Las actividades técnicas de este modelo favorecen el progreso fluido del proyecto, entre las que se destaca la de documentar los conocimientos y todo lo referido a la ejecución con el fin de que puedan ser utilizados dándole un valor agregado a futuros proyectos.

El modelo de proceso MoProPEI se estructura de forma jerárquica. Cuenta con cuatro niveles y su jerarquía depende del nivel de especificidad, es decir, el mismo va creciendo a medida que la jerarquía disminuye. Los niveles son:

- **Subprocesos**: es la división de mayor nivel de generalidad, la cual está integrada por fases, y permite distinguir entre las tareas técnicas y las de gestión del proyecto.
- **Fases**: subdivide las tareas en base a sus finalidades generales. Las fases están compuestas por actividades que poseen objetivos asociados.
- **Actividades**: subdivide las tareas en base a sus objetivos específicos.
- **Tareas**: trabajo asociado a un conjunto de técnicas posibles que generan, a partir de un conjunto de elementos de entrada, uno o más elementos de salida.

El proceso consta de 2 subprocesos, cada uno de los cuales tiene una serie de fases que están compuestas por actividades y estas por tareas, Figura 3-7:

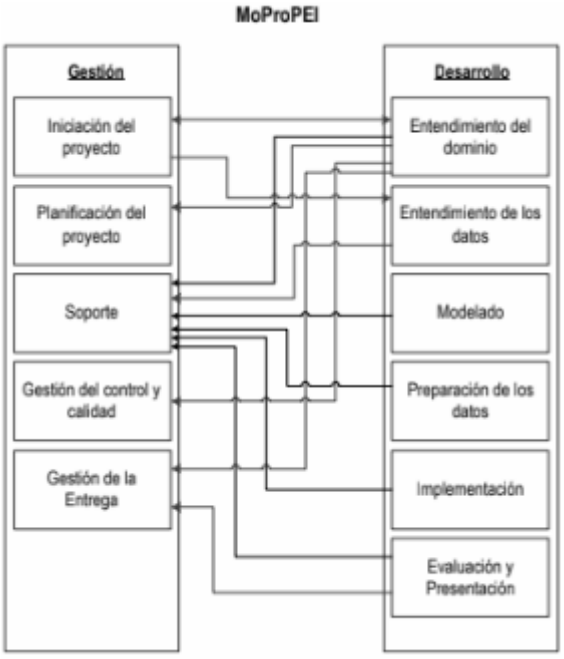


Figura 3-7: Estructura General – MoProPEI

Desarrollo: Vinculado con las actividades de entendimiento y preparación de los datos. Este subproceso se encarga de las tareas asociadas a la obtención de requisitos, la comprensión del negocio y de los problemas del mismo, la identificación de los recursos relevantes para el desarrollo del proyecto -particularmente de las fuentes de datos-, el análisis, comprensión y preparación de los datos existentes en la organización, la identificación y selección de los procesos, técnicas y herramientas a utilizar, la implementación y su posterior evaluación y producción de los resultados obtenidos, cuyo orden se define con el objetivo de reducir la cantidad de iteraciones entre las etapas del subproceso y favorecer la ejecución del mismo, logrando una mejor articulación entre las actividades involucradas.

Gestión: Orientado al control y la administración del proyecto. Este se concibe de forma transversal a las actividades de desarrollo, la ejecución de sus tareas no es de forma lineal, sino que se realizan en base al progreso del proyecto. Abarca la administración del proyecto, la comprensión de la situación del cliente, la identificación de los recursos y del modelo de ciclo de vida. En este subproceso se debe controlar la ejecución de las actividades, realizar las mediciones, definir la viabilidad del proyecto y formalizar el cierre del mismo. A su vez este subproceso de Gestión está conformado por cuatro fases, cada una de las cuales se compone de distintas tareas generales, que identifican un conjunto de actividades con un objetivo específico dentro del proyecto.

1.1. **La fase de Iniciación:** compuesta por cuatro actividades:

- **Definición de la comunicación**: tiene como elementos de entrada el discurso del cliente, el discurso del líder del proyecto y las políticas de las organizaciones intervinientes a partir de los cuales se define el protocolo del proyecto.
- **Exploración de conceptos iniciales**: sus entradas son las salidas de la actividad anterior y el discurso del cliente, produciendo como resultados parciales el reporte de recursos del cliente, el plan de adquisición de conocimiento y el conocimiento adquirido.
- **Evaluación de la situación**: sus elementos de entrada están conformados por todos los elementos de salida de la actividad previa. Se producen los reportes de recursos internos, posibilidad de tercerización, viabilidad e identifica riesgos y contingencias de los mismos.
- **Definición del ciclo de vida**: establece el patrón estructural mediante el cual el proyecto será ejecutado a partir de las características definidas en la actividad previa.

1.2. **La fase Planificación**: se compone de tres actividades generales:

- **Planificación de las actividades:** los elementos de entrada son los proyectos guías identificados, el ciclo de vida y el documento de requisitos definido en el proceso de desarrollo. Se elabora el mapa y calendario de actividades.
- **Planificación de recursos:** a partir de los proyectos guías, el calendario de actividades y el documento de requisitos, se define el plan de tercerización y el reporte de recursos requeridos.
- **Estimaciones y responsabilidades:** a partir de los elementos de salida de la actividad previa, junto con los proyectos guías, el mapa de actividades y el documento de requisitos, se realiza la estimación de costo y se definen los alcances del proyecto y las obligaciones que las partes intervinientes acuerdan (contrato del proyecto).

1.3. **La fase Soporte:** está conformada por 3 actividades:

- **Gestión del ciclo de vida**, en esta, a partir del calendario de actividades y el modelo de ciclo de vida escogido, se determinan los alcances del ciclo de vida y los elementos pendientes de realizar para próximas iteraciones del proyecto (en

caso de que hubiere) definidos en los reportes formales de inicio y fin de ciclo, los que son utilizados de manera global en el desarrollo del proyecto.

- **Gestión del desarrollo**, aquí se definen las responsabilidades de los recursos en el proyecto, teniendo como elementos de ingreso el calendario de actividades, el plan de tercerización (en caso de que hubiere), los riesgos y contingencias del proyecto y del problema de negocio así como los reportes de recursos requeridos y del progreso de las actividades. Como productos de salida se generan contratos y reportes utilizados de manera global en el desarrollo del proyecto.
- **Gestión de la configuración**, en esta, a partir de los avances en el proyecto, se producen ajustes en el documento del proyecto y se establece el manejo de versionado del mismo.

1.4. **La fase Control y Calidad:** integrada por 4 actividades:

- **Control de los recursos:** los elementos de entrada son los contratos de recursos y tercerización, el reporte de recursos requeridos y los problemas y objetivos de negocio, a partir de los que se generan los controles correspondientes a la incorporación de recursos y tercerización de tareas.
- **Mediciones del proyecto**: en esta, a partir del listado de métricas y los costos de las actividades del proyecto, se calculan y controlan las variables de interés para el proyecto, generando los reportes de métricas y de costos.
- **Control de las actividades:** aquí se realizan los controles generales del progreso de las actividades, riesgos y calidad del proyecto, a partir de las salidas de la actividad previa junto con los costos y esfuerzo estimado, los riesgos y contingencias identificados, el calendario de actividades y el reporte de responsabilidades del personal, brindando un control detallado del progreso y posibles desvíos en el desarrollo del plan de proyecto.
- **Gestión del cambio**: se realiza la evaluación, implementación y control de los cambios solicitados a través del desarrollo del proyecto dejando constancia de los mismos. Sus elementos de entrada son el documento de requisitos, los problemas de negocio y riesgos asociados y las solicitudes de cambio, en base a los cuales se producen los reportes de cambios implementados y de control de la integración del cambio.

1.5. **La fase de entrega:** compuesta por 2 actividades:

- **Formalización externa del cierre del proyecto:** en esta, se realiza una verificación y validación formal con el objetivo de constatar que las necesidades del cliente fueron satisfechas y que las obligaciones delegadas a terceros fueron correctamente cumplimentadas, generadas a partir del discurso del cliente, el contrato del proyecto, los contratos de recursos y tercerización y la planilla de criterios de éxito del problema de negocio, el documento de aceptación y el reporte de conclusión de contrataciones.
- **Formalización interna del cierre del proyecto:** aquí se genera una serie de reportes para el control y mejora propia del equipo de trabajo, según la experiencia adquirida a lo largo del proyecto. Sus elementos de entrada son los obtenidos en la actividad previa junto con los reportes de calidad del proyecto, riesgos acontecidos (en caso de que hubiere), calidad del proyecto, métricas y costos.

3.6 ASUM-DM

Fue desarrollada en el año 2015 por IBM Analytics, como una extensión del estándar CRISP-DM. Se caracteriza por ser un proceso iterativo que utiliza metodologías ágiles además de los principios tradicionales de la Ingeniería del Software para lo obtención de resultados óptimos. Consta de 5 grupos de fases globales que se desarrollan secuencialmente, a saber: analizar, diseñar, configurar y construir, desplegar, operar y optimizar, [34].

Es una metodología planteada para el desarrollo del proyecto en un equipo de trabajo, por ende debe existir un conocimiento base de los integrantes del grupo en las diferentes áreas del conocimiento del proyecto de manera que se tenga un lenguaje común.

3.7 TSDP

Es una metodología ágil, iterativa y flexible desarrollada por Microsoft en el año 2017 con el objetivo de ofrecer soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente, a fin de que las empresas perciban las ventajas de su programa de análisis [34]. Tiene una sección dedicada al desarrollo ágil de proyectos en Ciencias de Datos, en la que se explican los pasos a seguir para planificar un sprint, agregar ítems de trabajo al mismo (características, historias de usuario, tareas o bugs) y crear una plantilla de ítem de trabajo dentro de las etapas del ciclo de vida. Presenta 5 fases con la particularidad de que no deben realizarse secuencialmente.

3.8 Fases comunes de las metodologías

Las metodologías descriptas anteriormente tienen las siguientes fases en común, Tabla 3-1.

- **Comprensión del negocio:** se evalúa el problema que se abordará y el contexto organizacional.
- **Selección y preparación de los datos:** se procede a realizar la limpieza y transformaciones necesarias para crear el set de datos que se utilizará.
- **Aplicación de las técnicas de minería:** comprende análisis de regresión, árboles de decisión, redes neuronales, etc. y modelado de los nuevos patrones.
- **Evaluación de los resultados obtenidos:** se analiza la posibilidad de implementar el modelo obtenido y, de no ser factible, se realiza nuevamente el proceso.
- **Implementación y difusión** del nuevo conocimiento dentro de la organización.

Tabla 3-1: Fases en común-Elaboración propia

FASES	KDD	CRISP-DM	Catalyst	SEMMA	MoProPEI	ASUM	TSDP
Comprensión del Negocio	Comprensión del dominio	Comprensión del negocio	Modelado del negocio		Análisis del dominio Comprensión del problema de negocio	Comprender el negocio	Entendimiento del negocio
Selección y Preparación de los Datos	Crear el conjunto de datos. Limpieza y transformación. Construcción del set de datos	Entendimiento y Preparación de los datos	Preparación de los datos	Muestreo Comprensión Modificación	Análisis de los datos. Exploración de los datos. Evaluación de los datos.	Comprender los datos.	Adquisición de datos y comprensión
Aplicación de las Técnicas de Minería	Selección del algoritmo. Minería de Datos	Modelado	Modelado inicial	Modelado	Modelado del problema. Configuración del modelo	Diseñar. Construir el modelo	Modelado
Evaluación de los resultados obtenidos	Interpretación	Evaluación	Refinamiento del modelo	Valoración	Implementación del modelo	Implementar	
Implementación y difusión	Utilización del nuevo modelo	Despliegue	Comunicación del nuevo conocimiento		Presentación de los resultados	Desplegar	Despliegue

3.9 Metodología a utilizar en el TFM

Para el presente trabajo se ha optado por utilizar la Metodología MoProPEI teniendo en cuenta las fortalezas de la misma, a las que hace referencia [33], y que son descriptas a continuación.

- **Se centra en la generación del conocimiento**: el objetivo final del proyecto en producir piezas de conocimiento importantes para la toma de decisiones.
- **Se adapta a las necesidades del proyecto**: permite ajustar la ejecución del proyecto a partir de las características del mismo.
- **Se enfoca en la gestión:** fortalece la planificación, administración, documentación de todos los aspectos necesarios para el desarrollo de un proyecto de Explotación de Información.
- **Reduce el trabajo redundante**: a través de la planificación y el ordenamiento de las tareas, reduce la cantidad de iteraciones para el desarrollo del proyecto.
- **Se enfoca en la sistematización**: identifica y define las tareas a realizar de forma específica, contribuyendo a lograr el objetivo final.

Asimismo, a fines de documentar los conceptos educidos durante las distintas fases, se elabora un conjunto de plantillas en base a lo propuesto por [10]. Cada plantilla se asocia a un concepto educido y contiene una descripción detallada del mismo, permitiendo la evolución del concepto a lo largo del proyecto. De esta forma, a partir de la información contenida en las plantillas, el profesional de Ciencia de Datos puede realizar las actividades del proyecto tal como identificar los repositorios de datos, determinar el modelo y seleccionar las herramientas a utilizar.

Capítulo 4

Análisis del caso de estudio

4 Análisis del caso de estudio

En este capítulo se hace una breve descripción del problema, se enuncian algunas de las normativas necesarias para comprender la permanencia del alumno en la institución (sección 4.1), se describe la recopilación de los datos (4.2), se presentan características de la población en estudio sección (4.3), se detallan algunas de las políticas implementadas a fin de aumentar la retención de los estudiantes en la institución (sección 4.4).

4.1 Descripción del problema

La Universidad Nacional del Nordeste se creó en el año 1956. Una de las características es que tiene 3 sedes, ubicadas en dos provincias lindantes, Corrientes y Chaco. Son doce las Unidades Académicas que integran la institución, una de ellas es la Facultad de Humanidades, sita en Resistencia, Capital de la provincia del Chaco. Esta Unidad Académica inició sus actividades en el año 1958, actualmente se dictan en la misma un total de catorce carreras de grado, tres de ellas con título de pregrado.

La deserción universitaria y el rezago o desgranamiento de las carreras de grado son situaciones que preocupan a las autoridades y docentes de la UNNE. Estos representan pérdidas financieras para la sociedad y para la institución, así como desmotivación para los alumnos que no logran la meta. Por eso es importante identificar las principales causas, endógenas y exógenas, que motivan el abandono y el desgranamiento de los estudiantes que inician una carrera universitaria, con el fin de diseñar y elaborar políticas eficaces que permitan aumentar la retención estudiantil.

En el año 2017 en la UNNE, institución que ofrece 49 carreras de grado y 15 de pregrado, se registraron 11.988 ingresos y 3.154 egresos. En la Facultad de Humanidades específicamente se registraron 1.300 ingresos y 250 egresos [35]. Estos datos ponen de manifiesto la brecha existente entre ingresos y egresos.

La carrera en estudio es el Profesorado en Ciencias de la Educación. La misma tiene dos planes de estudios activos, aprobados en 1983 y 2000, respectivamente. El último es el vigente y con el que se trabaja en el TFM [36]. Para la obtención del título, el alumno debe aprobar un total de 35 asignaturas, distribuidas en 5 años.

El procedimiento a realizar para el ingreso a la UNNE es el siguiente:

En octubre se convoca a los aspirantes a preinscribirse a la propuesta formativa elegida. En diciembre del mismo año el interesado debe acercarse a la Unidad Académica con el fin de presentar los requisitos documentales establecidos en la normativa [37]. Cumplidos los requisitos, el estudiante es considerado alumno de la carrera.

En marzo del año académico al que corresponde el ingreso, se abren las inscripciones al cursado de las asignaturas, según lo establecido en el Régimen Pedagógico [38]. El mismo también establece que la permanencia del alumno en la institución se rige por la Resolución del Consejo Superior [39]. Esta normativa determina que si el estudiante no aprueba un mínimo de 2 asignaturas durante el año académico anterior, deberá solicitar la readmisión o bien una excepción, justificando el motivo del no cumplimiento. El alumno ingresante deberá tener aprobada 1 asignatura en el año de ingreso para dar cumplimiento a la permanencia.

Una particularidad de la UA a la que pertenece la carrera en estudio es el alto número de alumnos que no realizan actividad académica, no rinden evaluaciones parciales, finales e incluso tampoco se inscriben a cursar asignaturas. En la Tabla 4-1, se detalla el número de alumnos por cohorte (2010 a 2018), sin actividad académica y los que no cumplieron con el requisito de permanencia en la institución, de la carrera en estudio.

Tabla 4-1: Análisis por cohorte según el número de asignaturas aprobadas-PCE-Elaboración propia

			Alumnos que no cumplieron con el requisito de permanencia agrupados año académico:								
Cohorte	**Ingreso**	**S/Activ.**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
2010	189	68-35%	**73-49%**	41-33%	36-29%	34-28%	30-23%	19-15%	13-10%	9-7%	4-3%
2011	186	52-28%		**83-61%**	60-44%	49-36%	39-29%	34-25%	15-11%	10-7%	7-5%
2012	176	59-33%			**78-66%**	40-34%	39-33%	25-21%	21-18%	13-11%	7-6%
2013	155	56-36%				**77-77%**	45-45%	43-43%	30-30%	23-23%	12-12%

2014	138	44-32%					**70-74%**	40-42%	36-38%	26-27%	11-11%
2015	113	44-39%						**45-65%**	33-47%	32-46%	24-34%
2016	136	40-29%							**56-58%**	41-42%	28-29%
2017	93	36-38%								**36-63%**	27-47%
2018	94	33-5%									**42-68%**

En la Tabla 4-2, se muestra el número ingresos y de egresos anuales, con estos datos queda de manifiesto la baja proporción de egresos anuales.

Tabla 4-2: Ingresos / Egresos-PCE Elaboración propia

Cohorte	Ingresos	Egresos	% Egresos
2010	189	46	24
2011	186	41	22
2012	176	33	19
2013	155	50	32
2014	138	39	28
2015	113	44	39
2016	136	42	31
2017	93	49	53
2018	94	41	43

La Tabla 4-3 permite inferir que, sobre un total de 1280 ingresantes durante el período 2010 a 2018, el total de egresados de esas cohortes es de 188, con lo cual al año 2018 el porcentaje de egresados de las cohortes en estudio es del 14,6%.

Tabla 4-3: Ingresos/Egresos por cohorte-PCE-Elaboración propia

Cohorte	Ingresos	E-2014	E-2015	E-2016	E-2017	E-2018	Total Egresos
2010	189	14	19	17	10	3	**63**
2011	186		14	19	16	5	**54**
2012	176			7	10	4	**21**
2013	155				16	20	**36**

2014	138					14	**14**
2015	113						
2016	136						
2017	93						
2018	94						
	1280						**188**

La Tabla 4-4 permite analizar los índices de abandono por cohorte durante el período 2010-2014. Se considera este período debido a que los datos fueron extraídos en el año 2018, la duración teórica de la carrera es de 5 años, por ende no se registraron egresos de las cohortes 2015 – 2018.

Tabla 4-4: Índice de abandono por cohorte

Cohorte	Ingresos	Total Egresos	Índice de rezago o deserción
2010	189	63	66,66
2011	186	54	70,96
2012	176	21	88
2013	155	36	76,77
2014	138	14	89,85

4.2 Recopilación de datos

En la institución, desde el año 1992 se utilizan sistemas de información que almacenan datos básicos de la población estudiantil. En el año 2003 se realizó la migración de estos datos que incluían la actividad académica del alumnado al Sistema de información Universitaria Guaraní, que recaba mayor cantidad y variedad de datos del alumno.

El SIU GUARANI se define como un sistema que registra las actividades de la gestión académica dentro de la universidad desde que el alumno se inscribe hasta que egresa. El mismo tiene como objetivo la administración de las tareas en forma óptima y segura, con la finalidad de obtener información consistente para los niveles operativos y directivos.

El Sistema de Información Universitaria (SIU) es una base de datos de tipo relacional, es una colección de elementos de datos con relaciones predefinidas entre ellos. Estos elementos se organizan como un conjunto de tablas con columnas y filas. Las tablas se utilizan para guardar información sobre los objetos que se van a representar en la base de datos. Cada

columna de una tabla guarda un determinado tipo de datos y un campo almacena el valor real de un atributo. Las filas de la tabla representan una recopilación de valores relacionados de un objeto o entidad. Cada fila de una tabla podría marcarse con un identificador único denominado clave principal, mientras que filas de varias tablas pueden relacionarse con claves foráneas. Se puede obtener acceso a estos datos de muchas formas distintas sin reorganizar las propias tablas de la base de datos.

Luego de evaluar el modelo de datos del sistema, se decide utilizar 13 de las tablas relacionadas Tabla 4-5.

Tabla 4-5: Descripción de tablas

DESCRIPCION DE TABLAS	
Nombre	**Descripción**
sga_personas	En esta tabla se graban los datos del alumno que no sufrirán cambios con el tiempo. Ej. Fecha y lugar de nacimiento, datos de los progenitores.
sga_datos_censales	En esta tabla se registran datos como domicilio de procedencia, domicilio de residencia, estudios cursados por los padres, situación laboral del alumno y de los padres.
sga_dat_cen_aux	Contiene datos como año de egreso del secundario.
sga_dat_cen_aux2	Esta contiene datos referentes a forma de costear los estudios.
sga_alumnos	Se registran aquí las carreras a las que se inscribió el estudiante y la cohorte a la que pertenece.
sga_cursadas	Esta tabla contiene datos referentes a las materias regularizadas y promocionadas.
vw_hist_academica	Esta es una vista de las asignaturas aprobadas y desaprobadas con datos tales como fecha, calificación, acta.
sga_perd_regul	Se registran en esta las veces que el alumno perdió su condición de regular, el motivo y la fecha en que se lo rehabilito como alumno regular.
sga_det_perd_regul	Se registran en esta las veces que el alumno solicito excepción para obtener la condición de regular.
sga_eval_parc_alum	Se registran las calificaciones de las evaluaciones parciales.
sga_insc_cursadas	Esta tabla contiene las inscripciones al cursado.
sga_atrib_mat_plan	Atributos de las asignaturas del plan (si se puede promocionar, rendir libre, si es pro mediable, año de cursado).
sga_titulos_otorg	Se registra en esta tabla a los alumnos que egresaron con datos como fecha de egreso, promedio académico y general, título obtenido, fecha de inicio del trámite.

Utilizando el lenguaje SQL, se extrae un conjunto de 1280 registros correspondiente a la matrícula de las cohortes 2010 a 2018 de la carrera Profesorado en Ciencias de la Educación Plan 2000.

Se extraen datos de tipo personal, familiar, económico social y académico Tabla 4-6.

Tabla 4-6: Tipo de datos extraídos

Tipo	Dato
Personal	Fecha de nacimiento Localidad colegio secundario

	Localidad de procedencia Localidad de residencia Nacionalidad Sector colegio secundario Sexo Título secundario Vive padre/madre Ultimo nivel de estudios de padre/madre
Familiar	Estado civil Familiares a cargo Tiene hijos
Económico social	Cómo costea los estudios Es remunerado Horas de trabajo Trabaja Relación entre carrera y trabajo Tiene beca
Académico	Cohorte Calidad Regular Número de materias aprobadas durante el 1º, 2º y 3º año Readmisiones/excepciones Fecha de inscripción al cursado de cada materia Fecha en la que regularizo Fecha en la que aprobó/desaprobó examen Total de materias aprobadas Total de materias regularizadas Total de inscripciones al cursado

4.3 Características de la población en estudio

Se analizó la información correspondiente a las cohortes 2010 a 2018 de la carrera Profesorado en Ciencias de la Educación Plan 2000 y se obtuvieron los siguientes resultados:

Con respecto al género el 76% son mujeres y el 24% varones Gráfico 4-1.

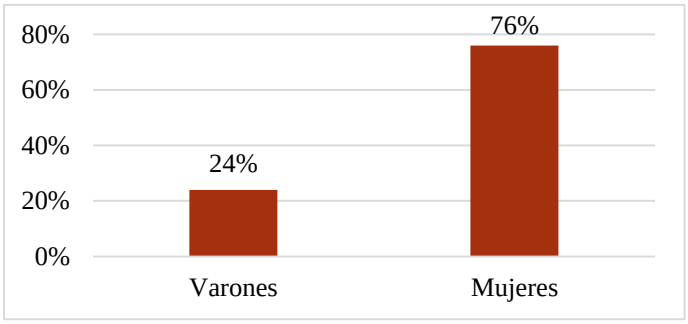


Gráfico 4-1: Alumnos por género

Con respecto a la edad de ingreso a la carrera el 56% tenía entre 18 y 20 años, el 25% entre 21 y 25, el 14% entre 26 y 35 y el 5% 36 años o más, Gráfico 4-2.

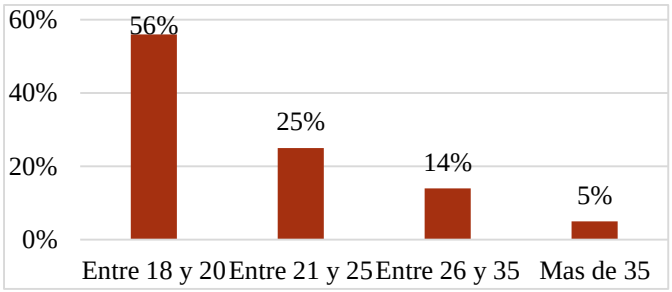


Gráfico 4-2: Alumnos según edad al ingresar a la carrera

En cuanto a la procedencia, el 80% declara proceder de la provincia del Chaco, el 15% de la provincia de Corrientes y el 5% restante de otras provincias, Gráfico 4-3.

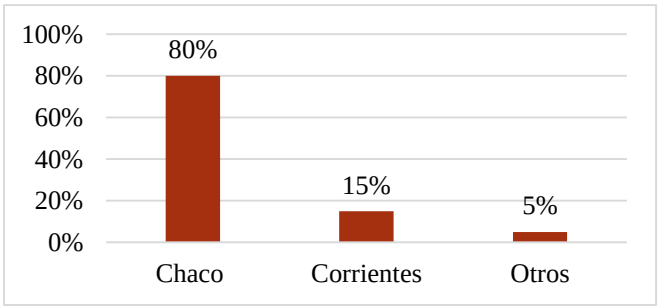


Gráfico 4-3: Alumnos según procedencia

De los procedentes de la provincia del Chaco, el 50% es de la ciudad capital, el 20% de localidades muy próximas y el 30% de localidades del interior de la provincia, Gráfico 4-4.

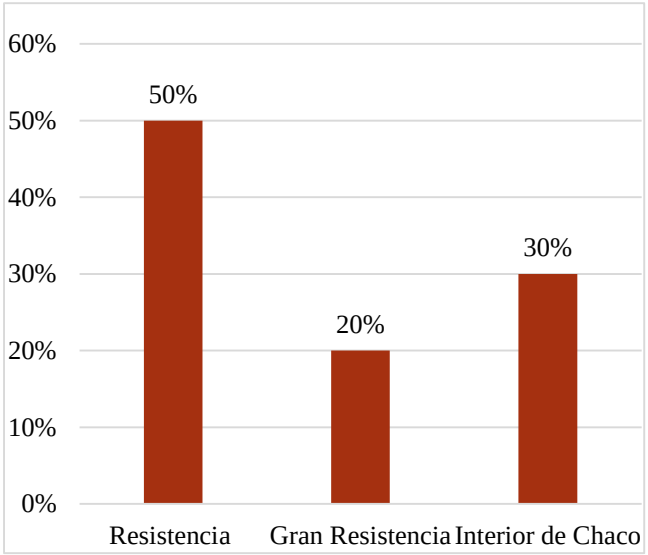


Gráfico 4-4: Alumnos según localidad (Chaco)

El 51% de los alumnos posee título secundario Bachiller, el 42% Educación Polimodal y el 4% Técnico, Gráfico 4-5.

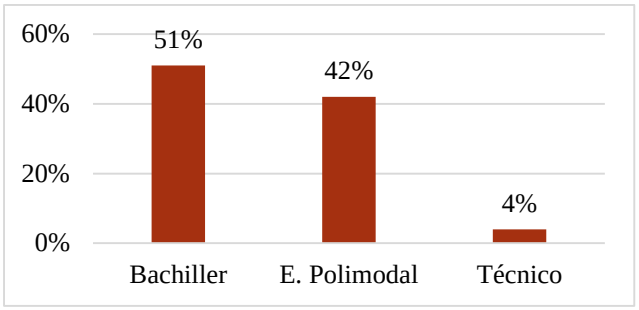


Gráfico 4-5: Alumnos según título secundario

4.4 Políticas universitarias implementadas para aumentar la retención

Con el objetivo de contribuir a minimizar la deserción, a lo largo de los últimos años se implementó en la Universidad Nacional del Nordeste una serie de programas tendientes a mitigar las problemáticas que influyan directa o indirectamente en el desempeño del estudiante y que contribuyan a mejorar las condiciones de enseñanza y aprendizaje a lo largo de su trayectoria académica.

Los problemas de rendimiento académico pueden ser consecuencia de la brecha entre las exigencias de la carrera y la formación adquirida en el nivel secundario. Así es que desde hace décadas se ofrece a los ingresantes diferentes cursos de ambientación a la vida universitaria. El objetivo de los mismos consiste en ofrecerles un espacio donde fortalezcan sus saberes y competencias para llevar adelante su inmersión en el ámbito universitario.

Específicamente en la Facultad de Humanidades se desarrolló en el año 2016 un Programa Institucional denominado Ingreso, Tutorías, Primeros [40], que tiene como principal objetivo articular acciones y recursos con los docentes y tutores pares para implementar un sistema de acompañamiento a los estudiantes en el proceso de incorporación a la universidad y en su tránsito por los primeros tramos.

A fin de paliar situaciones económicas desfavorables, aumentaron las becas adjudicadas en cantidad y variedad: se otorgan becas de transporte, comedor, conectividad, fotocopias de bibliografía entre otras. Desde el año 2019, se implementa en la UA un programa de acompañamiento académico a los alumnos con discapacidad.

En virtud de la confiabilidad de los datos que contiene el sistema SIU GUARANI y al procedimiento utilizado para la extracción de los mismos, se puede afirmar que existe un problema tangible en la carrera en estudio, manifestado por el alto índice de abandono.

No obstante las políticas implementadas por la institución orientadas a paliar las situaciones evidenciadas tales como deserción y/o rezago, a luz de los resultados aún poco alentadores obtenidos, resulta evidente que es necesario profundizar en la detección de factores en común de los estudiantes en riesgo de abandono.

A fin de abordar las problemáticas de la deserción y el rezago estudiantil este trabajo propone el desarrollo de un modelo predictivo, basado en la utilización de herramientas de minería de datos, que permita a nivel de gestión institucional, identificar predictivamente factores de riesgo que puedan interferir en la permanencia del estudiante en la carrera o determinar la deserción.

Capítulo 5

Desarrollo del modelo predictivo, utilizando la metodología MoProPEI, validación del caso de estudio

5 Desarrollo aplicando la metodología MoProPEI

Este capítulo tiene el objetivo de mostrar el desarrollo metodológico del TFM a través de la metodología MoProPEI, propuesta en [33] y seleccionada en el capítulo 3 del presente trabajo, utilizando datos de alumnos de la carrera PCE de la Facultad de Humanidades, cohortes **2010 a 2018**. En la sección 5.1 se desarrolla el Subproceso de Gestión y sus fases: Iniciación del Proyecto (sección 5.1.1), Planificación del Proyecto (sección 5.1.2), Soporte (sección 5.1.3), Gestión de Control y Calidad (sección 5.1.4), Gestión de la Entrega (sección 5.1.5). En la sección 5.2 se explica el Subproceso de Desarrollo con sus fases: Entendimiento del Dominio (sección 5.2.1), Entendimiento de los Datos (sección 5.2.2), Modelado (sección 5.2.3), Preparación de los Datos (sección 5.2.4), Implementación (sección 5.2.5) y Evaluación y Presentación (sección 5.2.6).

5.1 Subproceso de gestión

Este subproceso abarca todas las actividades asociadas a la interacción inicial con la institución (cliente), la estructuración del proyecto, planificación y administración de los recursos, control y calidad del desarrollo del proyecto [33].

5.1.1 Fase: Iniciación del proyecto

En esta fase se evalúan las características del trabajo a realizar, los recursos humanos involucrados y la valoración de su posibilidad de éxito.

5.1.1.1 Definición de la comunicación

Es importante definir las necesidades y canales de comunicación durante el desarrollo del trabajo.

Definir Protocolo de comunicación Externa

Entrada

- Discurso del cliente - Tabla 5-1

Después de la primera reunión formal con las autoridades de la Unidad Académica de la que depende la carrera en estudio, se elaboró el Acta Nº 1, en la que queda de manifiesto el discurso del cliente.

Tabla 5-1: Discurso del Cliente - Acta Reunión Nº 1

DISCURSO DEL CLIENTE-ACTA DE REUNIÓN Nº 1			
Convocador	Viviana Moschner	**Lugar**	Secretaría Académica-HUM-UNNE

Fecha	07/07/2018	**Hora Inicio**	09:00	**Hora Fin**	10:00
PARTICIPANTES					
Nombre y Apellido		**Área / Sector**		**Mail**	
Mariana Ojeda		Secretaría Académica		academica@hum.unne.edu.ar	
María Phipps		Secretaría de Asuntos Estudiantiles		sae@hum.unne.edu.ar	
Guadalupe Portillo		Asesoría Pedagógica		asesoría@hum.unne.edu.ar	
Soledad Almirón		Asesoría Pedagógica		asesoria@hum.unne.edu.ar	

ID	**TEMAS TRATADOS**
1	**Problema de negocio:** Las autoridades de la Facultad de Humanidades plantean la necesidad de conocer los motivos por los que la relación entre ingresos y egresos en dicha Unidad Académica es muy desproporcional. ¿Podrían existir factores personales, económicos o académicos que llevan al rezago o abandono de las carreras que se dictan en esta UA?
2	**Objetivo del trabajo:** Identificar características personales y/o académicas de alumnos en situación de riesgo de abandono en la carrera de grado Profesorado en Ciencias de la Educación de la Facultad de Humanidades de la UNNE.
3	**Disponibilidad de los interesados**: Los interesados estarán disponibles de lunes a viernes en el horario matutino (de 9 a 11 horas) previa coordinación por correo electrónico. Para concretar una reunión personal es indispensable solicitar vía correo electrónico con 48 horas hábiles de anticipación.
4	**Comunicación con los interesados/involucrados:** En la reunión inicial con las autoridades de la Unidad Académica y las del departamento al que pertenece la carrera en estudio se acordó realizar las comunicaciones internas a través de correo electrónico.
5	**Disponibilidad de los datos:** Se dispone de los datos obtenidos a través del SIU GUARANI y otros sistemas tal como SIU ARAUCANO.
6	**Rango de los datos:** 2010 a 2018
ID	**PROBLEMAS PLANTEADOS**
1	Se plantearon diferentes definiciones de alumno posible desertor: a) Alumnos que no hayan aprobado al menos dos materias en el año académico anterior y no hayan solicitado readmisión o excepción. b) Alumnos que no hayan regularizado o aprobado alguna actividad académica en los dos últimos años académicos. **c) Alumnos que no hayan realizado ninguna actividad académica, incluyendo inscripción a cursar o rendir materias, en los dos últimos años.**

Salida

- Protocolo de comunicación externa - Tabla 5-2

Además, en la primera reunión se establece el protocolo de comunicación con el cliente, el mismo se describe en la en la Tabla 5-2.

Tabla 5-2: Protocolo de Comunicación Externa

PROTOCOLO DE COMUNICACIÓN EXTERNA			
Analista	Viviana Moschner	**Fecha**	12/07/2018

ID	**Descripción**
CE-1	La comunicación con las autoridades de la UA se realizará cada 15 días por correo electrónico.
CE-2	En caso de requerir una reunión personal se deberá acordar por correo con 48 horas de anticipación.
CE-3	La solicitud de los documentos y/o resoluciones para conocer el funcionamiento de la organización serán solicitados a la Secretaría Académica.

Definir Protocolo de documentación Interna

Entrada

- Discurso del cliente - Tabla 5-1

Salida

- Protocolo de documentación interna - Tabla 5-3

En la Tabla 5-3 se explica el procedimiento que se utilizará para la obtención y resguardo de la documentación requerida.

Tabla 5-3: Protocolo de Documentación Interna

PROTOCOLO DE DOCUMENTACIÓN INTERNA			
Analista	Viviana Moschner	**Fecha**	12/07/2018
ID	**Descripción**		
DI-1	Los documentos, tales como estatuto, resoluciones, régimen pedagógico y otros, se solicitarán a la Secretaría Académica en formato digital. Los mismos serán resguardados por el líder del trabajo y enviados al equipo.		
DI-2	Los datos extraídos de los sistemas de información serán solicitados al técnico responsable del sistema. Se enviarán al líder del equipo en planillas Excel, quien los resguardará y hará el envío a los integrantes del grupo de acuerdo a las necesidades.		

Definir Protocolo de comunicación interna

Entrada

- Discurso del cliente - Tabla 5-1

Salida

- Protocolo de comunicación interna - Tabla 5-4

En la Tabla 5-4 se describe el protocolo que se utilizará para la comunicación con los interesados en el proyecto.

Tabla 5-4: Protocolo de Comunicación Interna

PROTOCOLO DE COMUNICACIÓN INTERNA			
Analista	Viviana Moschner	**Fecha**	12/07/2018
ID	Descripción		
CI-1	La comunicación con los expertos del negocio se realizará cada 15 días por correo electrónico.		
CI-2	En caso de requerir una reunión personal con los expertos en el negocio se acordará por correo con 48 horas de anticipación.		
CI-3	La comunicación con el responsable de los datos del sistema se hará personalmente, inicialmente una vez por semana o bien cada vez que el líder del proyecto lo requiera.		
CI-4	Con uno de los responsables del Sistema Araucano, la comunicación será por correo electrónico o teléfono en horario laboral, Lunes a Viernes de 7 a 12 horas.		
CI-5	La comunicación con el Censista de Datos del Proyecto se realizará una vez por semana.		

5.1.1.2 Exploración de conceptos iniciales

Planificar la adquisición de conocimientos

Entradas

- Discurso del cliente - Tabla 5-1
- Conceptos teóricos asociados al dominio - Tabla 5-5

Luego de realizar un estudio de las características de la organización se identifican los conceptos teóricos más relevantes, Tabla 5-5.

Tabla 5-5: Conceptos Teóricos Asociados al Dominio

CONCEPTOS TEÓRICOS ASOCIADOS AL DOMINIO			
Analista	Viviana Moschner	**Fecha**	18/07/2018

Concepto	**Descripción**	**Referencia**
CT-1	**Estudiantes:** es la suma de los nuevos inscriptos más los reinscriptos en una carrera determinada.	Manual de definiciones conceptuales y operativas –SIU ARAUCANO- [41].
CT-2	**Excepción:** pueden solicitar excepción al régimen de permanencia los estudiantes que, por razones de salud, laborales, obtención de una beca para estudio fuera del ámbito de la Universidad, no han podido cumplir con los requisitos de dicho régimen.	Régimen Permanencia UNNE [39].
CT-3	**Graduado**: son los estudiantes que completan todos los cursos y requisitos reglamentarios de la oferta a la que pertenecen.	Manual de definiciones conceptuales y operativas –SIU ARAUCANO- [41].
CT-4	**Nuevos Inscriptos:** aspirantes que, habiendo cumplido con los requisitos reglamentados por cada institución, son admitidos como estudiantes en una determinada carrera.	Manual de definiciones conceptuales y operativas –SIU ARAUCANO- [41].
CT-5	**No Regular:** estudiante que no pudo cumplir con los requisitos establecidos en el Régimen de Permanencia de la UNNE.	Régimen Permanencia UNNE [39]
CT-6	**Permanencia:** régimen establecido por la UNNE en el que se establecen requisitos mínimos para ser considerado alumno Regular.	Régimen Permanencia UNNE [39].
CT-7	**Readmisión:** acto administrativo por el cual el estudiante No Regular solicita se lo considere como alumno Regular de la carrera para poder realizar actividades académicas.	Régimen Permanencia UNNE [39].
CT-8	**Régimen Pedagógico**: conjunto de normas utilizadas para acordar los principios generales y particulares que regirán la actividad académica de docentes y alumnos a los fines de alcanzar el desarrollo integral de todo el proceso educativo.	Régimen Pedagógico [38].

CT-9	**Regular:** estudiante que pudo cumplir con los requisitos necesarios establecidos en el Régimen de Permanencia de la UNNE.	Régimen Permanencia UNNE [39].
CT-10	**Reinscripción:** la reinscripción puede significar: anotarse en una o más materias, inscribirse o rendir examen final, inscribirse para presentar tesis, tesina, trabajo final, etc.	Manual de definiciones conceptuales y operativas –SIU ARAUCANO [41].
CT-11	**Reinscriptos: e**studiantes a los que se les actualiza la inscripción en la misma carrera en un año académico posterior.	Manual de definiciones conceptuales y operativas –SIU ARAUCANO- [41].

- Estudio de la organización - Tabla 5-6

En la Tabla 5-6 se describen los conceptos que definen el funcionamiento de la organización en estudio (UNNE y Facultad de Humanidades).

Tabla 5-6: Estudio de la Organización

ESTUDIO DE LA ORGANIZACIÓN - HUM			
Analista	Viviana Moschner	**Fecha**	18/07/2018
ID	**Descripción**		**Referencia**
Inst-1	Según el estatuto de la UNNE, las Facultades son, dentro de la Universidad, unidades académicas, administrativas y de gobierno.		Estatuto UNNE [42].
Inst-2	Los órganos de Gobierno de las UA son: - Consejos Directivos - Decanos		Estatuto UNNE [42].
Inst-3	La Facultad de Humanidades actualmente ofrece 14 carreras de grado. 13 de las mismas se dictan en el Campus Resistencia y 1 de ellas en el Campus Sargento Cabral-Corrientes. También se dicta 1 carrera de grado en la Extensión Áulica de General Pinedo-Chaco. La UA ofrece el dictado de 8 carreras de Posgrado.		Oferta Académica
Alu-1	En octubre se abre la preinscripción online a las distintas ofertas del UNNE. El interesado debe completar el formulario homónimo.		Resolución CS Preinscripción UNNE
Alu-2	En diciembre el aspirante debe presentarse en la UA con el formulario de preinscripción impreso y con la documentación requerida según Resolución emitida por el CS. En dicho acto se concreta la inscripción a la propuesta elegida.		Resolución CS Ingreso UNNE
Alu-3	El aspirante puede inscribirse a más de una carrera en diferentes UA.		Régimen Pedagógico-HUM-[38].
Alu-4	En marzo el alumno debe proceder a la inscripción al cursado de las materias, utilizando la plataforma de autogestión del SIU GUARANI.		Régimen Pedagógico-HUM [38].

Alu-5	Las asignaturas son cuatrimestrales o anuales. Las mismas pueden cursarse con modalidad regular o promocional.	Régimen Pedagógico-HUM [38].
Alu-6	Si realiza el cursado con modalidad promocional y cumple con todos los requisitos establecidos en el Régimen Pedagógico, al finalizar el periodo tendrá la asignatura aprobada.	Régimen Pedagógico-HUM [38].
Alu-7	Si realiza el cursado con modalidad promocional y no aprueba los 3 parciales que la modalidad exige, podrá regularizar la asignatura. Para aprobar la misma deberá rendir el examen final.	Régimen Pedagógico-HUM [38].
Alu-8	A los fines de su permanencia como alumno Activo Regular se considerará lo establecido en el Régimen de Regularidad de la UNNE. La condición de alumno Activo Regular vale para la carrera en la que se haya inscripto y aprobado las materias. En caso de pérdida de tal condición, esta sólo podrá ser recuperada mediante los procedimientos de readmisión y excepción contemplados en el Régimen de Permanencia.	Régimen Pedagógico-HUM [38].

Salida

- Plan de adquisición de conocimientos - Tabla 5-7

En la Tabla 5-7 se detalla el procedimiento para adquirir los conocimientos necesarios.

Tabla 5-7: Plan de Adquisición de Conocimientos

PLAN DE ADQUISICION DE CONOCIMIENTOS			
Analista	Viviana Moschner	**Fecha**	18/07/2018
ID	**Descripción**		**Referencia**
AC-1	Se solicitará a la Secretaría Académica el Estatuto de la UNNE, el Régimen Pedagógico de la UA, las diferentes Resoluciones de Consejo Superior referidas a la permanencia de los estudiantes y el plan de estudios del PCE.		Protocolo de Documentación Interna
AC-2	Se harán reuniones con los expertos, Profesoras Guadalupe Portillo y Soledad Almirón, a fin de analizar los documentos solicitados y recabar información acerca de los pasos que deben realizar los alumnos desde la preinscripción y hasta su egreso.		Protocolo de Comunicación Interna
AC-3	Con el Director del departamento de Ciencias de la Educación se analizará el Plan de Estudios en vigencia. (Duración teórica en años, total de materias, tipos de actividades, carga horaria).		Protocolo de Comunicación Interna
AC-4	Con el técnico responsable del sistema se harán reuniones a fin de conocer el modelo de datos, determinar cuáles son las tablas que podrían ser útiles para obtener datos personales y académicos de los alumnos de la carrera en estudio.		Protocolo de Comunicación Interna

Implementar técnicas de adquisición de conocimientos

Entradas

- Plan de adquisición de conocimientos - Tabla 5-7

Salida

- Conocimiento adquirido - Tabla 5-8

En la Tabla 5-8 se detallan brevemente los conceptos más relevantes que serán utilizados en tareas posteriores.

Tabla 5-8: Conocimiento Adquirido

CONOCIMIENTO ADQUIRIDO			
Analista	Viviana Moschner	**Fecha**	08/08/2018
ID	**Descripción**		**Referencia**
CA-1	Luego de analizar el Régimen Pedagógico de la UA, se alcanzó la comprensión de los pasos que el alumno debe hacer desde la preinscripción y hasta graduarse.		Régimen Pedagógico-HUM [38].
CA-2	El aspirante, por lo general en octubre, debe completar un formulario online con el fin de preinscribirse a una o más carreras. En el mismo se solicitan: datos personales como fecha y localidad de nacimiento, estado civil, lugar de procedencia, lugar de residencia, forma de financiamiento de sus estudios, datos referidos a la institución en la que finalizó o finalizará los estudios de nivel secundario, datos de los progenitores.		Resolución CS Preinscripción UNNE [37].
CA-3	Durante todo el mes de diciembre, según la normativa, el aspirante deberá presentarse con la documentación establecida en la Resolución de Inscripción en la Unidad Académica para concretar la Inscripción a la/s propuesta/s elegida/s.		Resolución CS Ingreso UNNE [37].
CA-4	Luego del cursillo de nivelación, no eliminatorio, el alumno deberá inscribirse al cursado de las asignaturas utilizando la plataforma de autogestión del SIU GUARANI.		Régimen Pedagógico-HUM [38].
CA-5	Tendrá la opción de cursar con modalidad Regular o Promocional. Cada una de estas tiene requisitos específicos.		Régimen Pedagógico-HUM [38].
CA-6	Si cursó con modalidad Promocional y cumplió con los requisitos necesarios al finalizar el periodo lectivo, que puede ser cuatrimestral a anual, habrá aprobado la materia.		Régimen Pedagógico-HUM [38].
CA-7	Si cursó con modalidad Regular y cumplió con los requisitos necesarios al finalizar el periodo lectivo, habrá regularizado la asignatura. Para su aprobación deberá rendir examen final.		Régimen Pedagógico-HUM [38].
CA-8	La regularidad obtenida tiene una duración de 3 años. Luego de transcurrido este periodo, el alumnos quedará libre.		Régimen Pedagógico-HUM [38].
CA-9	Existen Resoluciones emitidas por CS que determinan las condiciones para permanecer en la carrera como alumno Activo Regular, así como también detallan los motivos aceptados para solicitar excepción, o bien en caso de haber perdido la permanencia, solicitar la readmisión.		Régimen Permanencia UNNE [39].
CA-10	El alumno Activo Regular puede realizar actividad académica, mientras que el alumno Activo No Regular no tiene ese derecho.		Régimen Pedagógico-HUM [38].
CA-11	En febrero de cada año, se procede a realizar el control de regularidad para saber en qué condición quedan todos los alumnos que hayan estado como activos regulares el año anterior.		Régimen Pedagógico-HUM [38].
CA-12	Si el alumno ingresó el año previo al del control, tuvo que aprobar al menos una asignatura del plan de estudios correspondiente para quedar con la condición de Activo Regular. Si el alumno ingresó antes, tuvo que aprobar como mínimo dos asignaturas para obtener esta condición.		Régimen Permanencia UNNE [39].

CA-13	El plan de estudios del Profesorado en Ciencias de la Educación tiene una duración teórica de 5 años. El alumno debe aprobar 32 asignaturas y acreditar 3 competencias. Las materias son cuatrimestrales con carga horaria de 72 o 92 horas, o anuales con carga horaria de 144 horas. El plan está diseñado para cursar en promedio 7 materias por año. La asignatura Práctica de la Enseñanza tiene, además de las evaluaciones parciales, instancias de observación y práctica en el aula.	Plan de Estudios PCE-2000 [36].

Generar el reporte de exploración inicial

Entradas

- Discurso del cliente - Tabla 5-1
- Conceptos teóricos asociados al dominio - Tabla 5-5
- Estudio de la organización -Tabla 5-6
- Conocimiento adquirido - Tabla 5-8

Salida

- Reporte de exploración inicial - Tabla 5-9

En la Tabla 5-9 se presenta el resultado de la exploración inicial.

Tabla 5-9: Reporte de Exploración Inicial

REPORTE DE EXPLORACIÓN INICIAL			
Analista	Viviana Moschner	**Fecha**	23/08/2018
ID	**Descripción**		
EI-1	Se adquirió conocimiento del funcionamiento de la organización y del problema de negocio. Ver Tabla 5-8.		
EI-2	Se cuenta con el apoyo de las autoridades de la Unidad Académica para la elaboración del trabajo.		
EI-3	Se cuenta con el apoyo de la Dirección de Estadísticas de la UNNE, responsable del SIU Araucano.		
EI-4	Dentro del grupo de interesados hay dos expertos en el dominio con los que se acordó una fluida comunicación. Ver Tabla 5-40.		
EI-5	Se cuenta con el acceso a la base de datos, SIU GUARANI, y apoyo del técnico responsable.		

5.1.1.3 Evaluación de la situación

Identificación de recursos externos

Entradas

- Reporte de exploración inicial - Tabla 5-9
- Análisis de los recursos existentes - Tabla 5-10

Determinar los recursos, humanos y materiales que existen en la organización y con los que se puede contar para la realización del trabajo, Tabla 5-10.

Tabla 5-10: Análisis de Recursos Existentes

ANÁLISIS DE LOS RECURSOS EXISTENTES			
Analista	Viviana Moschner	**Fecha**	26/08/2018
ID	**Recursos Humanos**		
RH-1	Dra. Paola Britos	Directora de tesis. Experta en Explotación de Información	
RH-2	Profesora Guadalupe Portillo	Experta	
RH-3	Profesora Soledad Almirón	Experta	
RH-4	Profesor/a Director Departamento Ciencias de la Educación	Interesado	
RH-5	Licenciada Paola Niemes	Responsable Estadísticas UNNE	
RH-6	E.E. y C. Viviana Moschner	Maestranda	
	Datos		
D-1	SIU Guaraní	Base de datos relacional	
D-2	SIU Araucano	Planillas Excel	

Salida

- Reporte de recursos externos - Tabla 5-11

Tabla 5-11: Reporte de Recursos Externos

REPORTE DE RECURSOS EXTERNOS				
Analista	Viviana Moschner		**Fecha**	26/08/2018
ID	**Recursos Humanos**			
RHE-1	Dra. Paola Britos	UNRN	Directora de tesis. Experta en Explotación de Información	
RHE-2	Licenciada Paola Niemes	Rectorado - UNNE	Responsable Estadísticas UNNE	

	Datos		
RDE-1	SIU Araucano	Excel	Estadísticas UNNE

Identificación de recursos internos

Entradas

- Reporte de exploración inicial -Tabla 5-9
- Análisis de los recursos existentes - Tabla 5-10

Salida

- Reporte de recursos internos - Tabla 5-12

Tabla 5-12: Reporte de Recursos Internos

REPORTE DE RECURSOS INTERNOS			
Analista	Viviana Moschner	**Fecha**	26/08/2018
ID	**Recursos Humanos**		
RHI-1	Profesora Guadalupe Portillo	HUM-UNNE	Experta
RHI-2	Profesora Soledad Almirón	HUM-UNNE	Experta
RHI-3	E.E. y C. Viviana Moschner	FACENA-UNNE	Maestranda
	Datos		
RDI-1	SIU Guaraní	HUM-UNNE	Base de datos relacional

Identificación de las suposiciones del proyecto

- Reporte de exploración inicial - Tabla 5-9
- Reporte de recursos externos - Tabla 5-11
- Reporte de recursos internos - Tabla 5-12

Salida

- Reporte de las suposiciones del proyecto - Tabla 5-13

Tabla 5-13: Suposiciones del Proyecto

SUPOSICIONES DEL PROYECTO			
Analista	Viviana Moschner	**Fecha**	10/08/2018

ID Suposición	**Descripción**
S-1	Los datos almacenados en la base de datos, en el período 2010 a 2018, son correctos y se tendrá acceso a los mismos.
S-2	Se cuenta con el apoyo de los interesados que pertenecen a la organización.
S-3	Se cuenta con la colaboración de los responsables de los Sistemas de Información.
S-4	La carga de los datos se ha realizado correctamente.
S-5	Los atributos rescatados brindan información fiable.

Identificación de riesgos del proyecto

Entradas

- Reporte de exploración inicial - Tabla 5-9
- Reporte de recursos externos - Tabla 5-11
- Reporte de recursos internos - Tabla 5-12

Salida

- Reporte de riesgos del proyecto – Tabla 5-14

Se caracterizan los posibles riesgos, de acuerdo a las particularidades del problema de negocio, Tabla 5-14.

Tabla 5-14: Reporte de Riesgos del Proyecto

REPORTE DE RIESGOS DEL PROYECTO			
Analista	Viviana Moschner	**Fecha**	12/07/2018
ID	**Descripción del Riesgo**		
RP-1	Se presentan dificultades debido al escaso conocimiento y experiencia del equipo en Explotación de la información.		

Definición del plan de contingencia

Entradas

- Reporte de exploración inicial - Tabla 5-9
- Reporte de riesgos del proyecto - Tabla 5-14

Salida

- Plan de contingencia - Tabla 5-15

Determinar las acciones a tomar para subsanar los riegos planteados, Tabla 5-15.

Tabla 5-15: Plan de Contingencia del Proyecto

PLAN DE CONTINGENCIA PARA RIESGO DEL PROYECTO			
Analista	Viviana Moschner	**Fecha**	12/11/2018
ID Plan de Contingencia	**Descripción**		
PC-1	Se propone capacitar en la disciplina a los técnicos del equipo de trabajo.		

Determinación de la viabilidad del proyecto

Definir la posibilidad de éxito del trabajo, detallando el nivel de factibilidad del mismo.

Entradas

- Reporte de exploración inicial - Tabla 5-9
- Reporte de recursos externos - Tabla 5-11
- Reporte de recursos internos - Tabla 5-12

Salida

- Reporte de viabilidad - Tabla 5-17

A continuación, en la Tabla 5-16, se indican las preguntas utilizadas para evaluar la viabilidad del problema a resolver [8].

Tabla 5-16: Preguntas asociadas a la caracterización

Categoría	ID	Pregunta asociada a la característica	Valor
D A T O S	P1	¿En qué medida los repositorios disponibles poseen datos actuales?	mucho
	P2	¿Qué tan representativos son los datos de los repositorios disponibles para resolver el problema de negocio?	mucho
	A1	¿En qué medida los repositorios se encuentran disponibles en formato digital?	todo
	A2	¿Qué cantidad de atributos y registros tienen los datos disponibles?	mucho
	A3	¿Cuánta confianza se posee en la credibilidad de los datos disponibles?	regular
	E1	¿Cuánto facilita la tecnología de los repositorios disponibles las tareas de manipulación de los datos?	mucho
P R O P	P3	¿Cuánto se entiende del problema de negocio?	mucho

	A4	¿En qué medida el problema de negocio no puede ser resuelto aplicando técnicas estadísticas tradicionales?	poco
	A5	¿Qué tan estable es el problema de negocio durante el desarrollo del proyecto?	mucho
TIPO DE PROYECTO	E2	¿Cuánto apoyan los interesados (stakeholders) al proyecto?	mucho
	E3	¿En qué medida la planificación del proyecto permite considerar la realización de buenas prácticas ingenieriles con el tiempo adecuado?	poco
EQUIPO DE TRABAJO	P4	¿Qué nivel de conocimientos posee el equipo de trabajo sobre Explotación de Información?	regular
	E4	¿Qué nivel de experiencia posee el equipo de trabajo en proyectos similares?	regular

En la Tabla 5-17 se muestran los cálculos de viabilidad de la situación planteada.

Tabla 5-17: Reporte de Viabilidad

REPORTE DE EVALUACIÓN DE VIABILIDAD												
Analista		Viviana Moschner							**Fecha**		22/05/2019	
Datos						**Problema de negocio**			**Tipo de Proyecto**		**Equipo de trabajo**	
P1	**P2**	**A1**	**A2**	**A3**	**E1**	**P3**	**A4**	**A5**	**E2**	**E3**	**P4**	**E4**
M	M	T	M	R	M	M	P	M	M	P	M	R
Peso												
8	9	4	7	8	6	7	10	9	8	7	6	6
Umbral												
P	P	P	P	P	N	P	P	P	N	N	P	N

Dimensiones		Viabilidad global	Resultado
Plausibilidad	7,2	**5,94**	**viable**
Adecuación	5,27		
Éxito	5,14		

El proyecto es viable, de acuerdo al valor global de viabilidad obtenido. Además, al analizar los valores dados para cada una de las dimensiones y sus umbrales, se puede afirmar que es factible de realizar, es adecuado en relación a los objetivos planteados y tiene una alta probabilidad de eficacia.

5.1.1.4 Definición del ciclo de vida

Selección del ciclo de vida

Entrada

- Objetivos del proyecto - Tabla 5-38
- Suposiciones del proyecto - Tabla 5-13
- Problemas del negocio - Tabla 5-39
- Reporte de riesgos del proyecto - Tabla 5-14
- Reporte de recursos externos - Tabla 5-11
- Reporte de recursos internos - Tabla 5-12
- Reporte de alternativas de ciclo de vida - Tabla 5-18

Tabla 5-18: Alternativas de Ciclo de Vida

ALTERNATIVAS DE CICLO DE VIDA			
Analista	Viviana Moschner	**Fecha**	02/09/2018
ID	**Modelo**	**Característica**	
A-1	Modelo en cascada	Cada fase se inicia cuando se termina la anterior.	
A-2	Modelo iterativo	Es la iteración de varios ciclos en cascada.	
A-3	Modelo en espiral	Acepta que los requerimientos cambien en cualquier momento.	
A-4	Modelo de desarrollo incremental	Se realiza construyendo módulos que cumplen las diferentes funciones.	
A-5	Modelo CRISP DM	El ciclo de vida de esta metodología, define las principales fases de un PEI, sus tareas y las relaciones entre las mismas.	

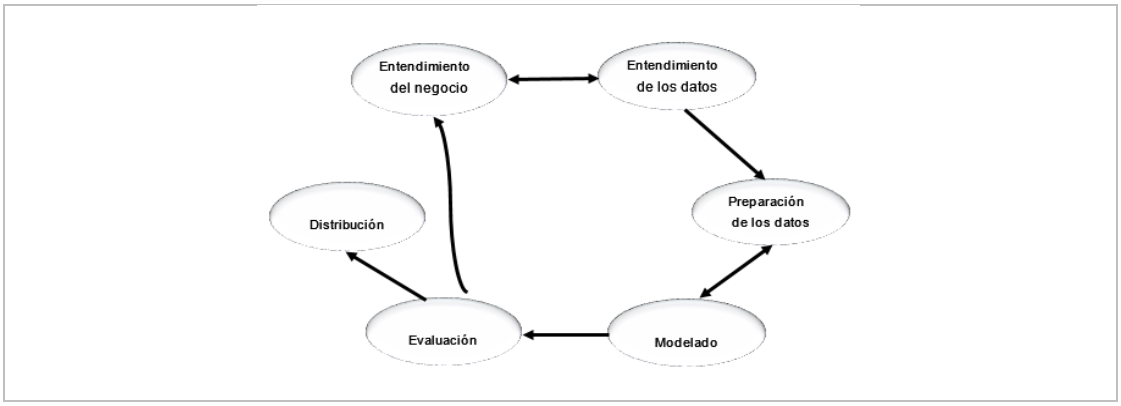
Salida

- Modelo de ciclo de vida - Tabla 5-19

En la Tabla 5-19, se presenta el ciclo de vida seleccionado.

Tabla 5-19: Modelo de Ciclo de Vida

MODELO DE CICLO DE VIDA			
Analista	Viviana Moschner	**Fecha**	12/09/2018
A5	**Modelo CRISP DM**		



5.1.2 Fase: Planificación del proyecto

5.1.2.1 Planificación de las actividades

Definir las actividades asociadas al proyecto

Entrada

- Reporte exploración inicial - Tabla 5-9
- Ciclo de vida seleccionado - Tabla 5-19

Salida

- Mapa y Calendario de Actividades - Tabla 5-20

Tabla 5-20: Mapa y Calendario de Actividades

MAPA Y CALENDARIO DE ACTIVIDADES			
Analista	Viviana Moschner	**Fecha**	02/11/2018
Fase	Actividad	Fecha Inicio	Fecha Fin
Subproceso de Gestión			
Iniciación	Definición de la comunicación	07/07/2018	12/07/2018
	Exploración de conceptos iniciales	08/07/2018	30/07/2018
	Evaluación de la situación	17/07/2018	23/08/2018
	Definición del ciclo de vida	02/09/2018	12/10/2018
Planificación	Planificación de actividades	02/11/2018	10/03/2019
	Planificación de recursos	02/11/2018	10/07/2019
	Estimaciones y responsabilidades	02/03/2019	20/08/2019
Soporte	Gestión del ciclo de vida	28/03/2018	12/09/2018
	Gestión del desarrollo		
	Gestión de la configuración	01/04/2020	30/10/2020

Gestión de entrega	Formalización del cierre del proyecto		
Subproceso de Desarrollo			
E. del dominio	Análisis del dominio	01/07/2018	22/07/2018
	Comprensión del problema del negocio	04/07/2018	30/07/2018
E. de los datos	Análisis de los datos	01/08/2018	10/11/2018
	Exploración de los datos	10/11/2018	30/03/2019
	Evaluación de los datos	01/05/2019	30/08/2019
Modelado	Modelado de los datos	01/04/2020	01/09/2020
	Configuración del modelo	01/09/2020	20/10/2020
Preparación de los datos	Construcción de la fuente temporaria de los datos	30/11/2019	12/03/2020
Implementación	Configuración de la implementación	06/04/2020	06/07/2020
	Implementación del modelo	12/10/2020	10/11/2020
Evaluación y presentación	Evaluación de los resultados	10/11//2020	12/03/2021
	Presentación de los resultados	10/09/2021	

Identificar las métricas a realizar

Entrada

- Reporte exploración inicial - Tabla 5-9

Salida

- Listado de Métricas - Tabla 5-21

Tabla 5-21: Listado de Métricas

LISTADO DE MÉTRICAS

Analista	Viviana Moschner	**Fecha:**	02/11/2019
ID	**Descripción**	**Tipo**	
NT	Tablas a ser consideradas para el proyecto.	Datos iniciales	13
NA	Número inicial de atributos a ser considerados para el proyecto.	Datos iniciales	65
NR	Número inicial de registros a ser considerados para el trabajo.	Datos iniciales	1900
NVN	Total de valores nulos o faltantes.	Exploración de datos	152
NVE	Valores erróneos o fuera del rango normal.	Calidad de datos	80
NANI	Cantidad de atributos nuevos que se deben construir en la tabla única para el proyecto de Explotación de Información.	Construcción de datos	8

Estimación del proyecto

Se definen ocho factores de costos para ser evaluados:

Tipo de objetivo de Explotación de Información (OBTY)

- Analiza el objetivo del proyecto a partir del tipo de proceso de Explotación de Información a ser aplicado.

Tabla 5-22: Valores del Factor de Costo OBTY

Tipo de objetivo de Explotación de Información (OBTY)	
Valor	**Descripción**
1	Se desea conocer los atributos que caracterizan el comportamiento o la descripción de una clase ya conocida.
2	Se desea dividir los datos disponibles en grupos sin poseer una clasificación conocida previamente.
3	Se desea conocer los atributos que caracterizan a grupos sin poseer una clasificación conocida previamente.
4	**Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre un comportamiento o la identificación de una clase conocida.**
5	Se desea conocer los atributos que poseen mayor frecuencia de incidencia sobre la identificación de una clase desconocida previamente.

Grado de apoyo de los miembros de la organización (LECO)

- El grado de apoyo y participación de los miembros de la organización se analiza por cada nivel. Se debe considerar en qué medida los participantes están dispuestos a asistir al equipo de trabajo.

Tabla 5-23: Valores del Factor de Costo LECO

Grado de apoyo de los miembros de la organización (LECO)	
Valor	**Descripción**
1	**Tanto los directivos como el personal poseen buena disposición para colaborar en el proyecto.**
2	Solo los directivos poseen buena disposición para colaborar en el proyecto mientras que el personal es indiferente al proyecto.
3	Solo la alta gerencia posee buena disposición para colaborar en el proyecto mientras que la gerencia media y el personal es indiferente.
4	Solo la alta gerencia posee buena disposición para colaborar en el proyecto pero la gerencia media no desea colaborar.

Cantidad y tipo de repositorios disponibles (AREP)

- Aquí se analizan las fuentes de datos disponibles. Interesa saber la cantidad de repositorios como la tecnología en la que se encuentran implementadas.

Tabla 5-24: Valores de Factor de Costo AREP

Cantidad y tipo de repositorios disponibles (AREP)	
Valor	**Descripción**
1	**Solo 1 repositorio disponible.**
2	Entre 2 y 4 repositorios con tecnología compatible para la integración.
3	Entre 2 y 4 repositorios con tecnología no compatible para la integración.
4	Más de 5 repositorios con tecnología compatible para la integración.
5	Más de 5 repositorios con tecnología no compatible para la integración.

Cantidad de tuplas disponibles en la tabla principal (QTUM)

- Este factor de costo evalúa la cantidad de registros disponibles en la tabla principal utilizada a ser utilizada en el proceso de explotación de la información.

Tabla 5-25: Valores del Factor de Costo QTUM

Cantidad de tuplas disponibles en la tabla principal (QTUM)	
Valor	**Descripción**
1	Hasta 100 tuplas en la tabla principal.
2	Entre 101 y 1.000 tuplas en la tabla principal.
3	**Entre 1.001 y 20.000 tuplas en la tabla principal.**
4	Entre 20.001 y 80.000 tuplas en la tabla principal.
5	Entre 80.01 y 5.000.000 tuplas en la tabla principal.
6	Más de 5.000.000 tuplas en la tabla principal.

Cantidad de tuplas disponibles en las tablas auxiliares (QTUA)

- Esta variable considera la cantidad aproximada de registros disponibles en las tablas auxiliares utilizadas para agregar complementaria a la tabla principal.

Tabla 5-26: Valores del Factor de Costo QTUA

Cantidad de tuplas disponibles en tablas auxiliares (QTUA)	
Valor	**Descripción**
1	No se utilizan tablas auxiliares.
2	**Hasta 1.000 tuplas en las tablas auxiliares.**
3	Entre 1.001 y 50.000 tuplas en las tablas auxiliares.
4	Más de 50.000 tuplas en las tablas auxiliares.

Nivel de conocimiento sobre los datos (KLDS)

- Este factor se refiere al nivel de documentación existente sobre los repositorios de datos. Se debe analizar si existen documentos que expliquen la tecnología utilizada, los atributos que componen las tablas y la forma en que los datos son creados, modificados o borrados.

Tabla 5-27: valores del Factor de Costo KLDS

Nivel de conocimiento sobre los datos (KLDS)	
Valor	**Descripción**
1	**Todas las tablas y repositorios están correctamente documentadas.**
2	Más del 50 % de los repositorios y tablas están correctamente documentados y existen expertos en los datos disponibles para explicarlos.
3	Menos del 50 % de los repositorios y tablas están correctamente documentados pero existen expertos en los datos disponibles para explicarlos.
4	Las tablas y repositorios no están documentadas pero existen expertos en los datos disponibles para explicarlos.
5	Las tablas y repositorios no están documentadas y existen expertos en los datos pero no están disponibles para explicarlos.
6	Las tablas y repositorios no están documentadas y no existen expertos en los datos para explicarlos.

Nivel de conocimiento y experiencia del equipo (KEXT)

- Este factor analiza la capacidad del equipo de trabajo para llevar a cabo el proyecto. El equipo debería tener un mínimo de experiencia y conocimiento en el desarrollo de proyectos de explotación de la información. No obstante, pueden o no poseer

experiencia en proyectos con objetivos similares, dentro del mismo tipo de negocio o bien usando datos similares.

Tabla 5-28: Valores del Factor de Costo KEXT

Nivel de conocimiento y experiencia del equipo (KEXT)	
Valor	**Descripción**
1	**El equipo ha trabajado en tipos de organizaciones y con datos similares para obtener los mismos objetivos.**
2	El equipo ha trabajado en tipos de organizaciones similares pero con datos diferentes para obtener los mismos objetivos.
3	El equipo ha trabajado en otros tipos de organizaciones y con datos similares para obtener los mismos objetivos.
4	El equipo ha trabajado en otros tipos de organizaciones y con datos diferentes para obtener los mismos objetivos.
5	El equipo ha trabajado tipos de organizaciones diferentes, con datos diferentes y otros objetivos.

Funcionalidad de las herramientas disponibles (TOOL)

- Esta variable evalúa las características de las herramientas disponibles para ser aplicadas en el proyecto.

Tabla 5-29: Valores del Factor de Costo TOOL

Funcionalidad de las herramientas disponibles (TOOL)	
Valor	**Descripción**
1	La herramienta posee funciones tanto para el formateo e integración de los datos (permitiendo importar más de una tabla de datos) como para aplicar a las técnicas de Minería de Datos.
2	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de Minería de Datos, y permite importar más de una tabla de datos en forma independiente.
3	La herramienta posee funciones tanto para el formateo como para aplicar las técnicas de Minería de Datos, pero solo permite importar una tabla de datos.
4	**La herramienta posee funciones solo para aplicar las técnicas de Minería de Datos, y permite importar más de una tabla de datos.**
5	La herramienta posee funciones solo para aplicar las técnicas de Minería de Datos, y solo permite importar una tabla de datos.

La fórmula lineal propuesta para el cálculo del esfuerzo en meses/hombre es:

PEM=0,80*OBTY+1,10*LECO-1,20*AREP-0,30*QTUM-0,70*QTUA+1,80*KLDS-0,90*KEXT+1,86*TOOL -3,30

Tabla 5-30: Estimación del Proyecto

ESTIMACIÓN DEL PROYECTO										
Analista		Viviana Moschner					**Fecha**		06/11/2019	
OBTY	LECO	AREP	QTUM	QTUA	KLDS	KEXT	TOOL	D	G	Total

4	1	1	3	2	1	1	4	5,84	0,876	6,716

Se obtiene para la fase de desarrollo un esfuerzo de 5,84 meses/hombre, como el subproceso de gestión representa un 15% del de desarrollo, el esfuerzo para el mismo es de 0,876 meses/hombre, por ello el total es de 6,716 meses/ hombre.

5.1.2.2 Planificación de los recursos

Planificar la necesidad de recursos

Entrada

- Reporte exploración inicial - Tabla 5-9
- Mapa y calendario de actividades - Tabla 5-20

Salida

- Reporte de recursos requeridos - Tabla 5-31

Tabla 5-31: Reporte de Recursos Requeridos

REPORTE DE RECURSOS REQUERIDOS			
Analista	Viviana Moschner	**Fecha**	02/11/2019
ID	**Recurso**	**Cantidad**	**Descripción**
Recursos Humanos			
RRHH_1	Experto	2	Expertos en el dominio del problema
RRHH_2	Científico de Datos	1	Experta en Ciencia de datos
RRHH_3	Maestranda	1	Data miner
Recursos Materiales			
RRM_1	PC	2	PC con Windows 10 – Modeler -RapidMiner

Planificar la capacitación de RRHH

Entrada

- Reporte exploración inicial - Tabla 5-9
- Mapa y Calendario de actividades - Tabla 5-20
- Reporte de recursos requeridos - Tabla 5-31

Salida

- Plan de capacitación de RRHH - Tabla 5-32

Tabla 5-32: Plan de Capacitación de RRHH

PLAN DE CAPACITACION DE RRHH			
Analista	Viviana Moschner	**Fecha**	02/03/2018
ID	**Descripción**		
C_RRHH_3	La maestranda realizará el curso de Ingeniería de explotación de la información, dictado por expertos en FACENA- UNNE- Marzo-2018		

5.1.2.3 Estimaciones y responsabilidades

Estimar el tiempo de desarrollo del proyecto

Entrada

- Reporte exploración inicial - Tabla 5-9
- Mapa y calendario de actividades - Tabla 5-20
- Reporte de recursos externos - Tabla 5-11
- Objetivos del negocio - Tabla 5-38
- Problemas del negocio - Tabla 5-39

Salida

- Tiempo estimado - Tabla 5-30

Definir las responsabilidades de las partes

Entrada

- Reporte exploración inicial - Tabla 5-9
- Objetivos del negocio - Tabla 5-38
- Problemas del negocio - Tabla 5-39
- Tiempo estimado - Tabla 5-30
- Reporte de riesgos del proyecto - Tabla 5-14
- Reporte de viabilidad - Tabla 5-17

Salida

- Contrato del proyecto - Tabla 5-33

Tabla 5-33: Contrato del Proyecto

CONTRATO DEL PROYECTO

Analista	Viviana Moschner	**Fecha**	25/03/2018
Descripción			
Se toma como contrato del proyecto el permiso proporcionado por el Decano de la Unidad Académica para utilizar los datos de la carrera en estudio para la realización del TFM.			

5.1.3 Fase: Soporte

5.1.3.1 Gestión del ciclo de vida

Formalizar el inicio del ciclo

Entrada

- Mapa y calendario de actividades - Tabla 5-20
- Ciclo de vida seleccionado - Tabla 5-19

Salida

- Reporte de inicio formal del ciclo - Tabla 5-34

Tabla 5-34: Reporte de Inicio Formal del Ciclo

REPORTE DE INICIO FORMAL DEL CICLO			
Analista	Viviana Moschner	**Fecha**	28/03/2018
Descripción			
Se toma como inicio formal la aprobación del Plan de TFM.			

5.1.4 Fase: Gestión de control y calidad

5.1.4.1 Control de los recursos

Controlar la capacitación de RRHH

Entrada

- Plan de capacitación de RRHH - Tabla 5-32

Salida

- Control de capacitación de RRHH - Tabla 5-35

Tabla 5-35: Control de capacitación de RRHH

CONTROL DE CAPACITACION DE RRHH

Analista	Viviana Moschner	**Fecha**	02/06/2018
ID	**Descripción**		
CC_RRHH_3	La maestranda asistió y aprobó el curso de Ingeniería de explotación de la información dictado por expertos en FACENA- UNNE.		

5.1.5 Fase: Gestión de la entrega

5.1.5.1 Formalización del cierre del proyecto

Verificación y validación del proyecto

Entrada

- Reporte de criterios de éxito - Tabla 5-41
- Problemas del negocio - Tabla 5-39

Salida

- Reporte de aceptación - Tabla 5-36

Tabla 5-36: Reporte de Aceptación

REPORTE DE ACEPTACIÓN			
Analista	Viviana Moschner	**Fecha**	24/06/2020
Descripción			
Se realizó una presentación del proyecto al grupo de expertos obteniéndose la aprobación del mismo. Se considera como Reporte de Aceptación la presentación del TFM.			

5.2 Subproceso de desarrollo

El subproceso de desarrollo comprende el conjunto de fases y actividades orientadas a la generación del producto resultante del trabajo, la identificación de patrones relevantes, así como su análisis y comprensión para la generación de piezas de conocimiento de interés que sean de valor para el proceso de toma de decisión [33].

5.2.1 Fase: Entendimiento del dominio

En esta fase se comprenden en detalle las características del dominio del negocio y se analizan los recursos disponibles para el desarrollo del mismo [33].

5.2.1.1 Análisis del dominio

Descripción de la terminología

Entrada

- Reporte de exploración inicial Tabla 5-9

Salida

- Glosario de términos - Tabla 5-37

Se registra en la Tabla 5-37 la terminología específica del dominio que no sea familiar para el equipo de trabajo, favoreciendo la comprensión de los distintos aspectos del negocio y la ejecución del mismo, así como también la interacción con los expertos/clientes (educción de requerimientos, presentación de reportes, etc.). La misma incluye una descripción del término, el tipo (definición, acrónimo o abreviación) y la referencia en la que fue visualizado.

Tabla 5-37: Glosario de Términos – Definiciones

DEFINICIONES			
Analista	Viviana Moschner	**Fecha**	01/07/2018
Término	**Descripción**	**Tipo**	**Referencia**
Activo	Alumno no egresado y que no haya solicitada la baja de la carrera.	definición	Régimen Pedagógico HUM [38].
CS	Consejo Superior	abreviatura	
Desertor	Alumno cuyo estado sea el de alumno Activo No Regular y que no registre actividad académica por un periodo de 2 o más años.	definición	
Desgranamiento	Situación que define a los alumnos que no han abandonado la carrera, pero no continúan con la cohorte de ingreso.	definición	
Egresado	Alumno que completa todas las actividades y requisitos reglamentarios de la carrera a la que pertenece y que ha tramitado su título.	definición	Manual de definiciones conceptuales y operativas – SIU ARAUCANO [41].

Excepción	Acto administrativo por el cual el alumno No Regular solicita la permanencia en la UA, explicando el motivo por el que no pudo cumplir con el requisito de permanencia.	definición	Resolución CS, de Permanencia en la UNNE [39].
No Regular	Alumno que no cumple con los requisitos de permanencia.	definición	Régimen Pedagógico HUM [38].
Pasivo	Alumno que haya solicitado la baja o bien no haya cumplido con la documentación obligatoria para el ingreso.	definición	Régimen Pedagógico HUM [38].
Permanencia	Según Resolución de C.S., permanecen los alumnos que hayan aprobado al menos 2 materias en el año académico anterior o bien hayan solicitado y se les haya otorgado la Readmisión.	definición	Resolución CS, de Permanencia en la UNNE [39].
Readmisión	Acto administrativo por el cual el alumno No Regular solicita su permanencia en la UA.	definición	Régimen Pedagógico HUM [38].
Regular	Alumno que cumple con los requisitos de permanencia según la Resolución de CS.	definición	Resolución CS, de Permanencia en la UNNE [39].
Rezagado	Alumno que no aprobó el total de las materias según lo establecido en el plan, pero aún continúa realizando actividad académica.	definición	

Identificación de objetivos

Entradas

- Discurso del cliente - Tabla 5-1
- Reporte de exploración inicial - Tabla 5-9

El principal objetivo de este trabajo es desarrollar un modelo predictivo que permita detectar situaciones potenciales de fracaso académico, deserción o desgranamiento en la población estudiantil de la carrera Profesorado en Ciencias de la Educación de la Facultad de Humanidades de la UNNE, con el fin de diseñar políticas educacionales estratégicas para minimizar el problema en estudio.

Salida

- Objetivos del trabajo - Tabla 5-38

Tabla 5-38: Objetivos del Negocio

OBJETIVOS DEL TRABAJO			
Analista	Viviana Moschner	**Fecha**	08/07/2018
ID Objetivo	**Descripción**		**Referencia**
O-1	El principal objetivo de este proyecto es desarrollar un modelo predictivo que permita detectar situaciones potenciales de deserción y desgranamiento en la población estudiantil de la carrera Profesorado en Ciencias de la Educación de la Facultad de Humanidades de la UNNE (2010-2018)		Reunión N° 1
O-2	Identificar características personales y/o académicas que influyen en el abandono de los alumnos de la carrera Profesorado en Ciencias de la Educación (2010-2018).		Reunión N° 1
O-3	Identificar factores personales y/o académicos que influyen en el rezago o desgranamiento de los alumnos de la carrera Profesorado en Ciencias de la Educación (2010-2018).		Reunión N° 1
O-4	Determinar atributos que influyen en la permanencia de los alumnos de la carrera en estudio.		Reunión N° 1

5.2.1.2 Comprensión del problema de negocio

Identificación de los problemas de negocio

Entradas

- Discurso del cliente - Tabla 5-1
- Estudio de la organización - Tabla 5-6
- Reporte de exploración inicial - Tabla 5-9
- Objetivos del negocio -Tabla 5-38

Salida

- Problemas del negocio - Tabla 5-39

Tabla 5-39: Problemas del Negocio

PROBLEMAS DEL NEGOCIO

Analista	Viviana Moschner	**Fecha**	08/07/2018
ID Problema	**Descripción**		**Referencia**
La Facultad de Humanidades es una Unidad Académica dependiente de la UNNE en la que se dictan 14 carreras de grado, 3 de ellas con título de pregrado.			
P-1	Existe una gran desproporción entre los ingresos y egresos anuales.		Reunión N° 1
P-2	La duración real de la carrera difiere mucho de la duración establecida en el plan de estudios.		Reunión N° 1
P-3	Un alto porcentaje de la población estudiantil está rezagada respecto a la cohorte a la que pertenece.		Reunión N° 1

Identificación de los expertos en el problema

Entrada

- Problemas de negocio - Tabla 5-39

Salida

- Listado de expertos en el problema de negocio - Tabla 5-40

Tabla 5-40: Expertos en el Problema de Negocio

EXPERTOS EN EL PROBLEMA DE NEGOCIO					
Analista	Viviana Moschner		**Fecha**	12/07/2018	
Posición	**Organización/ Sector**	**Rol en el Proyecto**	**Proceso de Negocio Asociado**	**Datos Contacto**	
				Nombre	**Email**
Asesoría Pedagógica	Hum	Experto	Entendimiento del Dominio	Guadalupe Portillo	asesoria@hum.unne.edu.ar
Asesoría Pedagógica	Hum	Experto	Entendimiento del Dominio	Soledad Almirón	asesoria@hum.unne.edu.ar

Identificación de los criterios de éxito del problema de negocio

Entradas

- Reporte de exploración inicial - Tabla 5-9
- Problemas de negocio - Tabla 5-39
- Listado de expertos en el problema de negocio - Tabla 5-40

Salida

- Criterios de éxito del problema de negocio - Tabla 5-41

Tabla 5-41: Criterios de Éxito del Problema de Negocio

CRITERIOS DE ÉXITO DEL PROBLEMA DE NEGOCIO			
Analista	Viviana Moschner	**Fecha**	22/07/2018
ID Criterio	**Descripción**	**Objetivo del Proyecto**	
CE-1	Detectar atributos personales que llevan al alumno a abandonar la carrera.	O-1	
CE-2	Identificar asignaturas que deben cursar más de una vez para poder regularizar o promocionar.	O-1	
CE-3	Identificar la/s asignatura/s que desaprueban con mayor frecuencia.	O-1	
CE-4	Detectar atributos personales que influyen en los alumnos rezagados.	O-2	
CE-5	Determinar patrones de comportamiento en los alumnos rezagados y/o posibles desertores.	O-2	

5.2.2 Fase: Entendimiento de los datos

5.2.2.1 Análisis de los datos

Descripción de las fuentes de datos

Esta actividad tiene como objetivo la obtención de los datos y la familiarización con la información obtenida. En algunos casos, en particular cuando el volumen de datos no es muy grande, se puede trabajar con los datos originales [30]. En esta propuesta se trabajará con los datos obtenidos del SIU GUARANI.

El consorcio SIU, en [43], define al SIU GUARANI como un Sistema que registra las actividades de la gestión académica dentro de la universidad desde que el alumno se inscribe hasta que egresa. El mismo tiene como objetivo la administración de las tareas en forma óptima y segura, con la finalidad de obtener información consistente para los niveles operativos y directivos.

El Sistema de Información Universitaria es una base de datos de tipo relacional, es decir, es una colección de elementos de datos con relaciones predefinidas entre ellos. Estos elementos se organizan como un conjunto de tablas con columnas y filas. Las tablas se utilizan para guardar información sobre los objetos que se van a representar en la base de datos. Cada columna de una tabla guarda un determinado tipo de datos y un campo almacena el valor real de un atributo. Las filas de la tabla representan una recopilación de valores relacionados de un objeto o entidad. Cada fila de una tabla podría marcarse con un identificador único denominado clave principal, mientras que filas de varias tablas pueden relacionarse con

claves foráneas. Se puede obtener acceso a estos datos de muchas formas distintas sin reorganizar las propias tablas de la base de datos.

Tabla 5-42: Repositorio de Datos

REPOSITORIO DE DATOS					
Analista	Viviana Moschner		**Fecha**		01/11/2018
ID Repositorio	**Nombre**	**Tipo**	**Descripción**	**Procesos afectados**	**Responsable**
BD-1	SIU GUARANI 2.9.5	Base de Datos de tipo relacional, motor INFORMIX	Base de datos utilizada en la UA desde el 2003, a la misma se migran en ese año los datos primarios con los que se trabajaban hasta el momento.		Viviana Moschner
BD-2	SIU ARAUCANO	Planillas Excel			Viviana Moschner

Descripción de las tablas

Se detallan en la Tabla 5-43 características de las tablas que utilizaremos para rescatar datos.

Tabla 5-43: Descripción de Tablas

DESCRIPCION DE TABLAS			
Analista	Viviana Moschner	**Fecha**	06/11/2018
ID Tabla	**Nombre**	**Descripción**	**Responsable**
T-1	sga_personas	En esta tabla se graban los datos del alumno que no sufrirán cambios con el tiempo. Ej. Fecha y lugar de nacimiento, datos de los progenitores.	
T-2	sga_datos_censales	En esta tabla se registran datos como domicilio de procedencia, domicilio de residencia, estudios cursados por los padres, situación laboral del alumno y de los padres.	
T-3	sga_dat_cen_aux	Contiene datos como año de egreso del secundario.	
T-4	sga_dat_cen_aux2	Esta contiene datos referentes a forma de costear los estudios.	
T-5	sga_alumnos	Se registran aquí las carreras a las que se inscribió el estudiante y la cohorte a la que pertenece.	
T-6	sga_cursadas	Esta tabla contiene datos referentes a las materias regularizadas y promocionadas.	
T-7	vw_hist_academica	Esta es una vista de las asignaturas aprobadas y desaprobadas con datos tales como fecha, calificación, acta.	
T-8	sga_perd_regul	Se registran en esta las veces que el alumno perdió su condición de regular, el motivo y la fecha en que se lo rehabilito como alumno regular.	
T-9	sga_det_perd_regul	Se registran en esta las veces que el alumno solicito excepción para obtener la condición de regular.	
T-10	sga_eval_parc_alum	Se registran las calificaciones de las evaluaciones parciales.	

T-11	sga_insc_cursadas	Esta tabla contiene las inscripciones al cursado.
T-12	sga_atrib_mat_plan	Atributos de las asignaturas del plan (si se puede promocionar, rendir libre, si es pro mediable, año de cursado).
T-13	sga_titulos_otorg	Se registra en esta tabla a los alumnos que egresaron con datos como fecha de egreso, promedio académico y general, título obtenido, fecha de inicio del trámite.

Identificar campos asociados al problema de negocio

En la Tabla 5-44 se detallan los campos asociados al problema del negocio.

Tabla 5-44: Campos Asociados al Negocio

CAMPOS ASOCIADOS AL NEGOCIO			
Analista	Viviana Moschner	**Fecha**	10/11/2018
ID Campo	**Nombre Campo**	**Tipo**	**Descripción**
C-1	calidad	alfanumérico	Código que indica calidad del alumno
C-2	carrera	alfanumérico	Código carrera
C-3	cohorte	numérico	Cohorte
C-4	colegio_secundario	numérico	Contiene el código del colegio
C-5	sexo	numérico	Código que indica género del alumno
C-6	fecha_nacimiento	date	Fecha nacimiento formato DD/MM/AAAA.
C-7	nacionalidad	numérico	Nacionalidad
C-8	localidad_col_sec	numérico	Contiene código de la localidad del colegio
C-9	provincia_col_sec	numérico	Contiene código de la provincia del colegio
C-10	sector_col_sec	alfanumérico	Sector del colegio
C-11	titulo secundario	numérico	Contiene código del título y con el mismo se extrae el nombre del mismo
C-12	anio_egreso_sec	numérico	Año de egreso (AAAA)
C-13	estado_civil	numérico	Código de estado civil
C-14	vive_con	numérico	Código indica con quien vive el alumno
C-15	loc_per_lect	numérico	Contiene el código de la localidad en la que reside, se extrae el nombre de la misma

C-16	loc_proc	numérico	Contiene el código de la localidad de la que procede, se extrae el nombre de la misma
C-17	existe_trab_alum	numérico	Situación laboral del alumno
C-18	fliares_cargo_alum	numérico	Número de familiares a cargo
C-19	hora_sem_trab_alum	numérico	Rango de horas semanales de trabajo
C-20	rel_trab_carrera	numérico	Nivel de relación del trabajo con la carrera
C-21	ult_est_cur_padre	numérico	Nivel de estudios alcanzado por el padre
C-22	ult_est_cur_madre	numérico	Nivel de estudios alcanzado por la madre
C-23	alu_est_civil_uh	numérico	Alumno estado civil unido de hecho
C-24	alu_trab_desjub	alfanumérico	Alumno está desocupado o jubilado
C-25	alu_trab_fami	alfanumérico	Alumno trabaja con la familia
C-26	alu_cos_est_ap_fam	alfanumérico	Costea sus estudios con el aporte de la familia
C-27	alu_cost_est_trab	alfanumérico	Costea sus estudios con su trabajo
C-28	alu_cost_est_beca	alfanumérico	Costea sus estudios con beca
C-29	alu_cost_est_plsoc	alfanumérico	Costea sus estudios con planes sociales
C-30	alu_cost_est_otra	alfanumérico	Costea sus estudios de otra forma
C-31	tiene_beca	alfanumérico	Recibe beca
C-32	alu_beca_muni	alfanumérico	Recibe beca municipal
C-33	alu_beca_prov	alfanumérico	Recibe beca provincial
C-34	alu_beca_tipo_eco	alfanumérico	Recibe beca tipo económica
C-35	alu_beca_tipo_ser	alfanumérico	Recibe beca de servicio
C-36	alu_beca_tipo_inv	alfanumérico	Recibe beca tipo investigación
C-37	alu_beca_eco_tran	alfanumérico	Recibe beca para transporte

C-38	alu_beca_eco_come	alfanumérico	Recibe beca para comedor
C-39	alu_beca_eco_habi	alfanumérico	Recibe beca para vivienda
C-40	tiene_beca_univ	alfanumérico	Recibe beca de la universidad
C-41	tiene_beca_inter	alfanumérico	Recibe beca intercambio
C-42	existe_trab_alum	numérico	Trabaja el alumno
C-43	cant_hijos_alum	numérico	Número de hijos
C-44	remuneración	alfanumérico	El alumno recibe remuneración por su trabajo
C-45	padre_vive	alfanumérico	Vive el padre
C-46	madre_vive	alfanumérico	Vive la madre
C-47	fecha ingreso	date	Fecha de ingreso a la carrera
C-48	regular	alfanumérico	Es regular
C-49	sede	numérico	Sede a la que pertenece
C-50	plan	alfanumérico	Año del plan de estudios
C-51	total readmisiones	numérico	Número de readmisiones
C-52	fecha_u_reincorp	date	Fecha de la última readmisión o excepción otorgada
C-53	total_inscripciones_cursado	numérico	Número de inscripciones a cursar
C-54	fecha_ultima_reg	date	Fecha de la última regularidad
C-55	fecha_ultima_aprob	date	Fecha de la última materia aprobada
C-56	fecha_ultima_desaprobada	date	Fecha de la última materia desaprobada
C-57	total_aprobadas	numérico	Total de asignaturas aprobadas
C-58	materia	alfanumérico	Código de la materia
C-59	fecha_regularidad	date	Fecha en que regularizó la materia

C-60	fecha aprobó	date	Fecha de aprobación
C-61	forma aprobación	alfanumérico	Forma de aprobación
C-62	inscripciones_a_cursar_mat	numérico	Número de inscripciones a cursar la materia
C-63	reprobo_mat	numérico	Número de veces que reprobó la asignatura
C-64	parciales_aprobo	numérico	Número de parciales aprobados de la materia
C-65	parciales_reprobo	numérico	Número de parciales desaprobados de la materia
	desde C-58 a C-65 para cada una de las 35 asignaturas		

5.2.2.2 Exploración de los datos

Integrar los datos en un medio digital

La tarea consiste en recolectar, explorar, verificar la calidad y cantidad de la información obtenida. Se confecciono un script para recolectar los datos personales de las distintas tablas relacionadas y otro para recabar datos de la situación académica de los alumnos de las cohortes en estudio.

Explorar los datos

Se describen en la Tabla 5-45, características de los datos de interés.

Tabla 5-45: Reporte de Datos Explorados

REPORTE DE DATOS EXPLORADOS			
Analista	Viviana Moschner	**Fecha**	19/11/2018
Atributo		**Tipo Dato**	
calidad		alfanumérico	A- Activo P- Pasivo E- Egresado
carrera		alfanumérico	01 a 34
colegio CUE		numérico	Código Único de Establecimiento
estado civil		alfanumérico	1- Soltero 2- Casado 3- Separado 4- Divorciado

		6- Viudo
es remunerado	alfanumérico	S- Si N- No
excepciones	numérico	Cantidad de excepciones otorgadas
familiares a cargo	alfanumérico	Número de familiares a cargo
fecha ingreso	date	DD/MM/AAAA Año desde 2010 a 2018
fecha nacimiento	date	DD/MM/AAAA
hijos	numérico	Cantidad de hijos
horas de trabajo	alfanumérico	1- hasta 20 horas 2- de 21 a 35 horas 3- de 36 o más horas
localidad colegio	alfanumérico	Nombre de la localidad del colegio
localidad procedencia	alfanumérico	Nombre de la localidad de la que procede
localidad residencia	alfanumérico	Nombre de la localidad en la que reside
madre vive	alfanumérico	S- Sí N- No D- Desconoce
nacionalidad	numérico	1- Argentino 2- Extranjero 3- Naturalizado 4- Argentino por opción
nombre colegio	alfanumérico	Nombre del colegio
nombre título	alfanumérico	Nombre del titulo
padre vive	alfanumérico	S- Sí N- No D- Desconoce
provincia colegio	alfanumérico	Nombre de la provincia del colegio
provincia procedencia	alfanumérico	Nombre de la provincia de la que procede
provincia residencia	alfanumérico	Nombre de la provincia en la que reside
readmisiones	numérico	Cantidad de readmisiones otorgadas
regular	alfanumérico	S- Sí N- No
relación del trabajo con la carrera	alfanumérico	1- Total 2- Parcial 3- No relacionada
sector colegio	alfanumérico	E- Estatal P- Privado
sede	alfanumérico	00000 Resistencia 00001 Gral. Pinedo 00002 Corrientes

Se inscribió anteriormente en otra carrera	alfanumérico	S- Sí N- N
tipo residencia	numérico	1- Casa 2- Departamento 3- Pensión/residencia 4- Otros
titulo secundario código	numérico	Código del título
trabaja	alfanumérico	1- Trabaja 2- No trabaja y buscó trabajo 3- No trabaja y no buscó
trabaja en negocio familiar	alfanumérico	S- Sí N- No
últimos estudios de la madre	numérico	1- No hizo estudios 2- Escuela primaria incompleta 3- Escuela primaria completa 4- Escuela secundaria incompleta 5- Escuela secundaria completa 6- Estudio superior incompleto 7- Estudio superior completo 10- Estudio universitario incompleto 11- Estudio universitario completo 12- Estudio de posgrado
últimos estudios del padre	numérico	1- No hizo estudios 2- Escuela primaria incompleta 3- Escuela primaria completa 4- Escuela secundaria incompleta 5- Escuela secundaria completa 6- Estudio superior incompleto 7- Estudio superior completo 10- Estudio universitario incompleto 11- Estudio universitario completo 12- Estudio de posgrado
unido de hecho	alfanumérico	S- Sí N- No
vive con	numérico	1- Solo 2- Con compañeros 3- Con familia de origen 4- Con su pareja e hijos 5- Otros
Costea sus estudios:		

con ayuda familiar	alfanumérico	S- Sí N- No
con beca	alfanumérico	S- Sí N- No
con otra	alfanumérico	S- Sí N- No
con planes sociales	alfanumérico	S- Sí N- No
con su trabajo	alfanumérico	S- Sí N- No
tiene beca	alfanumérico	S- Sí N- No
Beca:		
comedor	alfanumérico	S- Sí N- No
intercambio	alfanumérico	S- Sí N- No
investigación	alfanumérico	S- Sí N- No
nacional	alfanumérico	S- Sí N- No
municipal	alfanumérico	S- Sí N- No
otra	alfanumérico	S- Sí N- No
provincial	alfanumérico	S- Sí N- No
tipo económica	alfanumérico	S- Sí N- No
tipo servicio	alfanumérico	S- Sí N- No
transporte	alfanumérico	S- Sí N- No
universitaria	alfanumérico	S- Sí N- No
vivienda	alfanumérico	S- Sí N- No
Situación académica:		
regularizo_2010	numérico	Número de materias que regularizó en el año 2010
aprobo_2010	numérico	Número de materias que aprobó en el año 2010
promociono_2010	numérico	Número de materias que promocionó en el año 2010
desaprobo_2010	numérico	Número de materias que desaprobó en 2010
Se repiten los últimos 4 (cuatro) hasta 2018		
inscribio_CEN01	numérico	Número de inscripciones a cursar
reg_CEN01	date	Fecha en que regularizó
aprobo_CEN01	date	Fecha en la que aprobó la asignatura

reprobo_CEN01	numérico	Número de veces que reprobó la actividad
parciales_A_CEN01	numérico	Número de evaluaciones parciales aprobadas de la materia
parciales_R_CEN01	numérico	Número de evaluaciones parciales reprobadas de la materia
Se repite esta información para cada una de las materias del plan		

5.2.2.3 Evaluación de los datos

Verificación de la calidad de los datos

Para este trabajo se usan exclusivamente los datos provenientes de las diferentes tablas del SIU GUARANI. Utilizando lenguaje SQL se desarrolló un script para rescatar datos personales, laborales y económicos de los alumnos de la carrera en estudio, obteniéndose una tabla maestra con los datos originales. En la Tabla 5-45 se detallan los datos obtenidos. Además se elaboró un SP para recabar los datos académicos de los alumnos y se los almacenó temporalmente en una planilla Excel. Los atributos de las planillas son analizados, seleccionados, limpiados y transformados para la obtención del set de datos.

Es de primordial importancia que los atributos seleccionados sean relevantes para la tarea de Minería de Datos. Por ello, por ejemplo, se decidió no incluir el atributo que corresponde al número de documento, legajo, apellido y nombres del alumno debido a que el algoritmo utilizado podría obtener un modelo falto de generalidad.

Entrada

- Campos asociados al negocio - Tabla 5-44
- Reporte de datos explorados - Tabla 5-45

Salida

- Reporte de la calidad de los datos - Tabla 5-46

Tabla 5-46: Reporte de Calidad de los Datos

REPORTE DE CALIDAD DE LOS DATOS			
Analista	Viviana Moschner	**Fecha**	30/06/2019

	Nº Nulos	Nº **Atípicos**
colegio secundario	54	
costea estudios	34	13
edad ingreso		18
estudios padre	127	
estudios madre	176	
título secundario	8	24

Identificación de campos riesgosos

Entrada

- Campos asociados al negocio - Tabla 5-44
- Reporte de datos explorados - Tabla 5-45
- Reporte de la calidad de los datos - Tabla 5-46

Salida

- Reporte de tipo de campos riesgosos - Tabla 5-47

Tabla 5-47: Reporte de Tipos Campos Riesgosos

REPORTE DE TIPO DE CAMPOS RIESGOSOS			
Analista	Viviana Moschner	**Fecha**	29/06/2019
Haciendo un análisis de los datos de interés para el trabajo se encontraron los siguientes inconvenientes:			
ID		**Descripción**	
TCR-1	Campos nulos	Campos nulos/vacíos de algunas de las variables seleccionadas.	
TCR-2	Atributo muy específico	Datos muy específicos referentes a una característica.	
TCR-3	Varios atributos se refieren a uno	Hay más de un atributo que se refiere a la misma característica del alumno.	

- Reporte de campos riesgosos - Tabla 5-48

Tabla 5-48: Reportes de Campos Riesgosos

REPORTE DE CAMPOS RIESGOSOS

Analista	Viviana Moschner	**Fecha**	30/06/2019

ID	**Tipo riesgo**	**Campo**	**Descripción**
CR-1	TCR-3	estado civil	Se refieren al mismo atributo
CR-2	TCR-3	unido de hecho	
CR-3	TCR-3	hijos	Se refieren al mimo atributo
CR-4	TCR-3	familiares a cargo	
CR-5	TCR-2	titulo secundario código	Son muy específicos los títulos secundarios
CR-6	TCR-3	tiene beca	Se refieren al mismo atributo
CR-7	TCR-3	beca universitaria	
CR-8	TCR-3	beca nacional	
CR-9	TCR-3	beca provincial	
CR-10	TCR-3	beca municipal	
CR-11	TCR-3	beca intercambio	
CR-12	TCR-3	otra	
CR-13	TCR-3	beca tipo económica	
CR-14	TCR-3	beca tipo servicio	
CR-15	TCR-3	beca Investigación	
CR-16	TCR-3	beca transporte	
CR-17	TCR-3	beca vivienda	
CR-18	TCR-3	beca comedor	
CR-19	TCR-3	trabaja	Se pueden combinar y obtener un solo atributo
CR-20	TCR-3	trabaja en negocio familiar	
CR-21	TCR-3	es remunerado	

5.2.3 Fase: Modelado

En esta fase se seleccionan las técnicas más apropiadas y se aplican las mismas configurando los parámetros para la obtención de resultados.

5.2.3.1 Modelado del problema

Definir el problema de explotación de la información

Entrada

- Problemas de negocio - Tabla 5-39
- Campos asociados al negocio - Tabla 5-44
- Reporte de datos explorados - Tabla 5-45
- Reporte de campos riesgosos - Tabla 5-48

Salida

- Problemas de explotación de la información - Tabla 5-49

Tabla 5-49: Problemas de Explotación de la Información

PROBLEMAS DE EXPLOTACIÓN DE LA INFORMACION			
Analista	Viviana Moschner	**Fecha**	08/07/2018
ID Problema	**Descripción**		**Referencia**
PE-1	Agrupar a estudiantes de acuerdo a calidad (Activo-Pasivo – Egresado) y situación de permanencia (Regular – No Regular).		Reunión N° 1
PE-2	Descubrir reglas en base a los datos personales de los alumnos que identifiquen el comportamiento de los estudiantes Activos No regulares, es decir, que no tienen actividad hace 2 o más años.		Reunión N° 1
PE-3	Descubrir patrones de los estudiantes Activos rezagados que no continúan con la cohorte a la que pertenecen.		Reunión N° 1

Modelar el problema de explotación de la información

Entrada

- Campos asociados al negocio - Tabla 5-44
- Reporte de campos riesgosos - Tabla 5-48
- Problemas de explotación de la información - Tabla 5-49

Salida

- Identificación de la solución - Tabla 5-50

Tabla 5-50: Identificación de la Solución

IDENTIFICACIÓN DE LA SOLUCIÓN			
Analista	Viviana Moschner	**Fecha**	12/03/2019

Como resultado de aplicar la técnica de Derivación de Procesos de Explotación de Información, se decide aplicar el proceso de **Descubrimiento de reglas de pertenencia a grupos**.
Se generan grupos ordenados para obtener el conjunto de reglas que definen el comportamiento de cada grupo identificado.

5.2.3.2 Configuración del modelo

Identificación de las herramientas alternativas

Entrada

- Campos asociados al negocio - Tabla 5-44
- Reporte de datos explorados - Tabla 5-45
- Reporte de campos riesgosos - Tabla 5-48
- Problemas de explotación de la información - Tabla 5-49
- Identificación de la solución - Tabla 5-50

Salida

- Herramientas alternativas - Tabla 5-51

Tabla 5-51: Reporte de Herramientas Alternativas

REPORTE DE HERRAMIENTAS ALTERNATIVAS				
Analista	Viviana Moschner		**Fecha**	12/03/2019
Herramientas	**Weka**	**IBM Modeler**	**RapidMiner**	**Knime**
Origen	La Universidad de Waikato de Nueva Zelanda, en 1993, inició el desarrollo de la versión original de Weka.	La primera versión, desarrollada por la compañía ISL (Integral Solutions Limited), del Reino Unido, se lanzó en 1994. Se llamaba Clementine y estaba basada en Unix. En 1998 SPSS adquirió ISL.	Desarrollada en el año 2001 por la Universidad de Dortmund, Alemania.	Desarrollada originalmente en la Universidad de Constanza, Alemania.
Características	Ofrece varias herramientas para tareas de clasificación, agrupamiento, regresión, asociación y para el pre procesamiento y	Se utiliza para realizar tareas analíticas y construir modelos predictivos.	Se destaca en el análisis predictivo. Esta herramienta se usa a través de bloques que pueden ser arrastrados a áreas de trabajo.	Integra varios componentes para aprendizaje automático y Minería de Datos utilizando el concepto de fraccionamiento de datos modular.

	visualización de datos.			
Lenguaje de programación	Java	Java	Java	Java
Sistema Operativo	Windows, Linux	Windows, Linux	Windows, Linux	Windows, Linux

Entrada

- Reporte de herramientas alternativas - Tabla 5-51
- Evaluación de herramientas - Tabla 5-52

Tabla 5-52: Evaluación de Herramientas

EVALUACIÓN DE HERRAMIENTAS ALTERNATIVAS						
Analista	Viviana Moschner					
Criterios evaluación: 1-Malo, 2- Débil, 3- Bueno, 4-Excelente						
HERRAMIENTAS			WEKA	RAPIDMINER	IBM Modeler	KNIME
1. Características Técnicas						
Soporte de metodología	Soporte del proceso	3	2	3	3	3
Compatibilidad con fuentes de datos	Base de datos	8	2	3	4	3
	Otras fuentes	8	2	3	4	3
Integración	Soporte de distintas técnicas asociadas al proceso de explotación de la información	5	3	3	4	3
Multilenguaje	Soporte de distintos idiomas	2	2	3	3	2
Técnicas	Variedad de técnicas	18	2	2	3	2
Reporte y visualización	Permite generar reportes y visualizaciones	12	2	2	4	2
Multiplataforma	Soporta múltiples plataformas	5	3	3	3	3
Instalación remota	La administración y mantenimiento son remotos	5	2	3	3	3
Usuarios múltiples	Posee perfiles de usuario	2	2	2	3	2
Seguridad	Provee seguridad de la información configurada por perfiles	2	2	2	3	2
Backup	Metodología de backup	2	2	2	3	3
Amigable	Interfaz del usuario	10	2	2	4	2

Configuración	Permite la configuración del perfil	8	2	2	3	2
Documentación	Servicio de soporte y ayuda	5	2	2	4	3
Conexión	Soporta conexión por: Internet, FTP	2	2	2	3	3
Soporte de sistemas de mensajes	Soporta compartir información (por mail u otro medio)	3	2	2	3	2
Total			210	234	348	243
	Peso del grupo	40%	84	93,6	139,2	97,2
1. Características del proveedor						
Características del proveedor	Historia	30	2	3	4	3
Crecimiento	Perspectiva a futuro	10	2	3	4	3
Ubicación	Oficinas	30	2	3	3	2
Implementación	Otras implementaciones de la misma herramienta	5	2	2	4	2
	Contacto con otros clientes	5	2	2	3	2
Confidencialidad	Confidencialidad de la información	20	3	3	3	3
Total			220	290	345	260
	Peso del grupo	25%	55	72,5	86,25	65
2. Características del servicio						
Garantía del producto	Duración y alcance	30	3	2	3	3
Mejora	Brinda soporte a versiones previas	20	2	2	3	2
Licencia	Costo, alcances y soporte postventa	30	2	3	2	3
Soporte	Tiempo de respuesta y disponibilidad	20	2	3	3	2
Total			230	250	270	260
	Peso del grupo	20%	46	50	54	52
3. Características económicas						
Costo del software	Costo de la herramienta	30	2	3	1	3
Costo del hardware	Necesidad de mejorar o comprar nuevo hardware compatible con la herramienta	20	3	3	3	3
Otros costos	Costos adicionales al producto (backup, web servers, bases de datos, etc.)	20	3	3	3	3
Licencias	Política de licencia	10	3	3	2	3
Financiamientos	Existencia	10	3	3	3	3
Mejoras	Costo promedio de la mejora del producto	10	3	3	3	3

Total			270	300	230	300
	Peso del grupo	-15%	-40,5	-45	-34,5	-45
Final						
1. Características Técnicas		40%	84	93,6	139,2	97,2
2. Características del Proveedor		25%	55	72,5	86,25	65
3. Características del Servicio		20%	46	50	54	52
4. Características Económicas		-15%	-40,5	-45	-34,5	-45
TOTAL			144,5	171,1	244,95	169,2

Salida

- Herramientas seleccionadas, [44]. - Tabla 5-53

Tabla 5-53: Herramienta seleccionada

HERRAMIENTAS SELECCIONADAS			
Analista	Viviana Moschner	**Fecha**	12/03/2019
La herramienta **IBM Modeler** se identifica como la más adecuada para el trabajo, seguida de **RapidMiner**.			

Identificación de los algoritmos de Minería de Datos

Entrada

- Campos asociados al negocio - Tabla 5-44
- Problemas de explotación de la información - Tabla 5-49
- Identificación de la solución - Tabla 5-50
- Herramienta seleccionada - Tabla 5-53
- Algoritmos de Minería de Datos soportados - Tabla 5-54

Tabla 5-54: Algoritmos de MD Soportados

ALGORITMOS DE MINERIA DE DATOS SOPORTADOS			
Analista	Viviana Moschner	**Fecha**	29/03/2019
Algoritmos de Clasificación			
Arboles aleatorios	Se utiliza este algoritmo para desarrollar sistemas de clasificación que predicen o clasifican observaciones futuras basándose en un conjunto de reglas de decisión.		
Árbol C&R	Es un método de predicción y clasificación basado en árboles. Comienza por realizar un examen de los campos de entrada para buscar la mejor división. La división define dos subgrupos, que se siguen dividiendo en otros dos subgrupos sucesivamente hasta que se activa un criterio de parada.		

Red Neuronal	Las redes neuronales son redes interconectadas masivamente en paralelo y con organización jerárquica. Las mismas intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico.
C5.0	Este modelo divide la muestra en función del campo que ofrece la máxima ganancia de información, las sub muestras se vuelven a dividir por lo general basándose en otro campo. El proceso se repite hasta que sea imposible otra división. Luego se examinan las divisiones y se eliminan las que no contribuyen significativamente con el valor del modelo.
Algoritmos de Asociación	
Apriori	Es aplicado sobre bases de datos transaccionales. Permite encontrar de forma eficiente conjuntos frecuentes que sirven de base para generar reglas de asociación.
Carma	Este modelo extrae un conjunto de reglas de los datos sin necesidad de especificar campos de entrada ni objetivos, por ello las reglas generadas pueden ser utilizadas en una amplia variedad de aplicaciones.
Algoritmos de Segmentación	
K-medias	Es un método de agrupamiento cuyo objetivo es la partición de un conjunto de observaciones en n grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.
Kohonen	Se trata de un modelo de red neuronal con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro. En este hay neuronas que se organizan en muchas zonas, de forma que las informaciones captadas del entorno a través de los órganos sensoriales se representan internamente en forma de mapas bidimensionales.

Selección de los algoritmos de Minería de Datos

Entrada

- Campos asociados al negocio - Tabla 5-44
- Problemas de explotación de la información - Tabla 5-49
- Identificación de la solución - Tabla 5-50
- Herramienta seleccionada - Tabla 5-53
- Algoritmos de Minería de Datos soportados - Tabla 5-54

Salida

- Algoritmos de Minería de Datos seleccionados - Tabla 5-55

Tabla 5-55: Algoritmos de MD Seleccionados

ALGORITMOS DE MINERIA DE DATOS SELECCIONADOS			
Analista	Viviana Moschner	**Fecha**	29/03/2019
Algoritmos seleccionados			
Luego de realizar el procedimiento de derivación de explotación de la información se decide la utilización de los siguientes algoritmos.			

MODELER	
C5.0	Este modelo divide la muestra en función del campo que ofrece la máxima ganancia de información, las sub muestras se vuelven a dividir por lo general basándose en otro campo. El proceso se repite hasta que sea imposible otra división. Luego se examinan las divisiones y se eliminan las que no contribuyen significativamente con el valor del modelo.
Redes Neuronales	Las redes neuronales son redes interconectadas masivamente en paralelo y con organización jerárquica. Las mismas intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico.
RAPIDMINER	
Rule Induction	Este funciona de forma similar a una regla proposicional de aprendizaje, llamando en forma iterativa incremental una función de reducción de error y así amortiguando los valores de error hasta un 50%.

5.2.4 Fase: Preparación de los datos

5.2.4.1 Construcción de la fuente temporaria de datos

Seleccionar los datos

Entrada

- Campos asociados al negocio - Tabla 5-44
- Reporte de campos riesgosos - Tabla 5-48
- Datos seleccionados - Tabla 5-56

Tabla 5-56: Reporte de Datos Seleccionados

REPORTE DE DATOS SELECCIONADOS			
Analista	Viviana Moschner	**Fecha**	29/11/2019
Atributo		**Tipo Dato**	
calidad		alfanumérico	A- Activo P- Pasivo E- Egresado
carrera		alfanumérico	01 a 34
colegio CUE		alfanumérico	Código del colegio
es remunerado		alfanumérico	S- Sí N- No
estado civil		alfanumérico	1- Soltero 2- Casado 3- Separado 4- Divorciado 6- Viudo
excepciones		numérico	Cantidad de excepciones a la readmisión

familiares a cargo	alfanumérico	S- Sí N- No
fecha ingreso	date	DD/MM/AAAA Año desde 2010 a 2018
fecha nacimiento	date	DD/MM/AAAA
hijos	numérico	Número de hijos
localidad colegio	alfanumérico	Nombre de la localidad del colegio
localidad procede	alfanumérico	Nombre de la localidad de procedencia
localidad reside	alfanumérico	Nombre de la localidad de residencia
madre vive	alfanumérico	S- Si N- No D- Desconoce
nacionalidad	numérico	1- Argentino 2- Extranjero 3- Naturalizado 4- Argentino por opción
nombre título secundario	alfanumérico	Nombre del título
padre vive	alfanumérico	S- Sí N- No D- Desconoce
plan	numérico	1983-1989-1998-2000-2006
readmisiones	numérico	Cantidad de readmisiones obtenidas.
regular	alfanumérico	S- Sí N- No
provincia colegio	alfanumérico	Nombre de la provincia del colegio
provincia procedencia	alfanumérico	Nombre de la provincia de procedencia
provincia residencia	alfanumérico	Nombre de la provincia de residencia
sector colegio	alfanumérico	E- Estatal P- Privado
sede	alfanumérico	00000 Resistencia 00001 Gral. Pinedo 00002 Corrientes
sexo	numérico	1- M 2- F
Se inscribió anteriormente a otra carrera	alfanumérico	S- Sí N- No
tiene beca	alfanumérico	S- Sí N- No
título secundario	numérico	Código del título
trabaja	numérico	1- Trabaja 2- No trabaja y buscó trabajo 3- No trabaja y no buscó
trabaja en negocio familiar	alfanumérico	S- Sí N- No

último nivel de estudios de la madre		1- No hizo estudios 2- Escuela primaria incompleta 3- Escuela primaria completa 4- Escuela secundaria incompleta 5- Escuela secundaria completa 6- Estudio superior incompleto 7- Estudio superior completo 10- Estudio universitario incompleto 11- Estudio universitario completo 12- Estudio de posgrado
último nivel de estudios de la madre	numérico	1- No hizo estudios 2- Escuela primaria incompleta 3- Escuela primaria completa 4- Escuela secundaria incompleta 5- Escuela secundaria completa 6- Estudio superior incompleto 7- Estudio superior completo 10- Estudio universitario incompleto 11- Estudio universitario completo 12- Estudio de posgrado
unido de hecho	alfanumérico	S- Sí N- No
Costea sus estudios:		
con ayuda familiar	alfanumérico	S- Sí N-No
con su trabajo	alfanumérico	S- Sí N- No
con planes sociales	alfanumérico	S- Sí N- No
con beca	alfanumérico	S- Sí N- No
Beca:		
comedor	alfanumérico	S- Si N- No
intercambio	alfanumérico	S- Sí N- No
investigación	alfanumérico	S- Sí N- No
municipal	alfanumérico	S- Sí N- No

nacional	alfanumérico	S- Sí N- No
otra	alfanumérico	S- Sí N- No
provincial	alfanumérico	S- Sí N -No
tipo económica	alfanumérico	S- Sí N- No
tipo servicio	alfanumérico	S- Sí N- No
transporte	alfanumérico	S- Sí N- No
universitaria	alfanumérico	S- Sí N- No
vivienda	alfanumérico	S- Sí N- No
Asignaturas:		
regularizo_2010	numérico	Número de materias que regularizó en el año 2010
aprobo_2010	numérico	Número de materias que aprobó en el año 2010
promociono_2010	numérico	Número de materias que promocionó en el año 2010
desaprobo_2010	numérico	Número de materias que desaprobó en examen
Se repiten los últimos 4 hasta 2018		
Para cada asignatura:		
año_cursado_CEN01	numérico	Año/nivel al que pertenece la materia en el plan de estudios
inscribio_CEN01	numérico	Veces que se inscribió a cursar
reg_CEN01	date	Fecha en que regularizó
aprobo_CEN01	date	Fecha en la que aprobó la asignatura
reprobo_CEN01	numérico	Veces que reprobó la actividad
parciales_A_CEN01	numérico	Número de parciales que aprobó
parciales_R_CEN01	numérico	Número de parciales que reprobó
Se repite esta información para cada una de las materias del plan		

Definir la fuente temporaria de datos

Entrada

- Campos asociados al negocio - Tabla 5-44

- Reporte de campos riesgosos - Tabla 5-48
- Algoritmos de Minería de Datos seleccionados - Tabla 5-54

Salida

- Descripción de la fuente temporaria de datos - Tabla 5-57

Tabla 5-57: Descripción de la Fuente Temporaria de Datos

DESCRIPCIÓN DE LA FUENTE TEMPORARIA DE DATOS			
Analista	Viviana Moschner	**Fecha**	29/11/2019
Se obtienen 2 conjuntos de datos usando la herramienta SQL, se almacena cada set en una planilla Excel diferente.			
ID	**Descripción**	**Nº Filas**	**Nº Columnas**
E-Pers	Contiene datos personales.	1280	51
E-Acad	Inscripciones a cursadas, resultado de evaluaciones parciales, de cursadas, fecha de aprobación, número de exámenes desaprobados de cada una de las 35 asignaturas del plan de estudios.	1280	247

- Reporte de transformación de datos -Tabla 5-58

Tabla 5-58: Reporte de Transformación de Datos

REPORTE DE TRANSFORMACIÓN DE DATOS			
Analista	Viviana Moschner	**Fecha**	29/11/2019
ID	**Dato**	**Dato Transformado**	
T-1	Se combinan campos estado civil y unido de hecho.	estado civil	
T-2	Se genera el atributo que indica si tiene beca, sin especificar el origen, partiendo de los datos referidos a becas.	tiene beca	
T-3	Se combinaron datos familiares y cantidad de hijos.	familiares a cargo	
T-4	Se crea el atributo título secundario, agrupándolos con el objetivo de que no sean tan específicos.	título secundario	
T-5	Con la fecha de nacimiento se calcula la edad al ingresar a la carrera en estudio.	edad al ingresar	
T-6	Se calcula la edad al momento de extraer los datos.	edad al extraer datos	
T-7	Con los datos localidad de residencia y de procedencia se construye un atributo que nos indica si son iguales o difieren.	proc res	
T-8	Con los datos referidos a cómo costea sus estudios, tiene beca y trabaja se genera un nuevo atributo.	costea estudios	

T-9	Con la localidad en la que está ubicado el colegio secundario se crea el atributo que indica si es capital de provincia o interior.	colegio capital interior
T-10	Utilizando los estudios obtenidos por los padres se crea el atributo que indica si el alumno es la primera generación universitaria.	generación universitaria
T-11	Con los datos trabaja, trabaja en negocio familiar, es remunerado se genera el atributo trabaja.	trabaja
T-12	Con los datos del último examen rendido, última materia cursada, última readmisión/excepción se genera el atributo año última actividad que nos indica hasta que año estuvo activo.	año última actividad
T-13	Este campo contiene el año de la última materia regularizada.	año última regularizada
T-14	Con el número de inscripciones a una asignatura se construye el atributo que nos indica si recursó la materia.	inscribió_CENXX
T-15	Con este dato se crea el atributo que nos permite saber si el alumno rindió alguna de las evaluaciones parciales.	rindió_CENXX
T-16	Este dato indica la fecha en que aprobó la actividad CENXX.	aprobó_CENXX

Trabajando con los expertos en el dominio del negocio, se decide clasificar al conjunto de datos en base a los atributos calidad, número de inscripciones a cursar, número de asignaturas aprobadas, años de inactividad. Los datos quedan clasificados según la Tabla 5-59.

Tabla 5-59: Generación del Atributo Target

DESCRIPCIÓN DE GENERACION ATRIBUTO DESTINO- TARGET			
Analista	Viviana Moschner	**Fecha**	29/11/2019
ID	**Descripción**		
E	Calidad es igual a E-Aprobó todas las asignaturas e inició trámite de título.		Egresado
SA	Se inscribió a la carrera, pero no al cursado y tampoco rindió examen final.		Sin Actividad
R	Alumno con actividad académica reciente pero no acorde a la cohorte a la que pertenece.		Rezagado
PD	Alumno con actividad académica, pero no realiza actividad hace 2 años o más.		Posible Desertor

Generar la fuente temporaria de datos

Entrada

- Campos asociados al negocio - Tabla 5-44

- Descripción de la fuente temporaria de datos - Tabla 5-57
- Reporte de transformación de datos - Tabla 5-58

Salida

- Reporte de fuente temporaria de datos - Tabla 5-60

Tabla 5-60: Reporte de Fuente Temporaria de Datos

REPORTE DE FUENTE TEMPORARIA DE DATOS			
Analista	Viviana Moschner	**Fecha**	29/11/2019
Con los dos conjuntos de datos almacenados en planillas Excel, E-per y E-acad, se genera una fuente de datos única, FTD1			
Atributo		**Tipo Dato**	
calidad		alfanumérico	A- Activo P- Pasivo E- Egresado
carrera		alfanumérico	07
cohorte		numérico	2010 a 2018
colegio CUE		alfanumérico	Código del colegio
colegio capital o interior		numérico	1- Capital 2- Interior
costea estudios		numérico	1- Con ayuda familiar 2- Con su trabajo 3- Con su trabajo y ayuda familiar 4- Con su trabajo y beca 5- Con ayuda familiar y beca 6- Con planes sociales 7- Con becas
edad al ingresar a la carrera		numérico	xx
egreso secundario		numérico	xxxx
estado civil		alfanumérico	1- Soltero 2- Casado 3- Separado 4- Divorciado 6- Viudo 7- Unido de hecho
estudios madre		numérico	1 a 7, 10 a 12
estudio padre		numérico	1 a 7, 10 a 12
familiares a cargo		numérico	S- Sí N- No

género	numérico	1- Masculino 2- Femenino
inactivo	numérico	xx
generación universitaria	alfanumérico	Primer univ No es primer univ
madre vive	alfanumérico	S- Sí N- No D- Desconoce
nacionalidad	numérico	1- Argentino 2- Extranjero 3- Naturalizado 4- Argentino por opción
padre vive	alfanumérico	S- Sí N- No D- Desconoce
procedencia es igual a residencia	numérico	1- Sí 2- No
regular	alfanumérico	S- Sí N- No
sector colegio	alfanumérico	E- Estatal P- Privado
está en otra carrera	alfanumérico	S- Sí N- No
título secundario	numérico	1- Bachiller 2- Educación polimodal 3- Perito 4- Técnico 5- Otro
total inscripciones a cursar	numérico	Número de inscripciones al cursado
total aprobadas	numérico	Total de materias aprobadas
total regularizadas		
trabaja	alfanumérico	Trabaja No trabaja
ultima actividad	numérico	2010 a 2018
Asignaturas		
aprobo_1	numérico	Número de materias que aprobó en el 1º año
aprobó_2	numérico	Número de materias que aprobó en el 2º año
aprobó_3	numérico	Numero de materias que aprobó en el 3º año
regularizo_1	numérico	Numero de materias que regularizó en el 1º año
regularizo_2	numérico	Número de materias que regularizó en el 2º año
regularizo_3	numérico	Número de materias que regularizo en el 3º año
reprobo_1	numérico	Número de materias que desaprobó en el 1º año

reprobo_2	numérico	Número de materias que desaprobó en el 2º año
reprobo_3	numérico	Número de materias que desaprobó en el 3º año

5.2.4.2 Adecuación de la fuente temporaria de datos

Limpiar los datos

Para realizar esta tarea, hacemos usos de la pestaña de estadísticas de RapidMiner y del algoritmo auditoría de datos de la herramienta Modeler. Esto nos ayudará a identificar datos nulos, en blanco y outliers. Con los datos obtenidos podemos corregir los posibles errores y de esta manera aumentar la eficacia de los análisis a realizar.

En el caso de valores faltantes una de las técnicas recomendables es la de sustituir los mismos por la media de los datos.

Entrada

- Campos asociados al negocio - Tabla 5-44
- Reporte de calidad de los datos - Tabla 5-46
- Reporte de campos riesgosos - Tabla 5-48
- Reporte de fuente temporaria de datos - Tabla 5-60

Salida

- Reporte de limpieza de datos - Tabla 5-61

Tabla 5-61: Reporte de Limpieza de Datos

REPORTE DE LIMPIEZA DE DATOS			
Analista	Viviana Moschner	**Fecha**	06/04/2020
Se normalizan los datos fuera de rango, se imputan los datos en blanco o nulos, reemplazándolos por la media. No se modifica la estructura de la fuente temporaria de datos FT1.			

Formatear los datos

Entrada

- Descripción de los campos asociados al negocio - Tabla 5-44
- Reporte de fuente temporaria de datos - Tabla 5-60

Salida

- Reporte de datos transformados - Tabla 5-62

Tabla 5-62: Reporte de Datos transformados

REPORTE DE DATOS TRANSFORMADOS			
Analista	Viviana Moschner	**Fecha**	06/04/2020
Para la herramienta Modeler, el formato de los datos es el correcto, no necesitan adaptación.			

5.2.5 Fase: Implementación

5.2.5.1 Configuración de la implementación

Configurar algoritmos de Minería de Datos

Entrada

- Problemas de Explotación de Información - Tabla 5-49
- Identificación de la solución - Tabla 5-50
- Herramientas seleccionadas - Tabla 5-53

Salida

- Reporte de configuración de algoritmos Modeler - Tabla 5-63

Tabla 5-63: Reporte de Configuración de Algoritmos Modeler

REPORTE DE CONFIGURACION DE ALGORITMOS			
Analista	Viviana Moschner	**Fecha**	10/04/2020
MODELER			
Algoritmo	**Modelo**	**Reglas de parada**	**Valores perdidos en predictores**
RN	Perceptron multicapa	Precisión mínima 90%	Imputar perdidos
	Tipo de resultado	**Modo**	**Evaluación Modelo**
C5.0	Conjunto de reglas	Favorecer precisión	Calcular importancia del predictor

- Reporte de configuración de algoritmos RapidMiner - Tabla 5-64

Tabla 5-64: Reporte de Configuración de algoritmos RapidMiner

REPORTE DE CONFIGURACION DE ALGORITMOS			
Analista	Viviana Moschner	**Fecha**	

RAPIDMINER			
Algoritmo	**Criterio**	**Pureza**	**Beneficio mínimo de poda**
Rule Induction	Information gain	0.9	0.25

Entrenar algoritmos de Minería de Datos

Entrada

- Identificación de la solución - Tabla 5-50
- Herramientas seleccionadas - Tabla 5-53
- Algoritmos de Minería de Datos seleccionados - Tabla 5-55
- Reporte de fuente temporaria de datos - Tabla 5-60
- Reporte de configuración de algoritmos Modeler - Tabla 5-63
- Reporte de configuración de algoritmos – RapidMiner - Tabla 5-64

Salida

- Reporte de entrenamiento del algoritmo selección de características Modeler- Tabla 5-65
- Reporte de entrenamiento del algoritmo Redes Neuronales Modeler - Tabla 5-66
- Reporte de entrenamiento del algoritmo C5.0 Modeler - TABLA 5-66

  En primera instancia se utiliza el algoritmo selección de características, generándose un filtro con los atributos que tienen una importancia mayor a 0,667, Figura 5-1 y Figura 5-2.

Tabla 5-65: Reporte de Entrenamiento del Algoritmo Selección de Características

REPORTE DE ENTRENAMIENTO DEL ALGORITMO SELECCIÓN DE CARACTERÍSTICAS			
Analista	Viviana Moschner	**Fecha**	12/04/2020

ID	Datos	Tipo
SELECCIÓN DE CARACTERÍSTICAS		
Calidad Real	Calidad real	Entrada
Colegio capital interior	El colegio secundario está ubicado en capital de provincia o interior	Entrada
Costea estudios	Modo de costear estudios	Entrada
Está en otra carrera	Se inscribió a otra carrera con anterioridad	Entrada
Edad al ingresar a la carrera	Edad al ingresar a la carrera	Entrada
Egreso sec	Año de egreso del nivel secundario	Entrada
Estado civil	Estado civil del alumno	Entrada
Fliares a cargo	Número de familiares a cargo	Entrada
Estudios madre	Último nivel de estudios de la madre	Entrada
Estudios padre	Último nivel de estudios del padre	Entrada
Género	Género del alumno	Entrada
Generación universitaria	¿Es la primera generación universitaria?	Entrada
Inactivo	Años de inactividad	Entrada
Madre vive	Vive la madre	Entrada
Nacionalidad	Nacionalidad del alumno	Entrada
Padre vive	Vive el padre	Entrada
Res proc	¿Procedencia es igual a residencia?	Entrada
Título sec	Tipo de título secundario	Entrada
Total aprobadas	Total de materias aprobadas	Entrada
Total regularizadas	Total de materias regularizadas	Entrada
Trabaja	Trabaja	Entrada
Ultima actividad	Año en que realizó la última actividad	Entrada

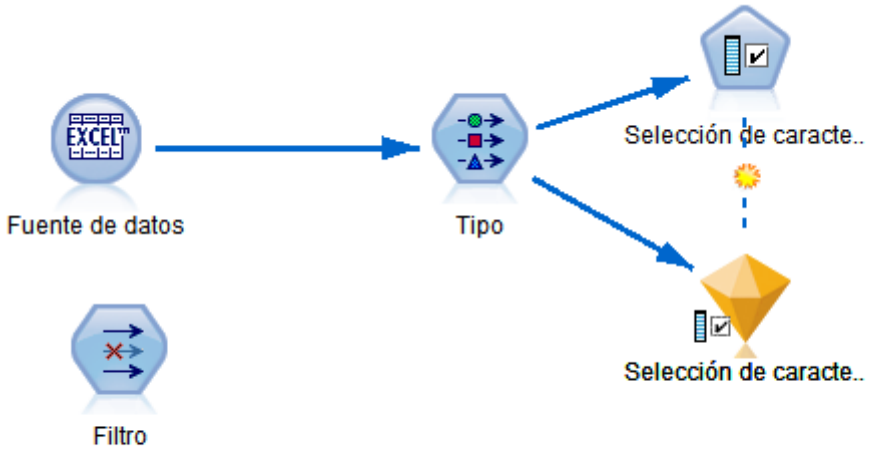
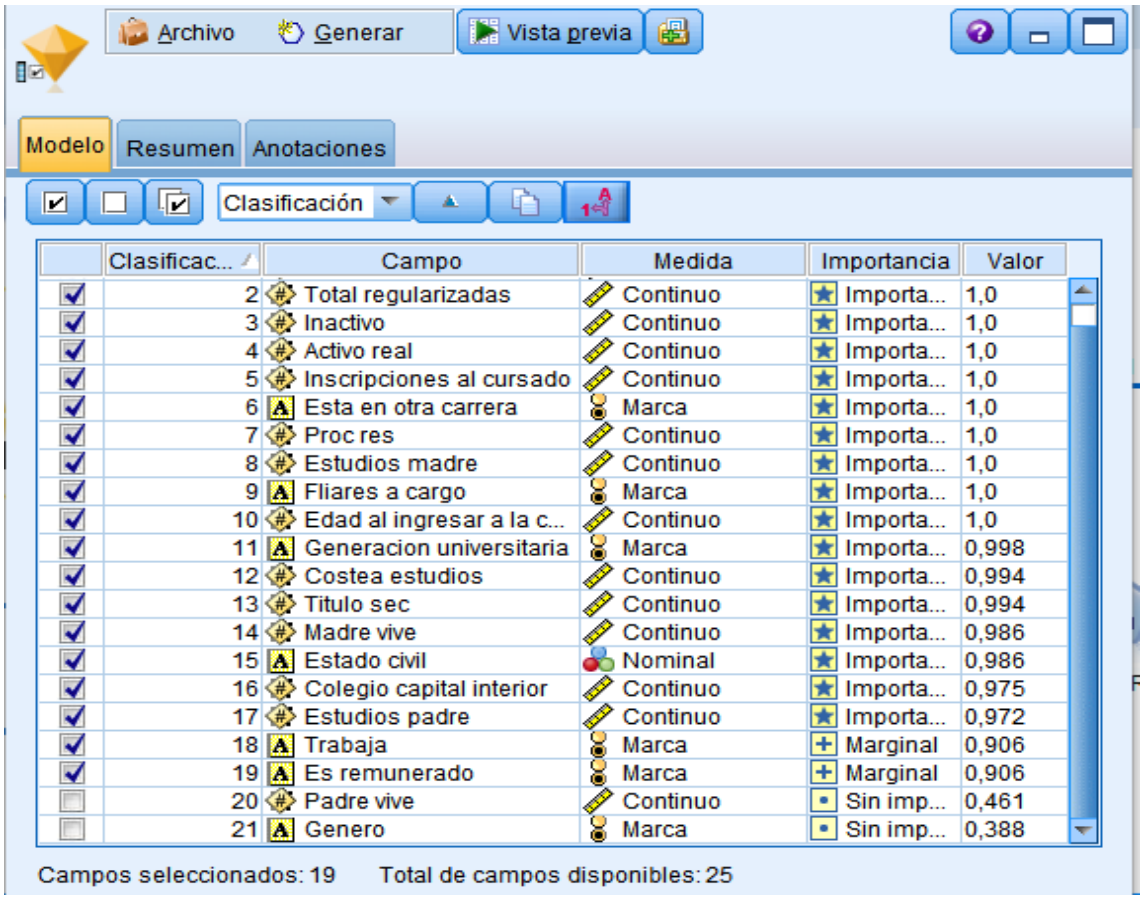


Figura 5-1: Selección de características



	Clasificac...	Campo	Medida	Importancia	Valor
☑	2	Total regularizadas	Continuo	Importa...	1,0
☑	3	Inactivo	Continuo	Importa...	1,0
☑	4	Activo real	Continuo	Importa...	1,0
☑	5	Inscripciones al cursado	Continuo	Importa...	1,0
☑	6	Esta en otra carrera	Marca	Importa...	1,0
☑	7	Proc res	Continuo	Importa...	1,0
☑	8	Estudios madre	Continuo	Importa...	1,0
☑	9	Fliares a cargo	Marca	Importa...	1,0
☑	10	Edad al ingresar a la c...	Continuo	Importa...	1,0
☑	11	Generacion universitaria	Marca	Importa...	0,998
☑	12	Costea estudios	Continuo	Importa...	0,994
☑	13	Titulo sec	Continuo	Importa...	0,994
☑	14	Madre vive	Continuo	Importa...	0,986
☑	15	Estado civil	Nominal	Importa...	0,986
☑	16	Colegio capital interior	Continuo	Importa...	0,975
☑	17	Estudios padre	Continuo	Importa...	0,972
☑	18	Trabaja	Marca	Marginal	0,906
☑	19	Es remunerado	Marca	Marginal	0,906
☐	20	Padre vive	Continuo	Sin imp...	0,461
☐	21	Genero	Marca	Sin imp...	0,388

Figura 5-2: Selección de características (Modeler)

Las técnicas de modelado requieren que no haya valores blancos o nulos. Se podría optar por eliminar las filas con atributos ausentes, descartar las variables con esa característica, imputar los valores ausentes, utilizando algún algoritmo o bien reemplazar estos ausentes con la media del atributo para el conjunto. Con IBM SPSS - Modeler se decidió imputar los valores utilizando el algoritmo CRT, generándose un supernodo de valores perdidos, Figura 5-4.

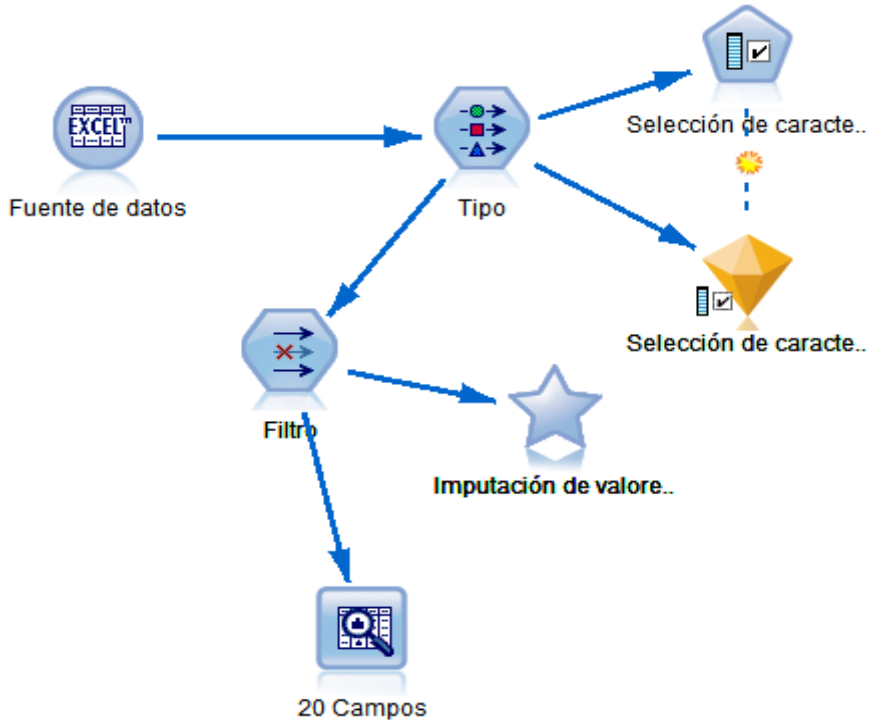


Figura 5-3: Auditoría de datos

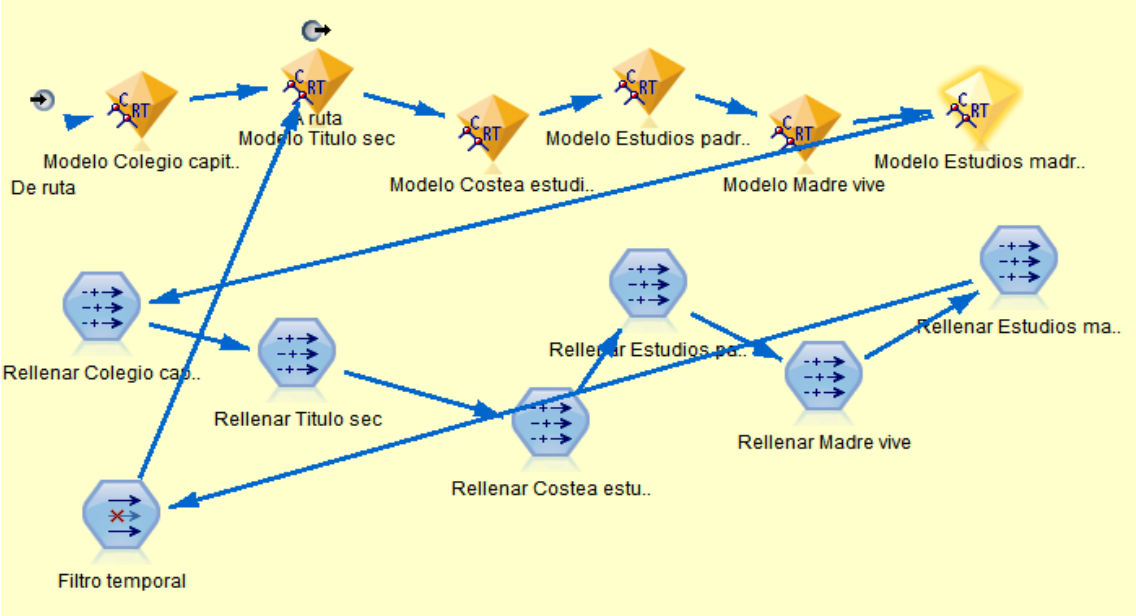


Figura 5-4: Supernodo de valores perdidos

Finalmente se aplica el algoritmo Redes Neuronales con los atributos de entrada y de salida que se especifican en la Tabla 5-66.

Tabla 5-66: Reporte de Entrenamiento del algoritmo Redes Neuronales

REPORTE DE ENTRENAMIENTO DEL ALGORITMO REDES NEURONALES			
Analista	Viviana Moschner	**Fecha**	12/04/2020

ID	Datos	Tipo
Calidad Real	Calidad real	Destino -Target
Colegio capital interior	El colegio secundario está ubicado en capital de provincia o interior	Entrada
Costea estudios	Modo de costear estudios	Entrada
Está en otra carrera	Se inscribió a otra carrera con anterioridad	Entrada
Edad al ingresar a la carrera	Edad al ingresar a la carrera	Entrada
Estado civil	Estado civil del alumno	Entrada
Fliares a cargo	Número de familiares a cargo	Entrada
Es remunerado	¿Recibe remuneración?	Entrada
Estudios madre	Último nivel de estudios de la madre	Entrada
Estudios padre	Último nivel de estudios del padre	Entrada
Generación universitaria	¿Es la primera generación universitaria?	Entrada
Inactivo	Años de inactividad	Entrada
Madre vive	Vive la madre	Entrada
Res proc	¿Procedencia es igual a residencia?	Entrada
Título sec	Tipo de título secundario	Entrada
Total aprobadas	Total de materias aprobadas	Entrada
Total regularizadas	Total de materias regularizadas	Entrada
Trabaja	Trabaja	Entrada

Al aplicar el algoritmo de clasificación supervisada RN, se obtiene una precisión del 96,6% en la clasificación, Figura 5-5.

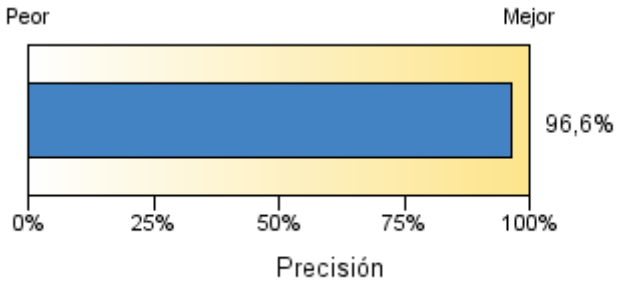


Figura 5-5: Precisión de la clasificación

Además se obtiene la siguiente Matriz de confusión, Figura 5-6.

Clasificación para Calidad _R

Porcentaje correcto global = 96,7%

Observado	Predichos		
	E	PD	R
E	98,4%	0,0%	1,6%
PD	0,0%	96,9%	3,1%
R	2,8%	2,8%	94,3%

Porcentaje de filas
100,00
80,00
60,00
40,00
20,00
0,00

Figura 5-6: Matriz de confusión (%)

Clasificación para Calidad _R

Porcentaje correcto global = 96,7%

Observado	Predichos		
	E	PD	R
E	183	0	3
PD	0	381	12
R	5	5	166

Porcentaje de filas
100,00
80,00
60,00
40,00
20,00
0,00

Figura 5-7: Matriz de confusión (recuento de casillas)

Los valores localizados en la diagonal principal de la matriz de confusión son las clasificaciones correctas y en la que se observa el mayor número de registros, esto indica una clasificación eficaz. Mientras que los valores de la diagonal secundaria representan los errores entre las clases, como ejemplo se observa en la primera fila de la matriz en donde el modelo clasifica por error al 1,6 % en la clase R cuando deberían pertenecer al grupo E.

Los datos clasificados son utilizados para generar el conjunto de reglas para cada grupo. Para realizar la validación interna de los modelos predictivos, se utiliza la estrategia de dividir

aleatoriamente el conjunto de datos, el 70% se utiliza para desarrollar el modelo y el 30 % restante para validarlo.

Tabla 5-67: Reporte de Entrenamiento del Algoritmo C5.0 - Modeler

REPORTE DE ENTRENAMIENTO DEL ALGORITMO C5.0			
Analista	Viviana Moschner	**Fecha**	12/04/2020
ID	**Datos**	**Tipo**	
C5.0			
$N-Calidad Real	Calidad real	Destino -Target	
Colegio capital interior	El colegio secundario está ubicado en capital de provincia o interior	Entrada	
Costea estudios	Modo de costear estudios	Entrada	
Está en otra carrera	Se inscribió a otra carrera con anterioridad	Entrada	
Edad al ingresar a la carrera	Edad al ingresar a la carrera	Entrada	
Egreso sec	Año de egreso del nivel secundario	Entrada	
Estado civil	Estado civil del alumno	Entrada	
Fliares a cargo	Número de familiares a cargo	Entrada	
Es remunerado	¿Es remunerado?	Entrada	
Estudios madre	Último nivel de estudios de la madre	Entrada	
Estudios padre	Último nivel de estudios del padre	Entrada	
Generación universitaria	¿Es la primera generación universitaria?	Entrada	
Madre vive	Vive la madre	Entrada	
Res proc	¿Procedencia es igual a residencia?	Entrada	
Título sec	Tipo de título secundario	Entrada	
Trabaja	Trabaja	Entrada	

Resultados para el campo de resultado Calidad _R

Comparando $N-Calidad _R con Calidad _R

'Partición'	1_Entrenamiento		2_Comprobación	
Correctos	487	96,06%	235	94,38%
Erróneos	20	3,94%	14	5,62%
Total	507		249	

Matriz de coincidencias para $N-Calidad _R (las filas muestran las reales)

'Partición' = 1_Entrenamiento	E	PD	R	$null$
E	129	0	4	0
PD	2	253	3	1
R	6	4	105	0

'Partición' = 2_Comprobación	E	PD	R
E	52	1	0
PD	0	130	5
R	3	5	53

Evaluación del rendimiento

'Partición' = 1_Entrenamiento	
E	1,278
PD	0,656
R	1,419
'Partición' = 2_Comprobación	
E	1,491
PD	0,567
R	1,316

Figura 5-8: Informe de precisión

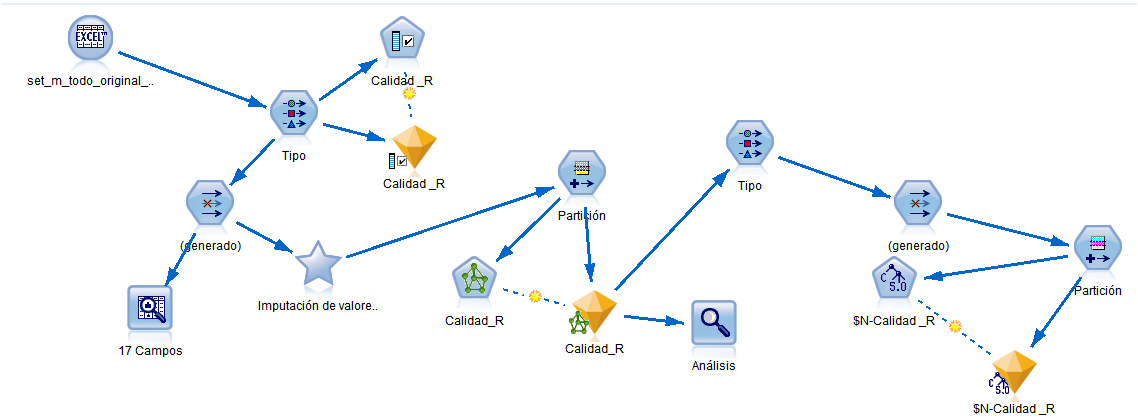


Figura 5-9: Flujo del proceso en IBM SPSS Modeler

Del conjunto de reglas obtenidas se destaca la siguiente para el grupo Egresado: si el alumno se inscribió a otra carrera con anterioridad y no trabaja, entonces es Egresado, Tabla 5-68.

Tabla 5-68: Regla para Egresados con IBM SPSS-Modeler

REGLA PARA CALIDAD_R=E			
Analista	Viviana Moschner	**Fecha**	12/08/2020

Regla	
1	si Se inscribió a otra carrera (antes) = Si y Trabaja = No trabaja entonces E

La regla Nº 4, indica que si es la primera carrera a la que se inscribe en la UA, no tiene familiares a cargo, costea sus estudios con ayuda familiar o con su trabajo, el nivel de estudios del padre es secundario completo, incompleto o menor nivel, entonces es Posible Desertor, Tabla 5-69.

Tabla 5-69: Regla para Posible Desertor con IBM SPSS Modeler

REGLA PARA CALIDAD_R=PD			
Analista	Viviana Moschner	**Fecha**	12/08/2020
Regla			
4	si Se inscribió a otra carrera (antes) = No y Familiares a cargo = No y Estudios padre <= 5 y Costea estudios <= 3 entonces PD		

La regla para el alumno rezagado, establece que si es la primera carrera en la que se inscribe, no tiene familiares cargo, trabaja y el padre tiene estudios universitarios, completos, incompletos o de posgrado, entonces es Rezagado, Tabla 5-70.

Tabla 5-70: Regla para Rezagados con SPSS-Modeler

REGLA PARA CALIDAD_R=R			
Analista	Viviana Moschner	**Fecha**	12/08/2020
Regla Nº			
4	si Se inscribió a otra carrera (antes) = No y Familiares a cargo = No y Estudios padre > 5 y Trabaja = Trabaja entonces R		

Si se incluyen los atributos académicos, total de materias regularizadas, aprobadas y desaprobadas durante el primer, segundo y tercer año de cursado, se obtienen las siguientes reglas:

Si durante el primer año aprobó más de 4 materias, durante el segundo más de 6 y durante el tercero más de 3, entonces es Egresado, Tabla 5-71.

Tabla 5-71: Regla para Egresado SPSS Modeler (c/datos académicos)

REGLA PARA CALIDAD_R=E			
Analista	Viviana Moschner	**Fecha**	12/08/2020
Regla			
16	si Aprobó 3º año > 0 y Aprobó 3º año > 3 y Aprobó 2º año > 6 y Aprobó 1º año > 4 entonces E		

Si durante el segundo año aprobó menos de 4 asignaturas, reprobó 1 y durante el tercer año no aprobó ninguna entones es Posible Desertor, Tabla 5-72.

Tabla 5-72: Regla para Posible Desertor SPSS Modeler (c/datos académicos)

REGLA PARA CALIDAD_R=PD			
Analista	Viviana Moschner	**Fecha**	12/08/2020
Regla			
1	si Aprobó 3º año <= 0 y Aprobó 2º año <= 3 y Reprobó 2º año <= 1 entonces PD		

Si durante el segundo año aprobó más de 3 materias y durante el tercer año entre y 3 y regularizo durante el primer año menos de 11, entonces es Rezagado, Tabla 5-73.

Tabla 5-73: Regla para Rezagado SPSS Modeler (c/datos académicos)

REGLA PARA CALIDAD_R=R			
Analista	Viviana Moschner	**Fecha**	12/08/2020
Regla			
9	si Aprobó 3º año > 0 y Aprobó 3º año <= 3 y Regularizó 1º año <= 11 y Aprobó 2º año > 3 y Aprobó 3º año <= 2 entonces R		

El set de datos que se obtuvo al aplicar el algoritmo de clasificación supervisada, Redes Neuronales, se utiliza con la herramienta RapidMiner. En la Tabla 5-74, se especifican los atributos de entrada y salida que se utilizan con esta herramienta.

Tabla 5-74: Reporte de Entrenamiento del Algoritmo Rule Induction-RapidMiner

REPORTE DE ENTRENAMIENTO DEL ALGORITMO RULE INDUCTION			
Analista	Viviana Moschner	**Fecha**	12/04/2020
ID	**Datos**	**Tipo**	
C5.0			
$N-Calidad Real	Calidad real	Destino -Target	
Colegio capital interior	El colegio secundario está ubicado en capital de provincia o interior	Entrada	
Costea estudios	Modo de costear estudios	Entrada	
Está en otra carrera	Se inscribió a otra carrera con anterioridad	Entrada	
Edad al ingresar a la carrera	Edad al ingresar a la carrera	Entrada	
Estado civil	Estado civil del alumno	Entrada	
Fliares a cargo	Número de familiares a cargo	Entrada	
Estudios madre	Último nivel de estudios de la madre	Entrada	
Estudios padre	Último nivel de estudios del padre	Entrada	
Generación universitaria	¿Es la primera generación universitaria?	Entrada	
Madre vive	Vive la madre	Entrada	
Res proc	¿Procedencia es igual a residencia?	Entrada	
Título sec	Tipo de título secundario	Entrada	
Trabaja	Trabaja	Entrada	

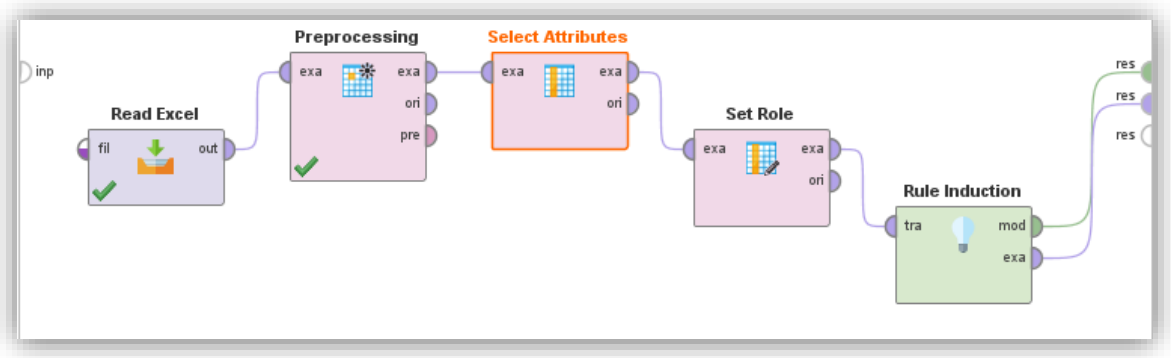


Figura 5-10: Flujo del proceso con RapidMiner

Las reglas que se obtienen al aplicar el algoritmo son las siguientes:

El alumno que con anterioridad se inscribió a otra carrera y no trabaja es Egresado, Tabla 5-75.

Tabla 5-75: Regla para Egresado con RapidMiner

REGLA PARA CALIDAD_R=E

Analista	Viviana Moschner	**Fecha**	12/08/2020
Se inscribió a otra carrera (antes) = Sí y Trabaja = No trabaja Entonces E			

El alumno que no se inscribió a otra carrera con anterioridad y costea sus estudios con ayuda familiar o con su trabajo es Posible Desertor, Tabla 5-76.

Tabla 5-76: Regla para Posible Desertor con RapidMiner

REGLA PARA CALIDAD_R=PD			
Analista	Viviana Moschner	**Fecha**	12/08/2020
Se inscribió a otra carrera (antes) = No y Costea estudios ≤ 3.50 Entonces PD			

Se observa que en el conjunto de reglas obtenidas con ambas herramientas, algunos de los factores que inciden en la condición del alumno, son coincidentes, entre ellos: costea sus estudios, trabaja, se inscribió en otra carrera de la Facultad (con anterioridad).

Capítulo 6

Resultados

6 Resultados

6.1 Modelo Predictivo para identificar factores potenciales de riesgo de abandono

El análisis predictivo es la rama de la Minería de Datos que trata con la predicción de las probabilidades y las tendencias futuras. El elemento central del análisis predictivo es una variable que puede ser medida por un individuo u otra entidad para predecir el comportamiento futuro [45]. Consiste en la extracción de conocimiento existente en los datos y en su utilización para predecir tendencias futuras y patrones de comportamiento pudiendo aplicarse sobre cualquier evento desconocido ya sea en el pasado, presente o futuro. Este se fundamenta en la identificación de relaciones entre variables en eventos pasados, para luego explotar dichas relaciones y predecir posibles resultados en futuras situaciones [45].

Realizar una predicción no es lo mismo que realizar un pronóstico. Con un pronóstico se podría predecir la cantidad de alumnos que abandonen un curso, mientras que con el análisis predictivo se podría predecir cuáles de ellos probablemente abandonen el curso.
A continuación se describe con qué nivel fueron alcanzados los objetivos propuestos.
Los objetivos específicos planteados en el TFM fueron:

- *Relevar conceptos, técnicas y herramientas vinculadas con la Explotación de Información.*
- *Analizar las metodologías de Explotación de Información disponibles, seleccionando la más adecuada para el caso de estudio.*
- *Describir el problema, contextualizando las distintas normativas institucionales vigentes en lo referente a la permanencia del estudiante.*
- *Desarrollar y validar un modelo predictivo, utilizando la metodología MoProPEI, para identificar factores potenciales de riesgo de abandono de los estudiantes del Profesorado en Ciencias de la Educación, (2010 a 2018).*

Se profundizó en el estudio del estado del arte de las técnicas y herramientas relacionadas con la Explotación de Información.

Se describieron las metodologías de Explotación de Información disponibles detallando en particular los subprocesos, fases y actividades y tareas de la metodología MoProPEI, con la que se desarrolló el presente TFM. Se destacaron como fortalezas de la misma, la producción de piezas de conocimiento para la toma de decisiones, la planificación, administración y

documentación de todos los aspectos necesarios para el desarrollo de un proyecto de Explotación de Información.

Se definió que la brecha existente entre ingresos y egresos anuales es la problemática que fundamenta el desarrollo del presente trabajo, situación que se registra y observa en el Profesorado en Ciencias de la Educación en cada cohorte, entre los Años 2010 al 2018.

Se desarrolló y validó el modelo predictivo utilizando la metodología MoProPei. Se obtuvieron plantillas que detallan los conocimientos adquiridos, las mismas podrán ser utilizadas en proyectos similares.

Se clasificó a la población estudiantil, en conjunto con el personal experto en la temática, en base a criterios predefinidos tales como: el total de materias aprobadas, regularizadas, inscripciones al cursado, año de la última actividad registrada; obteniéndose los siguientes grupos: Sin Actividad, Egresado, Posible Desertor y Rezagado, Gráfico 6-1.

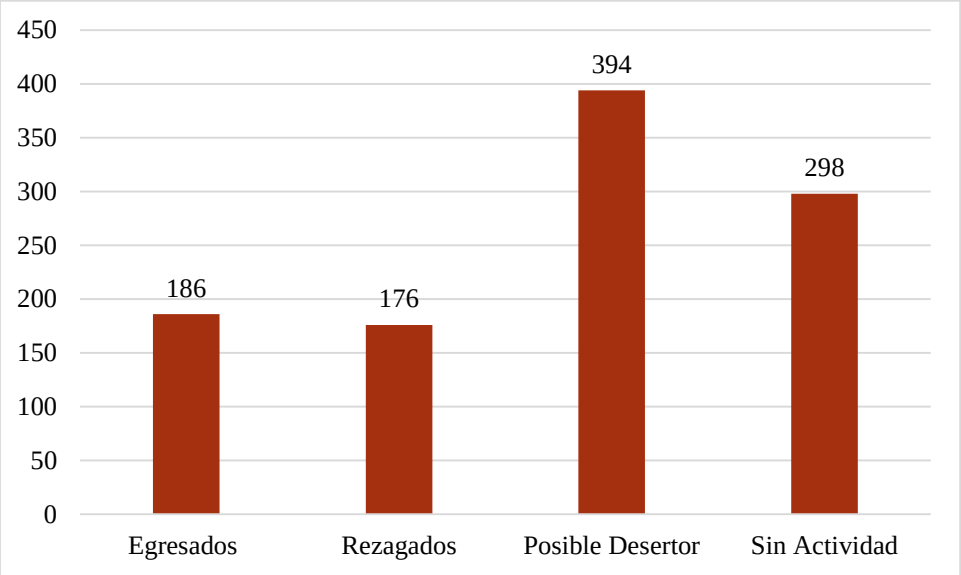


Gráfico 6-1: Clasificación de grupos

Sin Actividad: Integrado por aquellos alumnos que se inscribieron a la carrera, pero no al cursado de las asignaturas y tampoco rindieron examen final.

Egresado: Por aquellos estudiantes que aprobaron todas las actividades del plan de estudios e iniciaron el trámite de título.

Posible Desertor: Conformado por los alumnos con actividad académica, pero que no la tuvieron desde el año 2018.

Rezagado: Grupo integrado por aquellos alumnos con actividad académica reciente, pero no acorde a la cohorte a la que pertenecen.

Con el fin de evitar confusión al modelo desarrollado, se decidió eliminar al grupo de alumnos que no registró actividad académica (Sin Actividad). Este modelo se construyó a partir de 756 registros, el 70 % de ellos se utilizó para el entrenamiento y el 30% restante para la validación del mismo.

Se aplicó el algoritmo selección de características de la herramienta IBM SPSS Modeler, el que identificó los atributos que carecen de importancia, lo que permitió el descarte selectivo de los mismos. Se aplicó la técnica de aprendizaje supervisado, Redes Neuronales, obteniéndose una precisión[1] del 96,1%.

Se utilizó el set de datos obtenidos como resultado de la clasificación supervisada con RN y se aplicó el algoritmo C5.0 obteniéndose con ello el conjunto de reglas que caracteriza a cada uno de los grupos.

Del análisis del conjunto de reglas generadas, puedo afirmar que existen factores de tipo personal, académico y socio económico que podrían incidir en la deserción de los alumnos en el Profesorado en Ciencias de la Educación (2010-2018), Anexo II.

El grupo clasificado como Posible Desertor presentó las siguientes características: no se inscribió en otra carrera con anterioridad, costea los estudios con ayuda familiar o con su trabajo, no tiene familiares a cargo y el padre tiene como máximo nivel de estudios el secundario completo.

El conjunto de alumnos clasificado como Rezagado, presentó como factores que lo caracterizan: es la primera carrera de la UA en la que se inscribe, no tiene familiares a cargo,

[1] Precisión=VP/(VP+FP)

trabaja y el padre tiene nivel de estudios universitarios incompletos, completos o de posgrado.

Los factores que caracterizaron al grupo de Egresados son: no trabaja y se inscribió a otra carrera de la UA con anterioridad.

Con la aplicación del algoritmo Rule Induction, se obtuvo el conjunto de reglas que caracterizan a los grupos clasificados, Anexo II.

En este conjunto de reglas se detectó que tener familiares a cargo, ser la primera generación universitaria, costear los estudios con su trabajo o ayuda familiar, son características que distinguen al Posible Desertor.

Capítulo 7

Conclusiones y trabajos futuros

7 Conclusiones y futuras líneas de investigación

En la práctica, en la UA que comprende a la carrera en estudio, se llevan a cabo numerosas actividades y programas, con el fin de promover la retención en las carreras que de ella dependen. Se puede afirmar que las políticas elaboradas no logran paliar la problemática en estudio debido a que no han podido individualizar a los alumnos que se encuentran en riesgo de abandono.

El objetivo principal de este trabajo es detectar factores potenciales de deserción o desgranamiento, con el fin de que los nuevos programas de retención que se desarrollen puedan dirigirse a ese grupo específico de estudiantes.

En base a los patrones obtenidos en la presente investigación, con la aplicación de los algoritmos de clasificación e inducción, se puede afirmar que es posible generar conocimiento en el que las autoridades de la unidad académica puedan respaldarse para individualizar al grupo en riesgo de abandono y desarrollar nuevas políticas con el fin de prevenir la deserción y aumentar la retención estudiantil.

Este TFM permitió apreciar la importancia del proceso de recopilación, análisis y preparación de los datos, detallado en las fases Entendimiento de los datos y Preparación de los datos de la metodología utilizada.

Partiendo de la base de que los datos son fundamentales para los proyectos de Explotación de Información, ya que sin ellos no se podrían llevar a cabo estos procesos, es conveniente y aconsejable que las instituciones generen mecanismos que permitan capturar datos con valor significativo que puedan ser utilizados efectivamente en los procesos de generación de conocimiento.

Como líneas a ser abordadas en el futuro se propone adaptar el presente modelo, para su aplicación a otras carreras de la Unidad Académica y la elaboración e implementación de un tablero de control que permita a los docentes y expertos, hacer un seguimiento del rendimiento académico de los alumnos y detectar con antelación al alumno en riesgo de abandono, para ofrecer la ayuda oportuna o conveniente para cada situación tipificada en el presente trabajo.

Queda pendiente para futuras investigaciones analizar, evaluar y utilizar alguno de los siguientes modelos predictivos Support Vector Machine, Deep Neural Networks, Long

Short-Term Memory Networks (LSTM), Convolutional Neural Networks (CNN), Random Forest, XGBoost.

BIBLIOGRAFÍA

[1] "Egresados 2017-UNNE EN CIFRAS," 2017.

[2] R. E. López Briega, "Ciencia de datos," *Ciencia de datos - Libro online de IAAR*, 2017. [Online]. Available: https://iaarbook.github.io/datascience/.

[3] P. Altaria, J. M. Molina, A. Berlanga, and M. A. Patri-, *Ciencia de Datos*. .

[4] R. García-Martínez and P. Britos, "Towards an Information Mining Engineering," *Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching*, pp. 83–99, 2011.

[5] R. García-martínez, P. Britos, and D. Rodríguez, "Based on Intelligent Systems," pp. 402–410, 2013.

[6] P. Britos, "Procesos de explotación de información basados en sistemas inteligentes," *Universidad Nacional de La Plata, Argentina*, 2008.

[7] P. Britos and R. García-martínez, "Propuesta de Procesos de Explotación de Información," *XV Congreso Argentino de …*, pp. 1041–1050, 2009.

[8] D. Rodríguez, "Estimación Empírica de Carga de Trabajo en Proyectos de Explotación de Información," *… de Ciencias de la …*, pp. 664–673, 2010.

[9] S. Martins, "Derivación del Proceso de Explotación de Información desde el Modelado del Negocio," *Revista Latinoamericana de Ingenieria de Software*, vol. 2, no. 1, p. 53, 2015.

[10] D. M. Mansilla and P. García-Martínez, "Modelo de proceso para elicitación de requerimientos en proyectos de explotación de información.," 2012.

[11] S. K. David, A. T. M. Saeb, M. Rafiullah, and K. Rubeaan, "Classification Techniques and Data Mining Tools Used in Medical Bioinformatics," 2018.

[12] KNIME AG, "KNIME Analytics Platform | KNIME," *Knime*. 2019.

[13] Rapid-i, "The RapidMiner GUI Manual," *October*, 2009.

[14] IBM, "IBM SPSS Modeler CRISP-DM," *IBM Corporation*, 2016.

[15] V. Tinto, "Summary for Policymakers," in *Climate Change 2013 - The Physical Science Basis*, Intergovernmental Panel on Climate Change, Ed. Cambridge: Cambridge University Press, 1989, pp. 1–30.

[16] E. Himmel and E. Himmel K., "Modelos de análisis de la deserción estudiantil en la educación superior," *Calidad de la Educación*, 2002.

[17] J. B. Berger and J. M. Braxton, "Revising Tinto's interactionalist theory of student departure through theory elaboration: Examining the role of organizational attributes in the persistence process," *Research in Higher Education*, 1998.

[18] L.-E. Gonzalez-Fiegehen, "FORMACION UNIVERSITARIA POR

COMPETENCIAS (2007)," *Seminario Internacional CINDA*, 2007.

[19] L. E. González Fiegehen, "Repitencia y deserción universitaria en América Latina," *Informe sobre la Educación Superior en América Latina y el Caribe 2000-2005*, 2000.

[20] A. M. G. De Fanelli, "Acceso, abandono y graduación en la educación superior argentina," *Sistema de Información de Tendencias Educativas en América Latina*, 2006.

[21] S. Universitarias, "2017 - 2018 Volume 13," 2018.

[22] C. C. Russo, "Minería de datos aplicada a estrategias para minimizar la deserción universitaria en carreras de Informática de la UNNOBA," *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, 2019.

[23] J. G. A. Pautsch, H. D. Kuna, and A. E. Godoy, "Resultados Preliminares del Proceso de Minería de Datos Aplicado al Análisis de la Deserción en Carreras de Informática Utilizando Herramientas Open Source Objetivo principal Revisión conceptual," no. Md, pp. 1027–1036.

[24] S. Formia, L. C. Lanzarini, and W. Hasperue, "Characterization of university drop-out at UNRN using data mining. A study case," *XIX Congreso Argentino de Ciencias de la Computación*, pp. 681–690, 2013.

[25] K. B. Eckert and R. Suénaga, "Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos," *Formacion Universitaria*, 2015.

[26] "Minería de datos para un sistema de alerta temprana de deserción en carreras de Ingeniería," in *http://sedici.unlp.edu.ar/handle/10915/45515*, 2015.

[27] S. Perez, M. Giuliano, A. Sacerdoti, O. Spositto, and C. Gargano, "Abandono y egresos en las carreras de Ingeniería de la Universidad Nacional de la Matanza," in *Conferencia Latinoamericana sobre el abandono en la educación superior*, 2010, p. 11.

[28] P. E. Ramírez and E. E. Grandón, "Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados," *Formación universitaria*, 2018.

[29] K. Amaya, E. Barrientos, and J. Heredia, "Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos," in *Mining Techniques*, 2014.

[30] W. Hasperue, "Extraccion Del Conocimiento En Grandes Bases De Datos Utilizando Estrategias Adaptativas," 2013.

[31] J. A. Vanrell, "Un Modelo De Procesos Para Proyectos De Explotación De Información," *Iidia.Com.Ar*, p. 115, 2011.

[32] J. M. Moine, S. Gordillo, and A. S. Haedo, "Análisis comparativo de metodologías para la gestión de proyectos de minería de datos," *Xvii Congreso Argentino De*

Ciencias De La Computación, 2011.

[33] S. Martins, P. Pesado, and R. García-Martínez, "Propuesta de Modelo de Procesos para una Ingeniería de Explotación de Información: MoProPEI," *Revista Latinoamericana de Ingenieria de Software*, 2015.

[34] G. Fois, G. A. Agüero Crovella, and P. V. Britos, "Evaluación comparativa de las metodologías Team Data Science Process TDSP y Analytics Solutions Unified Method for Data Mining ASUM-DM desde la perspectiva de la ciencia de datos," in *Investigación Formativa en Ingeniería*, 2020.

[35] T. E. C. En, D. D. E. Imagen, and S. Y. Multimedia, "Estudiantes 2017 - UNNE EN CIFRAS," 2017.

[36] P. Y. Planificación and D. E. L. A. Asignatura, "Universidad Nacional del Nordeste Facultad de Humanidades Universidad Nacional del Nordeste Facultad de Humanidades," no. 3500, pp. 1–29, 2020.

[37] R. Gutiérrez, "Universidad Nacional del Nordeste Res_720_INGRESO_2017," *Histórica*, vol. III, No. 1, pp. 1–15, 1979.

[38] E. L. Consejo and D. D. E. La, "Universidad Nacional del Nordeste Facultad de Humanidades Dirección Gestión Académica Resolución N ° 005 / 17 – CD Universidad Nacional del Nordeste Facultad de Humanidades Dirección Gestión Académica," no. 3500, 2016.

[39] UNNE, "Resolución C.S. 162/03," *C.S.*, p. Artículo 2, 2003.

[40] "RES_CD_692_16.pdf." .

[41] L. Mirás *et al.*, "Sistema Araucano. Manual de Definiciones Conceptuales y Operativas," *Departamento de Información Universitaria*, p. 42, 2014.

[42] P. Rocha, "Estatuto_UNNE," pp. 1–34.

[43] M. Menéndez and M. De Lujan Gurmendi, "6º Simposio Argentino De Informática En El Estado," *6° Simposio Argentino de informática en el estado*, 2013.

[44] A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," in *Proceedings of the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015*, 2015.

[45] R. Bansal, A. Mishra, and S. N. Singh, "Mining of educational data for analysing students' overall performance," in *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, 2017, pp. 495–497.

ANEXO I - Proceso de Derivación de Modelos

Técnica Tabla Término-Categoría-Definición del Dominio

Por medio de esta técnica, se identifican los elementos relevantes pertenecientes al modelo de negocio. Los términos detectados serán introducidos en la tabla en orden alfabético. Un término pertenece a alguna de las siguientes categorías: Concepto, Atributo o Relación. Para la obtención de la tabla TCDD, se requiere como elemento de entrada a la descripción del dominio del negocio.

El procedimiento consiste en identificar en la descripción del dominio del negocio, los elementos relevantes para la comprensión del mismo. Por cada término detectado, se procede a la categorización del mismo. La Tabla A 1, resume los pasos necesarios para la implementación de esta técnica.

Tabla A 1: Técnica Tabla Término-Categoría-Definición del Dominio

Entrada	**Descripción del dominio del negocio**
Salida	**Tabla Término-Categoría-Definición del Dominio**
Paso	**Descripción**
1	Identificar en el dominio del negocio aquellos términos relevantes para la comprensión del mismo
2	Identificar la categoría a la que pertenece cada término identificado
3	Definir cada término en base al dominio del negocio analizado

Como resultado de aplicar esta técnica, se obtiene la tabla TCDD, Tabla A 2.

Tabla A 2: Tabla Termino Categoría Definición del Dominio

Término	**Categoría**	**Definición**
alumno	concepto	Persona que se inscribió a la carrera PCE plan 2000
año egresó secundario	atributo	Año en que egresó del secundario
año última regularizada	atributo	Última materia regularizada (año)
año ultima rendida	atributo	Última materia rendida (año)

beca	atributo	Indica si el alumno tiene alguna beca
calidad	atributo	Código que indica si es alumno activo o egresado
cohorte	atributo	Año en que el alumno ingresó a la carrera
colegio	atributo	Código que identifica a la institución de la que el alumno egresó del nivel secundario-(CUE)
costea estudios	atributo	Modo en que el alumno costea sus estudios
edad	atributo	Edad al ingresar a la carrera
egresó	relación	La persona egresó de un determinado colegio secundario
es	relación	Es egresado o alumno activo
es	relación	Es alumno regular a no regular
estado civil	atributo	Código que indica el estado civil del alumno
familiares a cargo	atributo	Número de familiares que el alumno tiene a cargo
género	atributo	Indica género del alumno
indica	relación	Indica año en que egreso del secundario
identifica	relación	Identifica el título secundario con el que egresó
Identifica	relación	Identifica el género del alumno
identifica	relación	Identifica el estado civil del alumno
inscripto en otra carrera	atributo	Está inscripto en otra carrera (antes)
localidad colegio secundario	atributo	Localidad colegio secundario
localidad procede	atributo	Localidad de la que el alumno procede
localidad reside	atributo	Localidad en la que el alumno reside
madre vive	atributo	Indica si la madre vive
padre vive	atributo	Indica si el padre vive
pertenece	relación	El alumno pertenece a determinada cohorte
procede	relación	El alumno procede de la localidad
reside	relación	El alumno reside en la localidad
regular	atributo	Código que indica si el alumno cumplió con el requisito de permanencia en la Facultad, es regular o no regular

regularizo	relación	El alumno regularizo un número de asignaturas
rindió	relación	Rindió un determinado número de materias
se inscribió	relación	Se inscribió a cursar actividades
tiene	relación	El alumno tiene beca
tiene	relación	El padre tiene un nivel de estudios
tiene	relación	El alumno tiene familiares a cargo
título secundario	atributo	Código del título secundario que obtuvo el alumno
trabaja	atributo	Indica si el alumno trabaja
total aprobadas	atributo	Número de materias aprobadas
total inscripciones a cursar	atributo	Número de inscripciones al cursado
último nivel estudios padre	atributo	Código que indica el último nivel de estudios que obtuvo el padre
último nivel de estudios madre	atributo	Código que indica el último nivel de estudios que obtuvo la madre

Técnica Tabla Concepto-Atributo-Relación-Valor del Dominio

Por medio de esta, se define la estructura del dominio de negocio, proporcionando un listado con los conceptos que se manipulan en el mismo. Cada concepto quedará descripto a través de los atributos que lo componen, el identificador que describe su relación con el concepto y los valores posibles de cada atributo. Para la ejecución de la tabla CARVD, se utiliza como elementos de entrada la tabla TCDD y la descripción de los datos del negocio. Para su aplicación, se procede a identificar en la tabla TCDD, aquellos términos cuya categoría es Concepto, introduciéndolos en la columna etiquetada con el mismo nombre, luego se agrupan en la columna atributo, aquellos términos presentes en la tabla TCDD cuya categoría es Atributo, asignándolos en la fila correspondiente al concepto con el cual dicho atributo se relaciona, es decir, existe un término en la tabla TCDD, cuya categoría es Relación, que vincula al concepto con el atributo a asignar. El identificador del término de categoría Relación previamente identificado, será asignado en la columna relación, obteniéndose la estructura general del negocio. A partir de la descripción de los datos, se procede a identificar los posibles valores de cada atributo en la columna valor. Como resultado de aplicar el

procedimiento definido al caso de estudio propuesto. La Tabla A 3, resume los pasos necesarios para la implementación de esta técnica.

Tabla A 3: Técnica Tabla Concepto-Atributo-Relación-Valor del Dominio

Entradas	**Tabla Término-Categoría-Definición del Dominio** **Descripción de los datos del negocio**
Salida	**Tabla Concepto-Atributo-Relación-Valor del Dominio**
Paso	**Descripción**
1	Identificar en la tabla TCDD, aquellos términos cuya categoría es Concepto registrarlos en la columna homónima
2	Identificar en la tabla TCDD, aquellos términos cuya categoría es Atributo y se relacionen con uno de los conceptos previamente identificados, registrándolos en la columna atributo de la fila perteneciente al concepto asociado
3	Identificar en la tabla TCDD, el término Relación que vincula al concepto con el atributo que lo compone y registrar su nombre en la columna relación
4	Registrar en base a la descripción de los datos del negocio, los valores posibles de cada atributo identificado

Como resultado de aplicar el procedimiento definido al caso de estudio propuesto, se obtiene la tabla CARDV, Tabla A 4.

Tabla A 4: Tabla Concepto-Atributo-Relación-Valor del Dominio (CARVD)

Concepto	**Atributo**	**Relación**	**Valor**
alumno	año egreso secundario	indica	num (1940 a 2009)
	año última regularizada	indica	num (2010-2018)
	año ultima rendida	indica	num (2010-2018)
	beca	tiene	S-N
	calidad	es	A-E
	cohorte	pertenece	num (2010-2018)
	colegio	identifica	num-Código Único de Establecimiento
	costea estudios	indica	
	edad	Indica	num 18 a 45
	estado civil	identifica	num-1 a 6
	familiares a cargo	tiene	num-1 a 6
	género	identifica	num 1-2

	inscripto en otra carrera	Indica	S-N
	localidad colegio secundario	indica	num- Código Postal
	localidad procede	procede	num- Código Postal
	localidad reside	reside	num- Código Postal
	madre vive	indica	Si-No
	padre vive	indica	Si-No
	regular	es	R-N
	título secundario	identifica	num-1-6
	trabaja	indica	Si-No
	total aprobadas	indica	Num 0-35
	total inscripciones a cursar	indica	Num 0-40
	último nivel estudios padre	identifica	num-1-12
	último nivel estudios padre	identifica	num 1-12

Técnica Tabla Concepto-Relación del Dominio

Se utiliza esta técnica para definir las interacciones entre los conceptos del dominio de negocio. Dicha tabla proporciona una lista de las relaciones entre los conceptos que se definen en el dominio. Cada relación quedará definida a través de los conceptos que la conforman, el nombre de la relación y una descripción de la misma. Para la aplicación de la tabla CRD, se dispone como productos de entrada las tablas TCDD y CARVD.

Para comenzar a aplicar la tabla CRD, se procede a identificar aquellos términos cuya categoría sea Relación, presentes en la tabla TCDD que indiquen un vínculo entre dos conceptos del problema identificados en la tabla CARVD. Por cada elemento identificado se registrará en la columna concepto, aquel concepto que genera dicha relación, en la columna concepto asociado aquel concepto vinculado, en la columna relación el nombre que describe el tipo de vínculo y la descripción que defina en base al dominio del negocio la relación identificada, en la columna homónima. Estos últimos dos elementos presentes en la tabla TCDD. La Tabla A 5, resume los pasos necesarios para la implementación de esta técnica.

Tabla A 5: Técnica Tabla Concepto-Relación del Dominio

Entradas	**Tabla Término-Categoría-Definición del Dominio** **Tabla Concepto-Atributo-Relación-Valor del Dominio**
Salida	**Tabla Concepto-Relación del Dominio**
Paso	**Descripción**
1	Identificar en la tabla TCDD, aquellos términos cuya categoría es Relación, que vinculen dos conceptos identificados en la tabla CARVD y registrar su nombre identificativo y su descripción en las columnas correspondientes
2	Identificar los conceptos que intervienen en cada relación registrada, y registrar al concepto que origina dicha relación y al concepto vinculado, en las columnas concepto y concepto asociado respectivamente

Como resultado de aplicar el procedimiento se obtiene, Tabla A 6.

Tabla A 6: Tabla Concepto-Relación del Dominio (CRD)

Concepto	**Concepto asociado**	**Relación**		**Descripción**
alumno				

Técnica Red Semántica del Modelo de Negocio

Por medio de esta técnica, se expresan las interacciones entre los elementos del modelo de negocio, las mismas representan el conocimiento del negocio, desde un enfoque orientado a las relaciones de sus componentes. La conceptualización obtenida de aplicar dicha técnica, proporciona una visión general, simple y precisa de los elementos del negocio, sus características y sus relaciones.

Para la aplicación de la RSMN, se utilizan como productos de entrada la tabla CARVD, que contiene la estructura del negocio, y CRD, que comprende las relaciones entre los conceptos del negocio.

La Tabla A 7, resume los pasos necesarios para la implementación de esta técnica y sus elementos de entrada y salida.

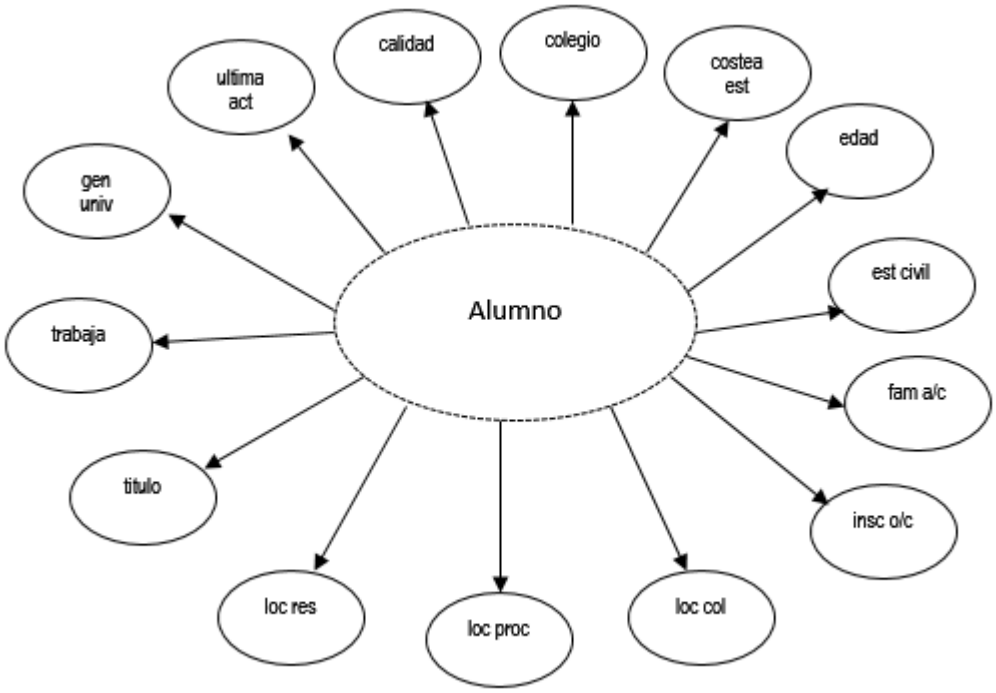
Tabla A 7: Técnica red Semántica del Modelo de Negocios

Entradas	**Tabla Concepto-Atributo-Relación-Valor del Dominio** **Tabla Concepto-Relación del Dominio**
Salida	**Red Semántica del Modelo de Negocios**

Paso	**Descripción**
1	Asociar los elementos (conceptos, atributos, relaciones y valores) identificados en la tabla CARVD obteniendo las estructuras de los conceptos implicados
2	A partir de las estructuras, se asocian aquellas que se identifica una relación en la tabla CRD, constituyendo la red semántica del modelo de negocio

Como resultado de aplicar el procedimiento definido al caso de estudio propuesto, se obtiene la RSMN.

Gráfico A 1: Red Semántica del Modelo de Negocio



Técnica Tabla Término-Categoría-Definición del Problema de Explotación de Información

Por medio de esta técnica, se identifican las ideas pertenecientes al problema de explotación de información, posteriormente utilizadas en formalismos orientados al problema de explotación. Los términos detectados serán introducidos en la tabla en orden alfabético.

Para la ejecución de la tabla TCDPEI, se dispone como productos de entrada la descripción del problema de explotación de información y de los datos del problema de negocio.

La Tabla A 8, resume los pasos necesarios para la implementación de esta técnica y sus elementos de entrada y salida.

Tabla A 8: Técnica Tabla Término-Categoría-Definición del Problema de Explotación de Información

Entradas	**Descripción del PEI** **Descripción de los datos del PEI**
Salida	**Tabla-Término-Categoría-Definición del PEI**
Paso	**Descripción**
1	Identificar en el problema de explotación de información los elementos relevantes para la comprensión del mismo
2	Por cada término detectado, identificar la categoría a la que pertenece cada elemento
3	Definir cada elemento en base al problema de explotación de información analizado

Como resultado de aplicar el procedimiento, se obtiene, Tabla A 9.

Tabla A 9:

Término	**Categoría**	**Definición del PEI**
alumno	concepto	Persona que se inscribió a la carrera PCE plan 2000
año egresó secundario	atributo	Año en que egresó del secundario
año última regularizada	atributo	Última materia regularizada (año)
año ultima rendida	atributo	Última materia rendida (año)
beca	atributo	Indica si tiene beca
calidad	atributo	Código que indica si es alumno activo o egresado
cohorte	atributo	Cohorte a la que pertenece
colegio	atributo	Código que identifica a la institución de la que el alumno egresó del nivel secundario-(CUE)
costea estudios	atributo	Modo en que el alumno costea sus estudios
edad	atributo	Edad al ingresar a la carrera
egresó	relación	La persona egresó de un determinado colegio secundario
es	relación	Es egresado a alumno activo
es	relación	Es regular o no regular

estado civil	atributo	Código que identifica el estado civil del alumno
familiares a cargo	atributo	Número de familiares que el alumno tiene a cargo
género	atributo	Género
grupo	concepto	Identifica grupo
inscripto en otra carrera	atributo	Inscripto en otra carrera
indica	relación	Indica año de egreso secundario
identifica	relación	Identifica el título secundario con el que egresó
identifica		Identifica el género del alumno
identifica	relación	Identifica estado civil del alumno
localidad colegio secundario	atributo	Localidad colegio secundario, interior o capital de provincia
localidad procede	atributo	Localidad de la que el alumno procede
localidad reside	atributo	Localidad en la que el alumno reside
madre vive	atributo	
padre vive	atributo	
pertenece	relación	El alumno pertenece a una corte
procede	relación	El alumno procede de la localidad
reside	relación	El alumno reside en la localidad
regular	atributo	Código que indica si el alumno cumplió con el requisito de permanencia en la Facultad
regla	concepto	Identifica la regla
regularizo	relación	El alumno regularizo un número de asignaturas
rindió	relación	El alumnos rindió asignatura
se inscribió	relación	Se inscribió a cursar materias
tiene	relación	El alumno tiene beca
tiene	relación	El padre tiene un nivel de estudios
tiene	relación	El alumno tiene familiares a cargo
título secundario	atributo	Código del título secundario que obtuvo el alumno
trabaja	atributo	Indica si el alumno trabaja

total aprobadas	atributo	Número de materias aprobadas
total inscripciones a cursar	atributo	Número de inscripciones al cursado
último nivel estudios padre	atributo	Código que indica el último nivel de estudios que obtuvo el padre
último nivel de estudios madre	atributo	Código que indica el último nivel de estudios que obtuvo la madre
subconjunto		El colegio secundario puede ser una variable que la defina en un grupo
subconjunto		El título secundario puede ser una variable que la defina en un grupo
subconjunto		La forma en que costea sus estudios puede ser una variable que la defina en un grupo
subconjunto		El trabajo del alumnos puede ser una variable que la defina en un grupo
subconjunto		Nivel de estudios de los padres puede ser una variable que determine el grupo
subconjunto		La localidad del colegio secundario, interior o capital de provincia
subconjunto		La localidad de residencia respecto a la localidad de procedencia puede ser un factor que influya en el grupo
subconjunto		Familiares a cargo
subconjunto		Está inscripto en otra carrera

El identificador del término de categoría relación, será asignado en la columna relación, obteniéndose la estructura general del negocio. A partir de la descripción de los datos, se procede a identificar los valores de cada atributo en la columna valor. Finalmente, se identifica en la columna E/S, los atributos que actúan como entrada y salida del problema de explotación de información.

La Tabla A 10, resume los pasos necesarios para la implementación de esta técnica y sus elementos de entrada y salida.

Tabla A 4: Técnica Tabla Concepto-Atributo-Relación-Valor Extendida del PEI

Entradas	**Descripción del PEI** **Descripción de los datos del PEI** **Tabla-Término-Categoría-Definición del PEI**
Salida	**Tabla Concepto-Atributo-Relación-Valor Extendida del PEI**

Paso	**Descripción**
1	Identificar en la tabla TCDPEI, aquellos términos de tipo Concepto y registrarlos en la columna homónima
2	Identificar en la tabla TCDPEI, aquellos términos Atributos y se relacionen con uno de los conceptos previamente identificados, registrándolos en la columna atributo de la fila perteneciente al concepto asociado
3	Identificar en la tabla TCDPEI, el término relación que vincula al concepto con el atributo que lo compone y registrar su nombre en la columna relación
4	Registrar en base a la descripción de los datos del negocio, los valores posibles de cada atributo identificado
5	Registrar los atributos de entrada y salida del problema analizado

Como resultado de aplicar el procedimiento definido se obtiene, Tabla A 11.

Tabla A 5: Tabla Concepto-Atributo-Relación-Valor Extendida del PEI

Concepto	**Atributo**	**Relación**	**E/S**	**Valor**
alumno	beca	tiene	entrada	S-N
	cohorte	pertenece		2010-2018
	colegio secundario	indica	entrada	CUE
	edad	indica	entrada	18-45
	inscripto en otra carrera	indica	entrada	SI-NO
	localidad colegio		entrada	CP
	localidad procede	procede	entrada	CP
	localidad reside	reside	entrada	CP
	nivel estudios padre	identifica	entrada	1-12
	nivel estudios madre	identifica	entrada	1-12
	título secundario	indica	entrada	CUE
	total inscripciones a cursar	indica	entrada	0-40
	total regularizadas	indica	entrada	0-35
	total aprobadas	indica	entrada	0-35
	calidad	identifica	entrada	A-E
	ultima actividad	indica	entrada	2010-2018
grupo	código de grupo		salida	SA-PD-E-A
regla				
	costea estudios			1-
	edad			18-45
	familiares a cargo			0-6
	inscripto en otra carrera			S-N

	localidad colegio			CUE
	localidad procede			CP
	localidad reside			CP
	nivel estudio de los padres			1-12
	trabaja			Si-NO

Técnica Tabla Concepto-Relación del Problema de Explotación de Información

Por medio de esta técnica, se definen las interacciones entre los conceptos del problema de explotación de información. Dicha tabla proporciona un listado de las relaciones entre los conceptos que se definen en el problema tratado. Cada relación quedará definida a través de los conceptos que la conforman, el nombre de la relación y una descripción de la misma. Para la aplicación de la tabla CRPEI, se dispone como productos de entrada las tablas TCDPEI y CARVEPEI.

La Tabla A 12, resume los pasos necesarios para la implementación de esta técnica y sus elementos de entrada y salida.

Tabla A 6: Técnica Tabla Concepto-Relación del PEI

Entradas	**Tabla-Término-Categoría-Definición del PEI** **Tabla Concepto-Atributo-Relación-Valor Extendida del PEI**
Salida	**Tabla Concepto-Relación del Problema de Explotación de Información**
Paso	**Descripción**
1	Identificar en la tabla TCDPEI, aquellos términos categoría relación que vinculen dos conceptos identificados en la tabla CARVEPEI y registrar su nombre identificativo y su descripción en las columnas correspondientes
2	Identificar los conceptos que intervienen en cada relación registrada, y registrar al concepto que origina dicha relación y al concepto vinculado, en las columnas concepto y concepto asociado respectivamente

Como resultado de aplicar el procedimiento definido al caso de estudio propuesto se obtiene, Tabla A 13.

Tabla A 7: Tabla Concepto-Relación del PEI

Concepto		**Concepto asociado**	**Relación**	**Descripción**
regla		grupo	define	La regla define el grupo al que pertenece al alumno
grupo		alumnos	integrado	El grupo está integrado por alumnos

Tabla A 8: Algoritmo de derivación de procesos de Explotación de la Información

Sub paso		SI	NO
7.1	¿En la RSPEI, puede identificarse un único nodo variable?		Hay dos nodos variables , entones se pasas al **sub paso 7.5**
7.5	¿En la RSPEI, se identifica un nodo variable el cual posee una arista identificada con la palabra INTEGRADO POR y se relaciona con otro nodo variable siendo este el nodo destino de la relación?	ir al **sub paso 7.6**	
7.6	¿El nodo variable de origen que identifica la relación del sub paso anterior NO se identifica con las palabras del sub paso siguiente?		**Ir al sub paso 7.8**
7.8	¿El nodo variable de origen que integra la relación, entre los nodos variables referidas en el sub paso 7.5 se identifica con la palabra REGLA**?**	Se aplica el proceso de **Descubrimiento de Reglas de Pertenencia a Grupos**	

ANEXO II – Reglas del modelo

Reglas C5.0

Reglas para E - contiene 5 regla(s)
Regla 1 para E (44; 0,773)
si Se inscribio a otra carrera (antes) = Si
y Trabaja = No trabaja
entonces E
Regla 2 para E (3; 1,0)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = Si
y Costea estudios > 3,28571
y Estudios padre > 3
y Estudios padre <= 3,88889
entonces E
Regla 3 para E (14; 0,5)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre <= 5
y Costea estudios > 3
y Generación universitaria <= 1
y Costea estudios <= 5
y Estudios padre > 2
y Costea estudios > 4
y Colegio capital o interior <= 1,26316
Entonces E
Regla 4 para E (2; 1,0)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre <= 5
y Costea estudios > 3
y Generación universitaria > 1
entonces E
Regla 5 para E (27; 0,481)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre > 5
y Trabaja = No trabaja
entonces E
Reglas para PD - contiene 7 regla(s)
Regla 1 para PD (12; 0,667)
si Se inscribio a otra carrera (antes) = Si
y Trabaja = Trabaja
y Costea estudios <= 2,12500
entonces PD
Regla 2 para PD (77; 0,688)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = Si

y Costea estudios <= 3,28571
entonces PD
Regla 3 para PD (9; 0,556)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = Si
y Costea estudios > 3,28571
y Estudios padre > 3
y Estudios padre > 3,88889
entonces PD
Regla 4 para PD (255; 0,584)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre <= 5
y Costea estudios <= 3
entonces PD
Regla 5 para PD (4; 0,75)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre <= 5
y Costea estudios > 3
y Generación universitaria <= 1
y Costea estudios <= 5
y Estudios padre <= 2
entonces PD
Regla 6 para PD (7; 0,571)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre <= 5
y Costea estudios > 3
y Generación universitaria <= 1
y Costea estudios <= 5
y Estudios padre > 2
y Costea estudios > 4
y Colegio capital o interior > 1,26316
entonces PD
Regla 7 para PD (26; 0,538)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre <= 5
y Costea estudios > 3
y Generación universitaria <= 1
y Costea estudios > 5
entonces PD
Reglas para R - contiene 4 regla(s)
Regla 1 para R (8; 0,5)
si Se inscribio a otra carrera (antes) = Si
y Trabaja = Trabaja
y Costea estudios > 2,12500

entonces R

Regla 2 para R (3; 0,667)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = Si
y Costea estudios > 3,28571
y Estudios padre <= 3
entonces R

Regla 3 para R (6; 0,833)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre <= 5
y Costea estudios > 3
y Generación universitaria <= 1
y Costea estudios <= 5
y Estudios padre > 2
y Costea estudios <= 4
entonces R

Regla 4 para R (9; 0,667)
si Se inscribio a otra carrera (antes) = No
y Familiares a cargo = No
y Estudios padre > 5
y Trabaja = Trabaja
entonces R

Valor predeterminado: PD

Reglas Rule Induction

RuleModel

```
if Familiares a cargo = Si and Título sec ≤ 1.50 then PD  (68 / 20 / 7)
if Costea estudios ≤ 2.50 and Generación universitaria ≤ 1.50 and Se inscribio a otra carrera (antes) = No and Trabaja = Trabaja then PD  (45 / 17 / 5)
if Se inscribio a otra carrera (antes) = No and Edad ingreso carr > 19.50 then PD  (126 / 46 / 25)
if Costea estudios ≤ 2.50 and Generación universitaria ≤ 1.50 and Edad ingreso carr > 24.50 then PD  (10 / 0 / 2)
if Se inscribio a otra carrera (antes) = Si and Edad ingreso carr > 23.50 then E  (1 / 2 / 13)
if Costea estudios ≤ 1.50 and Se inscribio a otra carrera (antes) = No and Estudios madre ≤ 3.50 then PD  (32 / 11 / 11)
if Se inscribio a otra carrera (antes) = Si and Costea estudios > 2.50 then E  (1 / 0 / 11)
if Residencia y Procedencia > 1.50 and Se inscribio a otra carrera (antes) = Si and Título sec > 1.50 then E  (4 / 1 / 15)
if Costea estudios ≤ 2.50 and Se inscribio a otra carrera (antes) = No and Madre vive ≤ 1.50 then PD  (82 / 41 / 57)
if Trabaja = Trabaja and Costea estudios > 3.50 then E  (1 / 2 / 8)
if Edad ingreso carr ≤ 18.50 and Familiares a cargo = No and Costea estudios ≤ 2 then R  (0 / 3 / 0)
if Costea estudios > 1.50 and Estudios madre > 4.50 and Residencia y Procedencia > 1.50 and Estado civil = Soltero and Familiares a cargo = No then R  (0 /
if Costea estudios ≤ 1.50 and Estudios madre > 4.50 and Estudios padre > 3.50 and Estudios padre ≤ 4.50 then E  (0 / 0 / 5)
if Estudios padre ≤ 3.50 and Estudios madre > 2.50 and Trabaja = No trabaja and Edad ingreso carr ≤ 18.50 then R  (1 / 6 / 2)
if Título sec ≤ 1.50 and Estudios madre ≤ 7.50 and Padre vive ≤ 1.50 and Estudios madre ≤ 4.50 then PD  (8 / 1 / 2)
if Estudios madre ≤ 3.50 and Colegio capital o interior ≤ 1.50 and Estado civil = Soltero then E  (0 / 1 / 5)
if Residencia y Procedencia ≤ 1.50 and Costea estudios > 4 and Familiares a cargo = No and Edad ingreso carr ≤ 18.50 then R  (0 / 4 / 1)
if Trabaja = No trabaja and Estudios padre > 2.50 then PD  (15 / 8 / 9)
else R  (0 / 5 / 5)

correct: 468 out of 754 training examples.
```