PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

La utilización de Business Intelligence como propuesta para mejorar los indicadores de deserción de los estudiantes de la Escuela de Ingeniería Informática.

Roberto Osvaldo Cordero Cerda

INFORME DE FINAL DE PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

Julio, 2017

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

La utilización de Business Intelligence como propuesta para mejorar los indicadores de deserción de los estudiantes de la Escuela de Ingeniería Informática.

PROFESORA GUÍA: PAMELA HERMOSILLA MONCKTON
PROFESOR CORREFERENTE: RODRIGO ALFARO ARANCIBIA

Julio, 2017

Resumen

El presente trabajo tiene como finalidad proponer herramientas y una metodología de Business Intelligence (BI), y los diversos beneficios y ventajas de su implementación dentro de la educación universitaria, tomando como caso de estudio el problema de la deserción de los estudiantes de la Escuela de Ingeniaría Informática de la Pontificia Universidad Católica de Valparaíso (PUCV). La universidad cuenta con un sistema que proporciona indicadores de gestión para directivos y jefes de docencia de las distintas Escuelas de la universidad, sin embargo, dichos indicadores no suelen entregar información detallada para resolver distintas problemáticas como es detectar a posibles estudiantes en riesgo de deserción.

Para la propuesta de solución se pretende utilizar un modelo basado en las mejores prácticas para el desarrollo de proyectos de BI, realizando un análisis de la situación de la carrera, un proceso de Extracción, Transformación y Carga (ETL) a partir de los datos de los estudiantes que ingresan, la creación de un datawarehouse centrado en los factores que hacen que los alumnos deserten, herramientas avanzadas de análisis y correlación multidimensional de datos como cubos de procesamiento analítico (OLAP), predicción y clasificación, mediante minería de datos, de estudiantes con posibilidades de desertar y herramientas para generar reportes y dashboards con información precisa y detallada. La integración de todos estos elementos conforma una plataforma de Business Intelligence, que permita a directivos y jefes de docencia tener la información necesaria para realizar acciones en favor de los estudiantes con posibilidades de desertar.

Palabras Clave: Deserción Universitaria, Business Intelligence, Datawarehouse, Datamart, ETL, OLAP, Minería de datos.

Abstract

The purpose of this study is to propose tools and a methodology of Buisness Intelligence (BI), and the various benefits and advantages of his implementation inside of the university education, taking as case the study of problem of the desertion from school of informatic ingieniring of the PUCV students. The university counts with a system that allows indicators of managment for directives and bosses of teaching in the differents schools of the university, however, such indicators of managment they dont usually deliver such detailed information to resolve different problematics how detect to possible students in risk of dissection.

For the proposal of solution it pretends to use a model base don the bests practices to the development of proyects of BI, doing an analysis of the situation of the career, a Extract Transform and Load (ETL) process starting of the data of students that they enter, the creation of a datawarehouse centered on the factors thay make that the students deserted, advanced tools of analysis and correlation multidimensional of data like On-Line Analytical Processing (OLAP) cubes, prediction and clasification through data mining, of students with chances of diserting and tools to generate reports and dashboards with accurate and detailed information. The integration of all this elements forms a platform of Buisness Intelligence, that allows to directives and bosses of teaching to have necessary information to do actions in favour of students with chances of discourse.

Keywords: University desertion, Business Intelligence, Datawarehouse, Datamart, ETL, OLAP, Data Mining.

A mis padres, mi hermano, mi perro Pipa y familia completa por todos los años de esfuerzo por entregarme lo mejor, un apoyo incondicional, un completo amor y creer siempre en mi. Simplemente los amo.

A mis amigos, de Valparaíso y San Fernando, por cada momento, instante, y segundo que hemos vivido durante este largo viaje llamado universidad. Tengo los mejores amigos del mundo.

Y por último, a mi profesora guía, Pamela, por creer en este proyecto, jugársela en cada momento y estar siempre cuando necesitara ayuda. Muchas gracias profesora

.

Este trabajo está dedicado para todos ustedes, ya que sin su ayuda no lo hubiese podido lograr.

Índice

Lista de Figuras

Lista de Tablas

1. Glosario

- **Business Intelligence (BI)**: Son el conjunto de estrategias y aspectos relevantes enfocados a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa.
- **Business Process Model and Notation (BPMN)**:Es una notación gráfica estandarizada que permite el modelado de procesos de negocio, en un formato de flujo de trabajo.
- **Datamart**: Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.
- **Datawarehouse**: Es una colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.
- **ETL**: Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o apoyar un proceso de negocio.
- **FoFu**: Cursos de Formación Fundamental.
- **Front-end y Back-end**: Front-end y back-end son términos que se relacionan con el principio y el final de un proceso.
- **Indicadores de Gestión**: Son expresiones cuantitativas de las variables que intervienen en un proceso, que permiten verificar o medir la cobertura de las demandas, la calidad de los satisfactores o productos y el impacto de la solución de la necesidad de la sociedad.
- **METADATA**: Ciertos ficheros gráficos pueden contener información descriptiva acerca de los datos que contienen.
- **Minería de datos**: Es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.
- **On-Line Analytical Processing (OLAP)**: Es una solución utilizada en el campo del BI, cuyo objetivo es agilizar la consulta de grandes cantidades de datos.
- **Sistemas de Soporte a las Decisiones (DDS)**: Es un sistema de información basado en computadora que combina modelos y datos en una tentativa para resolver problemas semiestructurados con un involucramiento pleno del usuario.
- **Unified Modeling Language (UML)**: Es el lenguaje de modelado de sistemas de software más conocido y utilizado en la actualidad.

2. Introducción

Actualmente, las organizaciones cada vez se dan cuenta de la importancia de la gestión de la información y las ventajas competitivas que pueden conseguir mediante su uso. Este proceso consiste en lograr de una manera eficiente el análisis de los datos de la organización y su entorno, a través de la explotación de la información por medio de distintas tecnologías de la información, facilitando la adaptación del Business Intelligence (BI) [1].

Según Nader [2] los sistemas de información utilizados en las universidades suelen responder a algunas inquietudes mediante la preparación de informes, sin embargo, se suele utilizar aproximadamente un 60 % del tiempo en la localización y preparación de los datos, como también de asignar personal para poder presentar las respuestas a las inquietudes solicitadas y dejar un 40 % o menos para que los interesados puedan transformar la información en conocimiento. Esta problemática se debe a que los actuales sistemas no fueron construidos con el fin de brindar síntesis, análisis, consolidación, búsqueda y proyecciones de la información obtenida.

No es posible que se use el 60 % del tiempo preparando información y tan sólo el 40 % analizándola. Estos porcentajes, como en toda organización, deberían ser al revés, es decir, destinar la mayor parte del tiempo a analizar la información y tan sólo una pequeña parte del tiempo a prepararla. Es en la toma de decisiones cuando se aporta valor, no en la preparación de la información. Aquí es donde entra BI, como una herramienta que proporciona conocimiento de forma rápida, optima y eficiente para todos los usuarios, en este caso directivos de las universidades que necesiten de información precisa, coherente y verdadera para tomar decisiones frente a distintos problemas dentro de su entorno universitario.

Este trabajo está organizado de la siguiente manera: Primero en la sección 3, se define el Objetivo General y los Objetivos Específicos del presente proyecto. En la sección 4, se entrega una Justificación a la realización de este trabajo. En la sección 5, se describe el problema de los indicadores de gestión dentro de la Pontificia Universidad Católica de Valparaíso (PUCV), además, de presentar un problema dentro de las universidades: la deserción. Se realiza un análisis de éste y como la Escuela de Ingeniería Informática responde ante este problema. En la sección 6, se presenta el Estado del Arte de distintos proyectos y herramientas de BI y su aplicación en distintas universidades alrededor del mundo. Además, se presentan estudios relacionados a entender los factores de la deserción universitaria y como preverla. En la sección 7, se presenta el Marco Teórico con la descripción del conjunto de conceptos, metodologías y definiciones de las herramientas y componentes de BI. En la sección 8, se detalla los distintos pasos a realizar para desarrollar un modelo basado en BI. En la sección 9, se presenta la aplicación de la solución propuesta, mediante los pasos detallados anteriormente, para resolver el problema identificado con anterioridad dentro de la PUCV. Finalmente, en la sección 10 se presentan las conclusiones obtenidas al realizar este proyecto y el trabajo que sigue a continuación de este.

3. Definición de objetivos

3.1. Objetivo General

- Proponer un modelo basado en Business Intelligence (BI) que permita predecir posibles estudiantes con mayor riesgo de deserción de la Escuela de Ingeniería Informática de la Pontificia Universidad Católica de Valparaíso.

3.2. Objetivos Específicos

- Estudiar los principales conceptos, técnicas, tecnologías utilizadas y enfoques metodológicos existentes en los procesos de proyectos de BI.
- Analizar los sistemas de información actuales que posee la Pontificia Universidad Católica de Valparaíso.
- Investigar la problemática actual de la deserción de los estudiantes universitarios y los diversos factores que afectan la permanencia dentro de la carrera universitaria, así como la respuesta de nuestra universidad ante dicho problema.
- Proponer una solución que ayude a la toma de decisiones de directivos y jefes de docencia, mediante índices de rendimiento y pronósticos de comportamiento de los estudiantes con mayor riesgo de deserción, para tomar acciones y reducir la tasa de deserción dentro de la carrera.
- Aplicar un modelo y herramientas BI, para el procesamiento de datos, minería de datos y visualización de la información representados a través de cubos OLAP, permitiendo reportes de fácil entendimiento y mayor flexibilidad para directivos y jefes de docencia.
- Validar el modelo de predicción desarrollado.

4. Justificación del Proyecto

Actualmente las universidades son instituciones con una responsabilidad social enorme, desarrollando y transmitiendo el conocimiento entre sus integrantes hacia la sociedad. Su compromiso de enseñanza y el alto costo, demandan procesos que se desarrollen de forma eficiente y eficaz, con un gran grado de calidad en ellos. El uso de la tecnología y distintos métodos para la gestión de procesos, serian clave para apoyar y mejorar el rendimiento de los procesos dentro de una universidad. Según Jing Luan [3], las instituciones de educación superior deben de emplear herramientas que ayuden a utilizar la información de forma ordenada y el conocimiento que acumulan sus procesos para ayudar a la dirección de las universidades.

Cada estudiante cuenta con una gran cantidad de características que lo diferencia de otro y esto dificulta el proceso de formación que las universidades quieren lograr. Los directivos y profesores son agentes fundamentales en la formación de los estudiantes y son capaces de utilizar datos académicos de estudiantes como una fuente de información para generar conocimiento, conocer bajo que condiciones se desarrolla el proceso docente y elaborar métodos y procedimientos acertados para alcanzar resultados y realizar un proceso educativo de calidad. Es por esta razón que es necesario una correcta gestión de la información y el conocimiento para identificar, captar, procesar y diseminar que datos son adecuados para obtener un modelo del proceso de enseñanza que facilite la toma de decisiones y estrategias orientadoras [4].

Actualmente la Escuela de Ingeniería Informática de la Pontificia Universidad Católica de Valparaíso (PUCV) cuenta con almacenes de información, en la cual se registran enormes volúmenes de datos de los distintos estudiantes que cursan y han cursado la carrera, que al poder ser analizados, promueve el despliegue de una estrategia de análisis de datos para el descubrimiento y gestión de conocimiento asociado al proceso educativo, contribuyendo así, a transformar las actuales estrategias de formación académica y formar nuevos enfoques y acciones para mejorar la productividad de profesores y estudiantes. El saber dónde y cómo se organiza toda la información a procesar es importante, y es en ese contexto que surgen conceptos como lo son los Datawarehouse (DW), herramienta utilizada en BI que se detallaran más adelante, que posibilita a directivos y profesores a formular consultas y analizar los datos en el momento, forma y cantidad necesaria sin necesidad de acudir al personal informático de la universidad. BI se define a grandes rasgos como un proceso interactivo para explotar y analizar información estructurada sobre un área para descubrir tendencias y patrones a partir de los cuales derivar ideas y extraer conclusiones [5].

Haciendo uso de metodologías, herramientas gerenciales y distintas técnicas de BI, las universidades cuentan con un recurso que si es bien utilizado puede ayudar en el esfuerzo por ser instituciones cada vez más competitivas. Entre las ventajas que aportan las herramientas BI en la educación podemos destacar:

- Control y reducción de gastos mediante el seguimiento presupuestario y el uso de la analítica.
- Mayor competitividad y posicionamiento, gracias a una buena gestión y retención del talento.

- Mejorar la gestión educativa, por mediciones de evaluación, seguimiento a alumnos, profesores y la calidad de la enseñanza.
- Mayor eficiencia en sus procesos.

Este proyecto va enfocado a aprovechar la explotación de los datos almacenados de los estudiantes de la Escuela de Ingeniería Informática de la PUCV, que semestre tras semestre se van acumulando, en relación a un problema que afecta a muchas universidades hoy en día, la deserción universitaria, del cual se hablará en la Sección 5.

Se pretende crear una inteligencia organizacional que posibilite la identificación, captación y procesamiento de datos adecuados, mediante un modelo Business Intelligence, que facilite la toma de decisiones y la concepción de estrategias, por parte de directivos, a través del modelo propuesto, que identifique a los estudiantes con mayor riesgo de deserción.

5. Situación en Estudio

5.1. La deserción como problema actual dentro de la educación universitaria

La deserción universitaria durante un tiempo fue considerada como un fenómeno normal y una muestra de la exigencia de la universidad y de la dificultad de una carrera en particular. La deserción de los estudiantes afecta en diversos ámbitos: los personales, con una condición de fracaso que afecta emocionalmente al alumno; los institucionales, donde se refleja una disminución en el rendimiento académico de la universidad; los sociales, contribuyendo a generar un desequilibrio social; alejando los objetivos que la sociedad ha entregado a la educación universitaria, y los económicos, agregando un costo que, la deserción, aumenta al sistema educacional [6].

Según cifras del Ministerio de Educación, la deserción en la educación superior universitaria sigue siendo elevada en Chile. Para el primer año, ésta supera el 25 %. Si se analiza esta tasa en un ámbito socio económico es más alta entre estudiantes provenientes de liceos municipales, con un 32,5 %, en cambio, estudiantes egresados de colegios particulares pagados alcanzaría un 21,9 % [7]. Esto haría que Chile pierda US$ 780 millones al año (CLP $545 mil millones), según un estudio elaborado por Unesco, donde dos tercios de esa cantidad de dinero proviene desde las familias de los estudiantes y el resto de aportes del Estado [8].

Otro punto importante es entender que existen distintos tipos de deserción en la educación. Según el informe *Estudio de la deserción estudiantil en la educación superior en Colombia* [9], existen distintos tipos de deserción no excluyentes entre si:

- **Deserción total**: Un completo abandono de los estudios de formación académica del estudiante.
- **Deserción discriminada por causas**: La existencia de diversas causas que determinan la decisión.
- **Deserción por universidad, facultad o carrera**: Cuando el estudiante realiza un cambio dentro del mismo sistema educativo.
- **Deserción por programa**: Cuando el estudiante realiza un cambio de programa en una misma facultad.
- **Deserción a primer semestre de carrera**: Esto sucede cuando el estudiante no logra una adaptación a la vida universitaria.
- **Deserción acumulada**: Una sumatoria de deserciones en una institución.

En el mismo estudio anterior, se presentan factores que permiten predecir si un estudiante ésta en riesgo de deserción. Estos factores [9] son clasificados en Individuales, Académicos, Institucionales y Socioeconómicos y se pueden detallar en la Fig. 1.

Individuales	Académicos	Institucionales	Socioeconómicos
• Edad, género, estado civil • Entorno familiar • Calamidad y problemas de salud • Integración social • Incompatibilidad horaria en actividades extra académicas	• Orientación profesional • Rendimiento académico • Calidad del programa • Métodos de estudio • Calificación en examen admisión • Insatisfacción en el programa u otros factores académicos • Numero de materias	• Normalidad académica • Tipo de colegio • Becas y forma de financiamiento • Recursos universitarios • Orden público • Entorno político • Relaciones con los profesores y otros estudiantes	• Estrato • Trabajo del estudiante • Situación laboral de los padres • Dependencia económica • Personas a cargo • Nivel educativo de los padres • Entorno macroeconómico del país

Figura 1: Factores determinantes de la deserción universitaria

5.2. La deserción en la Escuela de Ingeniería Informática de la PUCV

5.2.1. Indicadores de Gestión de la PUCV

Actualmente el sistema dentro de nuestra universidad almacena una gran cantidad de datos y el aumento constante del volumen de éstos hace que directivos y profesores de nuestra universidad se enfrenten a escenarios de incertidumbre y complejidad creciente.
Para plantear si existe un problema dentro de alguna organización se deben realizar las siguientes tres preguntas [5]:

- ¿Existe una gran cantidad de información y poco tiempo para analizarla?
- ¿Los sistemas de información existentes, ayudan a tomar decisiones rápidamente?
- ¿Los responsables de generar información directiva están desbordados por peticiones de información urgente, continua y no coordinada?

Responder la primera pregunta es sencillo, actualmente existe una gran cantidad de información dentro de nuestra universidad y analizarla es cada vez más difícil por los profesores y directivos de nuestra universidad al no contar con el tiempo necesario para transformar dicha información en conocimiento para realizar clases personalizadas o nuevas estrategias dentro de su Escuela.

En el presente, la PUCV cuenta con un sistema dentro del Navegador Académico, donde existen indicadores de gestión. Sin embargo, dichos indicadores solo son visibles para directivos y jefes de docencia. Los profesores no cuentan con un sistema de información que les permita tomar decisiones de manera rápida.

Actualmente son tres los indicadores a los que se pueden acceder mediante el rut del directivo o jefe de docencia:

- Indicadores ingreso

- Indicadores proceso formativo
- Indicadores comparativos

El detalle de cada uno de los indicadores mencionados con anterioridad se encuentra detallado en el Anexo A.

Para responder a la última pregunta sobre los responsables de generar información directiva, nos encontramos que los jefes de docencia de las distintas Escuelas no tienen el tiempo para enviar esta información de forma constante a los profesores, afectando la eficiencia que podría tener la información de manera rápida y oportuna en las manos de los profesores y así generar conocimiento para realizar clases y nuevas estrategias de enseñanza.

Analizando las respuestas a las tres preguntas realizadas con anticipación nos encontramos con tres problemas en relación a la calidad y eficiencia de los indicadores entregados por el Navegador Académico:

- Muchos de los reportes e indicadores proporcionados por el Navegador Académico, muestran una ausencia de información detallada, al encontrarse en un formato estático, lo que provoca que los profesores busquen herramientas externas para poder conseguir la información necesaria, gastando una gran cantidad de tiempo en preparar la información y obtener poco tiempo para analizarla. Esto responde a la primera pregunta planteada *¿Existe una gran cantidad de información y poco tiempo para analizarla?* para señalar un problema dentro de los sistemas de información de la PUCV.
- El actual sistema de indicadores y reportes, proporcionado por el Navegador Académico, cuenta con problemas en la visualización y exportación de los reportes entregados (HTML, PDF, CVC), imposibilitando su correcta visualización y análisis. Este problema, que va algo conectado al anterior, muestra que, aunque existan varias fuentes de información disponible, existen dificultades que hacen difícil la toma de decisiones por parte de directivos, jefes de docencia y profesores. Esto responde a la segunda pregunta planteada *¿Los sistemas de información existentes, ayudan a tomar decisiones rápidamente?*, lo cual se responde con un rotundo no.
- La información sólo se encuentra disponible para directivos y jefes de docencia. Por lo que la única forma de que otros usuarios obtengan esta información es pedirla con tiempo de anticipación o mandar solicitudes a la Casa Central de la PUCV, lo que impide una rápido análisis de la información obtenida. En este problema se plantea en la última pregunta, *¿Los responsables de generar información directiva están desbordados por peticiones de información urgente, continua y no coordinada?* En nuestra universidad los jefes de docencia tienen muy poco tiempo para atender consultas por parte de los profesores. Así también, las consultas a Casa Central suelen tardar en responderse y solo si son pedidas por los directivos de la Escuela.

Ya identificado el problema de lo poco preciso y detallado de los indicadores entregados, Actualmente el indicador que detalla la tasa de retención de la carrera de Ingeniería Civil Informática, entregado por el Navegador Académico, muestra las distintas tasas de retención de

las generaciones en un semestre base de referencia, junto a la información de matrículas de dicha generación. Un ejemplo de esto se puede observar en las Figuras 2 y 3.

Tasa de Retención por Cohorte

Carrera : INGENIERIA CIVIL INFORMATICA
Año / : 2016 / 1 Semestre

Promoción	Nivel	Total Ingresados	Matriculados	No Matriculados	Tasa Retención	Tasa de No Matriculados
2016	1	96	96	0	100.00 %	0.00%
2015	2	103	84	19	81.55 %	18.45%
2014	3	96	55	41	57.29 %	42.71%
2013	4	95	34	61	35.79 %	64.21%
2012	5	88	23	65	26.14 %	73.86%
2011	6	95	39	56	41.05 %	58.95%

Figura 2: Tasa de retención estudiantes de Ingeniería Civil Informática de la PUCV, hasta el año 2016

Carrera : INGENIERIA DE EJECUCION INFORMATICA
Año / : 2016 / 1 Semestre

Promoción	Nivel	Total Ingresados	Matriculados	No Matriculados	Tasa Retención	Tasa de No Matriculados
2016	1	110	110	0	100.00 %	0.00%
2015	2	103	81	22	78.64 %	21.36%
2014	3	100	63	37	63.00 %	37.00%
2013	4	103	38	65	36.89 %	63.11%
2012	5	95	27	68	28.42 %	71.58%
2011	6	88	6	82	6.82 %	93.18%

Figura 3: Tasa de retención estudiantes de Ingeniería en Ejecución Informática de la PUCV, hasta el año 2016

Como se puede observar en las figuras anteriores, existe una baja en la tasa de retención de los estudiantes de las carreras de la Escuela y debido al poco detalle de los indicadores es necesario conocer los distintos factores que llevan a cabo este porcentaje y así también, detectar de forma temprana posibles alumnos desertores. Actualmente en nuestra escuela solo se toman medidas cuando los alumnos reprueban constantemente varios ramos, o caen en instancias de terceras oportunidades. Las medidas a realizar van desde tutorías hasta implementar planes de apoyo a los estudiantes con talleres interpersonales de desarrollo y un fuerte seguimiento al desempeño del alumno.

Por esta razón es necesario obtener indicadores de retención detallados y una forma de detectar a tiempo posibles estudiantes desertores con información clara y precisa, para tomar acciones antes de que los alumnos con mayor probabilidad a desertar reprueben sus ramos y se retiren de la carrera.

6. Estado del arte

Quizás cuando uno escuche el término "Business Intelligence", lo primero que se le puede venir a la cabeza son empresas de retail o dedicadas a los negocios. Sin embargo, la masificación de esta tecnología ha sido enorme y sus usos se han ampliado desde la medicina [10], la acción social [11], deporte [12] y la educación, uno de los temas más complejos actualmente en nuestra sociedad.

Según Guitart y Conesa [13] "Hoy en día, los principales objetivos de los gestores de las universidades son mejorar el rendimiento de la gestión interna (disminuyendo gastos y optimizando procesos) e incrementar la calidad docente e investigadora de la universidad. Los gestores universitarios también necesitan sistemas analíticos para conocer de forma fiable que ha sucedido, está sucediendo o puede suceder en la universidad." Bajo esta premisa son muchos los estudios para encontrar un modelo y herramientas que puedan satisfacer las necesidades de mejorar el rendimiento de la gestión de la información dentro de las universidades y así, predecir y conocer el estado de las organizaciones educacionales.

6.1. Proyectos e iniciativas del uso de Business Intelligence dentro de universidades

6.1.1. Universidad de las Ciencias Informáticas

Un estudio realizado en la Universidad de las Ciencias Informáticas [14], en Cuba, detectó dificultades en el tratamiento y forma en que se utilizaban los datos para el soporte a las tomas de decisiones dentro de su establecimiento. Es por esta razón que se desarrolló un sistema basado en BI para brindar facilidades en el uso de datos e información académica y matrícula de los estudiantes para analizar y descubrir nuevas oportunidades y tomar decisiones con mayor información.

6.1.2. Universidad del Magdalena

Una solución BI ha sido investigada en la Universidad del Magdalena, en Colombia, para la gestión de sus recursos y espacios físicos [15]. Está solución tuvo como resultado la obtención de informes históricos y actuales de los procesos dentro de la universidad y gestionar el rendimiento de éstos.

6.1.3. Universidades Chilenas

Otras universidades chilenas comenzaron proyectos BI en sus instituciones, tales son el caso de la Universidad de Concepción, que ha de contratar los servicios de Kr. Consulting [16] para los primeros pasos de implementación de un sistema BI. Por otro lado, la Universidad del Desarrollo ha contratado los servicios de DATAMART [17] y ha adquirido licencias corporativas de Cubix Olap Analyzer 3.1 [18] plus Data Mining y Cubix Web Edition, implementando un sistema de Control de Gestión Financiero, Presupuestario y Académico. La Pontificia Universidad Católica de Chile ha de implementar un sistema BI, sobre la plataforma MicroStrategy [19], para comparar los ingresos y egresos reales de cada unidad académica o administrativa de la institución con su respectivo presupuesto. Otro proyecto, utilizando herramientas técnicas BI, para apoyar la

gestión Institucional de la Universidad de La Serena fue realizado durante el año 2012 [20], trayendo una mejora a la provisión de datos a la administración de la institución. Por último, la Universidad de Tarapacá en Arica, Chile, ha de implementar elementos de un sistema BI para dar soporte a los requerimientos de información y análisis, en sus procesos de admisión y matrícula [21].

6.1.4. Herramientas BI en universidades alrededor del mundo

Son muchas las universidades alrededor del mundo que han de utilizar herramientas BI para obtener un mejor acceso a información actual e histórica relevante que les permita identificar tendencias y patrones, así como monitorear el progreso de estudiantes en los distintos programas académicos y carreras.

Ejemplo hay varios, uno de estos es el proyecto de BI utilizando Stratebi, Sigma y Pentaho BI por la Universidad de Zaragoza para mejorar el análisis de sus datos [22]. La Universidad de Macquarie en Sidney, Australia y la Universidad de Constanza, en Baden-Württemberg, Alemania, han de utilizar la herramienta BI Yellowfin para eliminar los silos de conocimiento y mejorar la cooperación interdisciplinaria entre los departamentos y facultades de ésta. [23]. El Instituto de Tecnología de Georgia y Maximus University han de utilizar soluciones BI gracias al sistema InetSoft que ofrece una solución integral para toda la información y requisitos analíticos de directivos y académicos [24]. Otra herramienta BI utilizada en educación es Jaspersoft que permite optimizar la toma de decisiones de instituciones como la Universidad de Johns Hopkins, la Universidad de Vanderbilt y la Universidad del Este de Washington en Estados Unidos, la Universidad de Cranfield en el Reino Unido, el Instituto Católico de París en Francia y la Universidad de Valencia en España [25]. Por último, se han publicado distintos documentos que promocionan herramientas BI en el sector de la educación, no tan sola universitaria sino también en la educación básica y media, como lo son IBM Cognos [26] y BI Oracle [27].

6.2. Estudios relacionados a la deserción universitaria

Para estudiar el problema de la deserción dentro de la educación universitaria, es necesario conocer y entender los distintos factores que llevan a que un estudiante deje sus estudios. En [28], realizado en la Universidad Católica de la Santísima Concepción, se propone un modelo conceptual que explique la deserción o permanencia de un alumno como resultado de la motivación positiva o negativa, afectada por la integración académica y social. Otro trabajo interesante fue [29], realizado por la Facultad de Ciencias Físicas y Matemáticas del Departamento de Ciencias de la Computación de la Universidad de Chile, que propone una metodología basada en minería de datos que permita identificar de forma automática a estudiantes con mayor riesgo de deserción de las carreras de Ingeniería de la Universidad de Las Américas. Por último, fuera del país también se han hecho estudios para entender los distintos factores por los cuales los estudiantes desertan de sus carreras, en *Estudio de la deserción estudiantil en la educación superior en Colombia* [9], se clasifican los distintos factores asociados describiéndolos y ejemplificándolos.

7. Marco Teórico

7.1. Business Intelligence

En la actualidad vivimos en una sociedad de la información, donde nace la necesidad de tener mejores, más rápidos y más eficientes métodos para extraer y transformar los datos de una organización en información para ser utilizada por las personas que la necesiten y así descubrir conocimiento en éstas. Es aquí donde nace el Business Intelligence como respuesta a la necesidad de transformación de los datos en información, describiendo aproximadamente a BI como una evolución de los sistemas de soporte a las decisiones (DDS). Existen varias definiciones de lo que es BI es:

Según Josep Curto [30] *"Se entiende por Business Intelligence al conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización."*

Según Larissa T.Moss [31] *"Business Intelligence no es ni un producto ni un sistema. Se trata de una arquitectura y una colección de operacionales integrados, así como apoyo a las decisiones de aplicaciones y bases de datos que proporcionan el acceso fácil de la comunidad de negocios a los datos empresariales."*

Por último, una definición [32] más amplia de lo que es BI, según The Datawarehouse Institute es: *"Business Intelligence es un término paraguas que abarca los procesos, las herramientas, y las tecnologías para convertir datos en información, información en conocimiento y planes para conducir de forma eficaz las actividades de los negocios. BI abarca las tecnologías de datawarehousing los procesos dentro del negocio, consultas, informes, análisis y las herramientas para mostrar información y los procesos con el cliente".*

Son muchas las definiciones de lo que es BI, si queremos establecer una definición formal, sería la siguiente: Business Intelligence se define como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada en información estructurada para su explotación directa mediante reportes y análisis OLAP, para su posterior análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

Dentro de los beneficios de la implementación de un sistema BI, se puede destacar:

- Transformar los datos en información que genera conocimiento para la toma de decisiones que brindan beneficios para organización.
- Crear, manejar y mantener métricas e indicadores claves de rendimiento y de metas.
- Permitir una visión de la información única, conformada, histórica, persistente y de calidad.
- Aportar información actualizada.
- Reducir el diferencial de orientación de negocio dentro de la organización.

- Mejorar comprensión y documentación de los sistemas de información.
- Mejorar la competitividad de la organización dentro del mercado.

7.1.1. Niveles de Inteligencia de Negocios

Existen tres niveles de información en los sistemas BI que se asemejan a el esquema de trabajo multidimensional que existen dentro de las organizaciones y a la necesidad de información de cada una de estas, estas son Operacional, Táctica y Estratégica, como se especifica en la Figura 4.

Figura 4: Niveles de información de Business Intelligence

A nivel Operativo, BI permite que el personal, a nivel operacional, que utilicen información operativa la reciban de forma oportuna, exacta y de la forma correcta. Esta información operativa suelen ser hojas de cálculo que se actualizan constantemente en el tiempo.

BI a nivel Táctico permite que la gerencia media y los analistas de información utilicen herramientas de análisis y consulta en línea(OLAP) para obtener acceso a la información sin que exista intervención de terceros.

Por último, BI a Nivel Estratégico permite que la alta gerencia de alguna organización pueda monitorear y analizar tendencias, indicadores, metas y objetivos de la organización.

7.2. Componentes de un sistema Business Intelligence

Para que una organización se desarrolle correctamente bajo un sistema BI, debe de tener una arquitectura bien definida. La Figura 5 muestra los componentes de un sistema BI [5].

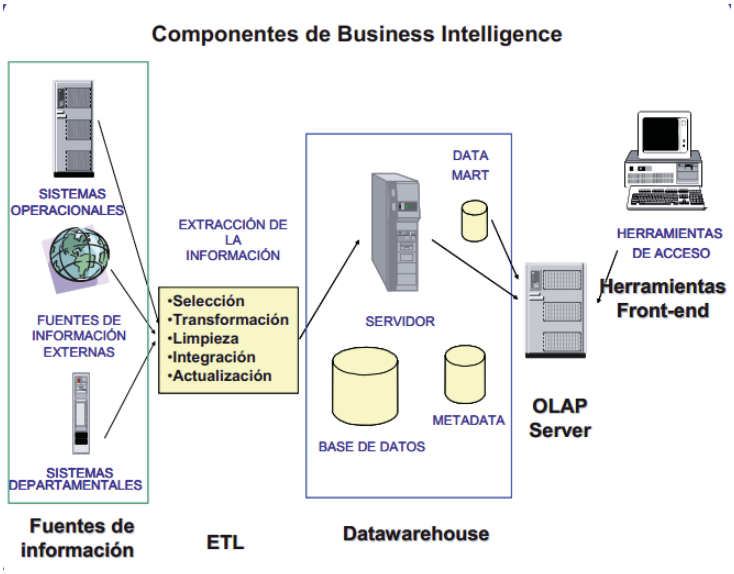


Figura 5: Componentes de una Arquitectura Business Intelligence

7.2.1. Fuentes de información

En el componente Fuentes de información se definen todas las fuentes que se utilizaran en la obtención de los datos transaccionales de la organización. Estas pueden ser archivos de texto, hojas de cálculo, páginas web, mails, bases de datos de distintos sistemas ERP, CRM, MRP o cualquier fuente de información externa que servirán para alimentar el datawarehouse.

El primer paso es analizar si la información que se dispone es la realmente necesaria para alimentar los modelos de negocio que se pretende adoptar. Una vez decididas las fuentes de información se debe verificar la calidad de los datos, ya que, si la información no es precisa ni válida, no puede responder de las decisiones basadas en ella [33].

7.2.2. Proceso de extracción, transformación y carga (ETL)

El proceso de ETL consume entre el 60 % y el 80 % del tiempo de un proyecto de Business Intelligence, por lo que es un proceso clave en la vida de todo proyecto [34], donde se toman los datos desde las fuentes de información y luego de unos subprocesos internos, preparan los datos listos para alimentar el datawarehouse.

El proceso de ETL se divide en 5 subprocesos:

- **Extracción**: Este proceso recupera los datos físicamente de las distintas fuentes de información. En este momento se dispone de los datos en bruto.
- **Limpieza**: Este proceso recupera los datos, comprueba su calidad, elimina los duplicados, corrige los valores erróneos y completa los valores vacíos para reducir los errores de carga. Así también, la limpieza de datos se divide en distintas etapas:

- **Depurar los valores**: Este proceso identifica los elementos individuales de información en las fuentes de datos y los aleja en los ficheros destino.
- **Corregir**: Este proceso corrige los valores individuales de los atributos usando algoritmos de corrección y fuentes de datos externas.
- **Estandarizar**: Este proceso aplica rutinas de conversión para transformar valores en formatos definidos aplicando procedimientos de estandarización definidos por las reglas del negocio.
- **Relacionar**: Este proceso busca y relaciona los valores de los registros, corrigiéndolos y estandarizándolos, basándose en reglas de negocio para eliminar duplicados.
- **Consolidar**: Este proceso analiza e identifica relaciones entre registros relacionados y los junta en una sola representación.

- **Transformación**: Este proceso recupera los datos limpios y de alta calidad para estructurarla en los distintos modelos de análisis. El resultado de este proceso es la obtención de datos limpios, consistentes y útiles.
- **Integración**: Este proceso valida que los datos que cargamos en el datawarehouse son consistentes con las definiciones y formatos del datawarehouse.
- **Actualización**: Este proceso es el que nos permite añadir nuevos datos al datawarehouse.

7.2.3. Datawarehouse

Un datawarehouse es un repositorio de datos que proporciona una visión global, común e integrada de los datos de la organización para soportar las distintas aplicaciones de toma de decisiones. Según Bill Inmon [35], un datawarehouse debe cumplir con las siguientes características:

- **Orientado a un área**: Cada parte del datawarehouse está organizado de manera que todos los elementos pertenecientes a la misma área queden unidos entre sí.
- **Integrado**: La información debe ser transformada en medidas, códigos y formatos comunes para que pueda ser útil para todos los sistemas operacionales, en otras palabras, los datos deben ser consistentes entre departamentos.
- **Indexado en el tiempo**: La información se mantiene de forma histórica y los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar y reflejen esas variaciones.
- **No volátil**: La información no se modifica ni se elimina. Una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.

Existen principalmente dos tipos de esquemas para estructurar los datos en un datawarehouse:

7.2.3.1. Esquema en estrella

Este esquema, ejemplificado [36] en la Figura 6, consiste en estructurar la información en procesos, vistas y métricas recordando a una estrella. Su forma de diseño consiste en una tabla de hechos, que es el centro de objeto de análisis, donde se encontraran los atributos destinados a medir y una o varias tablas de dimensión por cada punto de vista de análisis que participa de la descripción de ese hecho. La tabla de hechos sólo presenta relaciones con dimensiones.

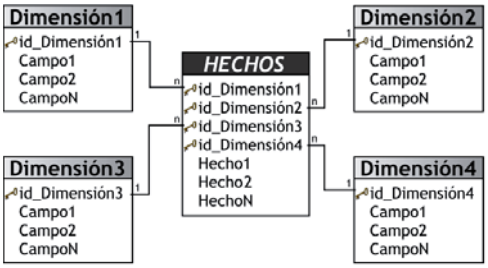


Figura 6: Ejemplo de Esquema Estrella

7.2.3.2. Esquema en copo de nieve

Este esquema, ejemplificado [36] en la Figura 7, esta derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas. Es por esta razón, que la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas y así aparecen nuevas uniones.

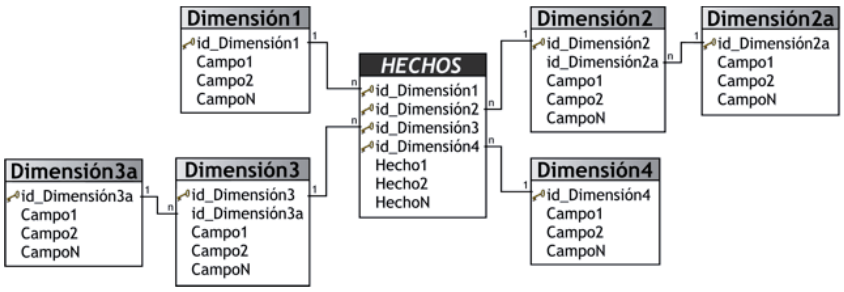


Figura 7: Ejemplo de Esquema Copo de Nieve

7.2.3.3. DataMart

Construir un datawarehouse corporativo puede generar complicaciones, ser costoso y requerir plazos de tiempo que las organizaciones no tienen, es por esta razón que se originaron los DataMart. Los DataMart están dirigidos a una comunidad de usuarios dentro de la organización, que puede estar formada por los miembros de un departamento o por los usuarios de un determinado nivel organizativo, además, de almacenar información de un número limitado de áreas en específico y normalmente se definen para responder a usos muy concretos. Los DataMart son

más pequeños que los datawarehouses, teniendo menos cantidad de información, menos modelos de negocio y son utilizados por un número inferior de usuarios.

Los DataMart [5] pueden ser independientes y alimentarse directamente de los orígenes de información, como lo indica la Figura 8, o ser dependientes y alimentarse desde el datawarehouse corporativo, como lo indica la Figura 9.

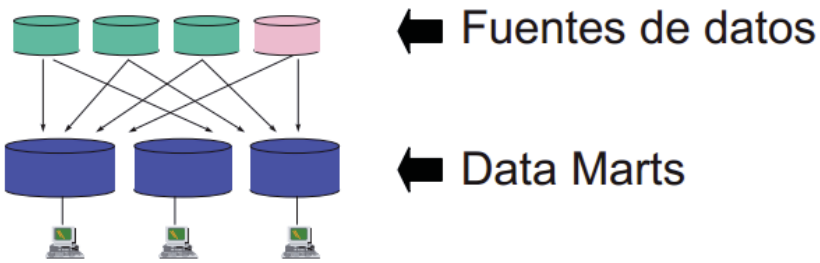


Figura 8: DataMart Independiente

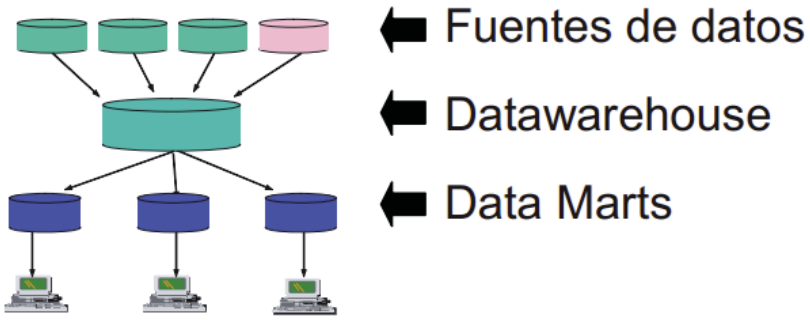


Figura 9: DataMart Dependiente

7.2.4. Almacén multi-dimensional (OLAP)

Los datos del data warehouse se transforman en miembros de dimensiones y en valores para las métricas de los Cubos de Información OLAP. El término OLAP fue presentado en 1993 por Codd [37] y es un tipo de procesamiento de datos que se caracteriza por permitir el análisis multidimensional. La ventaja de un almacén multi-dimensional es que está previamente definida la manera en que se realizarán las consolidaciones de datos, tanto por la propia estructura del cubo, como por la existencia de niveles jerárquicos en las dimensiones.

Dependiendo de la base de datos a utilizar, OLAP se subdivide dentro de dos grandes categorías:

- **OLAP Multidimensional (MOLAP)**: Son herramientas de almacenamiento de datos en un sistema de base de datos multidimensional propio.

- **OLAP Relacional (ROLAP)**:son herramientas que simulan un modelo multidimensional con una base de datos relacional basada en un esquema estrella o copo de nieve.

7.2.5. Herramientas Front-end

Es la herramienta disponible para el usuario final. Con ella se puede explotar la información del BI de distintas maneras:

- Mediante reportes pre-definidos, de acuerdo a cada nivel de la organización.
- Realizando un análisis en línea, también llamado reportes Ad-hoc, utilizando directamente los cubos, e interactuando con las dimensiones ubicándolas en el eje de análisis más adecuado.
- Con Cuadros de Mando y Dashboard para monitorear el desempeño de un área específica o la totalidad de la empresa, con una visión operativa, táctica y estratégica.

7.3. Modelos en Minería de Datos

Existen distintos modelos [38] y algoritmos que se pueden aplicar dentro del proceso de Minería de Datos, por lo que es importante tener una clasificación de los métodos existentes. La elección del método depende del problema en estudio o el tipo de datos a utilizar, el proceso de extracción de datos se rige por las aplicaciones, por esta razón, los métodos utilizados se pueden clasificar de acuerdo con el objetivo de los análisis.

7.3.1. Modelos Descriptivos

Los modelos descriptivos, simétricos o aprendizaje no supervisado permiten formar grupos de datos rápidamente. Las observaciones son generalmente clasificadas en grupos que no son conocidos con anterioridad, los elementos de las variables pueden estar conectados entre sí de acuerdo a vínculos desconocidos de antemano, de esta manera, todas las variables disponibles son tratados en el mismo nivel y no hay hipótesis de causalidad.

7.3.2. Modelos Predictivos

Los modelos predictivos, o de aprendizaje supervisado, se basan en entrenar a un modelo o método para poder describir y predecir el comportamiento de diversos datos y sus futuras tendencias desconocidas. La extracción de patrones determina potenciales oportunidades o riesgos para quien analice la información. Los modelos predictivos establecen relaciones entre variados factores y condiciones significativas dentro una gran cantidad de datos, por lo que el análisis del resultado de dicho proceso se considera crítico para la toma de decisiones de la empresa. En el presente proyecto se ha de trabajar con dos métodos predictivos:

7.3.2.1. Perceptrón Multicapa

El perceptrón multicapa (MLP) es una red neuronal multicapa que se deriva del concepto del perceptrón simple, presentando varias capas de neuronas, en lugar de una sola neurona. Estas capas son llamadas capas ocultas, puesto que no pertenecen ni a la entrada ni a la salida de

la red. Esta red tiene la característica de no ser lineal, es decir, es capaz de clasificar entradas que pertenecen a dos o más clases que no son linealmente separables, siendo capaz de clasificar entradas que pertenecen a dos o más clases que no son linealmente separables.

7.3.2.2. Redes Bayesianas

Las redes bayesianas son una técnica que pertenece al grupo de las técnicas de clasificación y consiste en un modelo gráfico que utiliza arcos para formar una gráfica acíclica y es aplicado en aquellas situaciones en que la incertidumbre se asocia con un resultado que se puede expresar en términos de probabilidad. Esta técnica busca determinar relaciones causales que expliquen un fenómeno y es aplicado en aquellos casos que son de carácter predictivo.

7.3.3. Máquina de soporte vectorial

Las Máquinas de Vectores Soporte (Support Vector Machine, por sus siglas en ingles SVM) son una de las más poderosas técnicas de aprendizaje automático, que a pesar de su sencillez ha demostrado ser un algoritmo robusto y que se adapta bien a problemas de la vida real [39]. Las SVM fueron propuestas por Vladimir Vapnik en 1992 y permiten resolver problemas de clasificación y de regresión [40].

7.3.3.1. Optimización mínima secuencial

La Optimización Mínima Secuencial (Sequential Minimal Optimization, por sus siglas en ingles SMO) es un algoritmo, inventado por el científico John Platt, que resuelve rápidamente el problema de programación cuadrática sin ningún tipo de almacenamiento de la matriz extra y sin usar pasos numéricos de optimización de programación cuadrática en absoluto [41]. Este algoritmo nace a partir de la idea del método de descomposición a su extremo, al optimizar un subconjunto mínimo de solo dos puntos para cada iteración. Lo principal de esta técnica reside en el hecho de que el problema de optimización para dos puntos admite una solución analítica, eliminando la necesidad de usar un optimizador de programación cuadrática [42].

8. Modelo propuesto para el desarrollo de proyectos BI

Este modelo fue propuesto por Larissa T. Moss [31] y cubre todos los aspectos del ciclo de vida de un proyecto de BI, presentando una organización de sus actividades, adecuando las características específicas y contexto del negocio. El modelo consta de dieciséis pasos distribuidos en seis etapas, entregando de forma clara y ordenada cada uno de los pasos a seguir para el desarrollo de un proyecto BI, resultando ideal para ilustrar cada una de las etapas del desarrollo.

8.1. Etapa de justificación

- **Primer paso:** Realizar una evaluación del negocio, esto consiste en definir el problema de forma clara y la oportunidad de negocio que existe, así como la solución de BI propuesta.

8.2. Etapa de planificación

- **Segundo paso**: Realizar una evaluación de la infraestructura empresarial, determinando con qué infraestructura técnica y no técnica cuenta la organización. Ejemplos de infraestructura son: bases de datos, hardware, software, redes, guías, estándares, procedimientos y metodologías.
- **Tercer paso**: Realizar una planificación del proyecto determinando el ámbito del proyecto y las posibles necesidades del personal, el presupuesto requerido, la tecnología necesaria y revisar los patrocinadores del proyecto a nivel interno requeridos para cada etapa del proyecto.

8.3. Etapa de análisis del negocio

- **Cuarto paso**: Realizar una definición de los requerimientos del proyecto.
- **Quinto paso**: Realizar un análisis de los datos evaluando la calidad de los datos existentes y sus fuentes.
- **Sexto paso**: Construir un prototipo de la aplicación no funcional de lo que se desea implementar.
- **Séptimo paso**: Realizar un análisis del repositorio del metadata técnico y hacerlo corresponder con el de negocios. En este paso se determinará cuál metadata capturar y almacenar.

8.4. Etapa de diseño

- **Octavo paso**: Realizar un diseño de la base de datos, de acuerdo a las necesidades de información obtenidas en los requerimientos, además, se diseñarán los esquemas de la base de datos multidimensional.
- **Noveno paso**: Realizar el diseño del proceso de ETL de sus distintas fuentes de datos hacia la base de datos multidimensional.
- **Décimo paso**: Diseñar el repositorio del metadata determinando si el repositorio del metadata se construye o se utiliza una licencia.

8.5. Etapa de construcción

- **Undécimo paso**: Desarrollar el proceso de ETL según las necesidades de transformación de datos que se determinaron en el paso de análisis de datos y el diseño del ETL. En este paso se determinar si será necesario adquirir una herramienta de ETL.
- **Duodécimo paso**: Desarrollar la aplicación, teniendo en cuenta que prototipo diseñado inicialmente cumplío con los requerimientos establecidos. El desarrollo de la aplicación comienza en sus aspectos de acceso y análisis de los datos.
- **Decimotercer paso**: Construir una aplicación de minería de datos utilizando una herramienta apropiada para el proyecto.
- **Decimocuarto paso**: Desarrollar el repositorio de la metadata integrándose un equipo de desarrolladores que se encargue de su construcción.

8.6. Etapa de despliegue

- **Decimoquinto paso**: Establecer un cronograma de capacitación, con el fin de que todos los usuarios del sistema desarrollado aprendan a utilizarlo. En este punto las labores de soporte y mantenimiento comienzan tanto del sistema como de las bases de datos y el repositorio de la metadata.
- **Decimosexto paso**: Evaluar el sistema entregado; sus herramientas, técnicas y guías utilizadas y así determinar si realmente fueron útiles o si deben ser descartadas para futuros proyectos. Además, se analizarán los retrasos en las distintas fases del proyecto, si hubo sobrepaso en el presupuesto inicial, las disputas presentadas y cómo se resolvieron.

9. Aplicación de Solución

Como solución integral se pretende realizar un sistema BI que brinde un indicador que permita clasificar a los alumnos en riesgo de deserción de la Escuela de Ingeniería Informática de la PUCV, entregando facilidades para realizar un análisis de los datos óptimo y eficiente, descubrir patrones entre estudiantes y así generar estrategias para combatir la deserción por parte de directivos y jefes de docencia. Para la construcción y diseño de la solución de BI se basó en el modelo propuesto por Larissa T. Moss [31], acerca de las mejores prácticas para el desarrollo de proyectos de BI, presentado en la sección 8. Un esquema de la solución propuesta esta graficada en la Figura 10

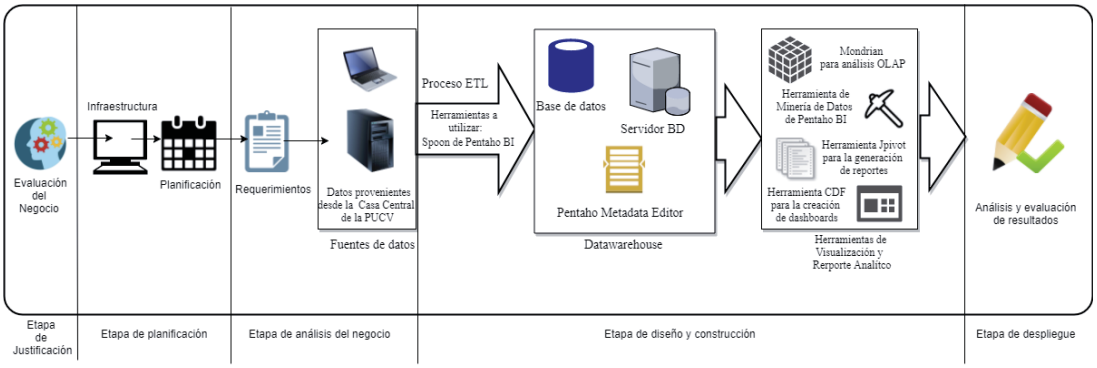


Figura 10: Componentes de la solución BI a realizar

9.1. Etapa de Justificación

Durante esta etapa se realiza una evaluación del negocio identificando la situación actual y la problemática dentro de la Escuela de Ingeniería Informática, justificando la realización del proyecto y la explotación de los datos que se encuentran disponibles, realizando un estudio de la situación actual de los indicadores entregados por la universidad y las necesidades de directivos y jefes de docencia para tomar acciones en relación al problema de la deserción en la carrera viendo la oportunidad y beneficios que trae su implementación dentro de la Escuela, como se observa en la Sección 4 y la Sección 5.

9.2. Etapa de planificación

Durante la etapa de planificación se realiza una estimación de la duración del desarrollo del proyecto, tomando en cuenta la duración de un año correspondientes a los ramos semestrales de Proyecto 1 y 2, donde en el primero se abarca toda la teoría, datos, herramientas, metodologías y plan de implementacióna para el segundo semestre.

9.3. Etapa de Análisis del Negocio

Para la etapa de análisis del negocio, sección 8.3, se ha definido ya el requerimiento de un indicador de deserción con mayor precisión y detalle, por lo que ahora es necesario realizar un

análisis de datos el cual se detalla a continuación.

9.3.1. Análisis de datos

Durante la fase de análisis de datos se pretende recolectar diferentes datos de los estudiantes de la Escuela de Ingeniería Informática de la PUCV, obtenidos desde diferentes sistemas de información provenientes de la Casa Central.

La selección de datos a utilizar para este proyecto se tomó en cuenta principalmente a los factores determinantes de la deserción universitaria, descritos en la Figura 1 en la Sección 5.2, en la cual se presentan factores asociados que permiten predecir la deserción, estos factores son el académico, el socioeconómico, los institucionales e individuales. Por otro, para elegir la mejor selección de atributos para este modelo se ha tomado en cuenta un proyecto similar, para determinar riesgos de deserción en alumnos de la Universidad de las Américas en Santiago [29] y en el trabajo desarrollado por Martha Artunduaga [43] sobre variables que influyen en el rendimiento académico en la universidad.

Los atributos más destacados a utilizar en la solución propuesta se presentan en la Figura 11. Donde el atributo “Estado”, viene a ser la clase predictiva para realizar el indicador de deserción durante el proceso de minería de datos, dentro del sistema BI.

Tipo	Atributo	Descripción	Posibles Valores
Individuales	RUT	Número Identificador del alumno	Numeros del rut sin - y digito verificador.
	Nombre	Nombre y Apellidos del Estudiante	Todos el nombres del alumno
	Sexo	Genero del Estudiante	M (Masculino), F (Femenino)
	CodigoPostal Ciudad	Codigo Postal de residencia del estudiante	Valor numerico
	Región Ciudad	Región de la ciudad	Valores numericos, 1 hasta 15
	CiudadResidencia	Ciudad de Residencia del Estudiante	Nombre de la ciudad
	TipoResidencia	Indicador si vive con familia, amigos o solo.	1. Solo 2.Acompañado
Académicos	Egreso Enseñanza Media	Año en que el estudiante egreso de enseñanza M.	Año en que egreso.
	Cohorte	Fecha de Ingreso a la Universidad	Fecha del año de ingreso
	Codigo Carrera	Codigo para identificar la carrera	Valor numerico
	Carrera	Carrera al cual entro el estudiante	Nombre Carrera
	Preferencia	Preferencia de la carrera al momento de postular	Valor numerico
	Tipo de ingreso	Forma en que el estudiante entro a la Carrera	1. PSU 2.BEA 3.Cambio Interno 4.Traslado Universidad 5.Intercambio Graduado
	Nem	Puntaje NEM del estudiante	Rango de puntajes desde el 4 al 8.5 Numerados del 1 al 6
	PjePSULeng	Puntaje PSU Lenguaje	Rango puntaje desde 200 a 850. Numerados del 1 al 6
	PjePSUMat	Puntaje PSU Matematicas	Rango puntaje desde 200 a 850. Numerados del 1 al 6
	PjePSUCscHist	Puntaje PSU Ciencias o Historia	Rango puntaje desde 200 a 850 Numerados del 1 al 6
	Año Puntaje	Año de selección del puntaje para entrar	1. Puntaje utilizado por primera vez. 2. Puntaje utilizado por segundo año
	PromRank	Promedio Ranking del alumno	Rango puntaje desde 200 a 850. Numerados del 1 al 6
	PromNem	Promedio NEM del estudiante	Rango puntaje desde 200 a 850. Numerados del 1 al 6
	PromPuntaje	Promedio del puntaje del estudiante	Rango puntaje desde 200 a 850. Numerados del 1 al 6
	PromPonderado	Promedio Ponderado del estudiante	Rango puntaje desde 200 a 850. Numerados del 1 al 6
Institucionales	Nombre Colegio	Nombre Colegio del que salio el estudiante	Nombre del Colegio
	CiudadColegio	Ciudad donde se encuentra el Colegio	Nombre Ciudad
	TipoColegio	Tipo administraciión del Colegio	1. Municipal 2. Particular Sub 3. Particular 4. Desconocido
	TipoEnseñanza	Tipo de enseñanza del Colegio	1.H1 2.H2 3.T1 4.T2 5.T3 6.T4
Socio Económicos	NvlEduPadre	Nivel Educacional Padre	1.E Basica 2.E. Media Inc. 3. E.Media C 4. E. Tecnica I 5. E. Tecnica C 6. E Universitaria C 7. E. Universitaria I. 8.Estudios Postgrado
	NvlEduMadre	Nivel Educacional Madre	1.E Basica 2.E. Media Inc. 3. E.Media C 4. E. Tecnica I 5. E. Tecnica C 6. E Universitaria C 7. E. Universitaria I. 8.Estudios Postgrado
	TrabajoEstudiante	Trabajo del estudiante	SI, NO
Clasificador	Estado	El estado (clase) a predecir	1. Cursando 0. Desertor

Figura 11: Atributos a considerar relativos de los estudiantes

Luego de esto se da finalizada la etapa de análisis del negocio y datos, y se comienza la etapa de diseño.

9.4. Etapa de diseño

9.4.1. Diseño Base de Datos para Datawarehouse

Para la creación e implementación del modelo para dar soporte al Datawarehouse se ha de basar en un esquema estrella como se detalla en la Fig. 12, constituido por una tabla de “hechos”, en este caso sería la tabla “Estudiante” que es la que se desea medir, y las distintas tablas de dimensión “Ciudad”, “Colegio”, “DatosAcademicosEstudiante” y “NivelSocioEconomico” que son como queremos medir nuestro modelo. Las ventajas de este modelo sobre otro es que es más sencillo de implementar, entendible para todos y si existen errores, es más fácil detectarlos.

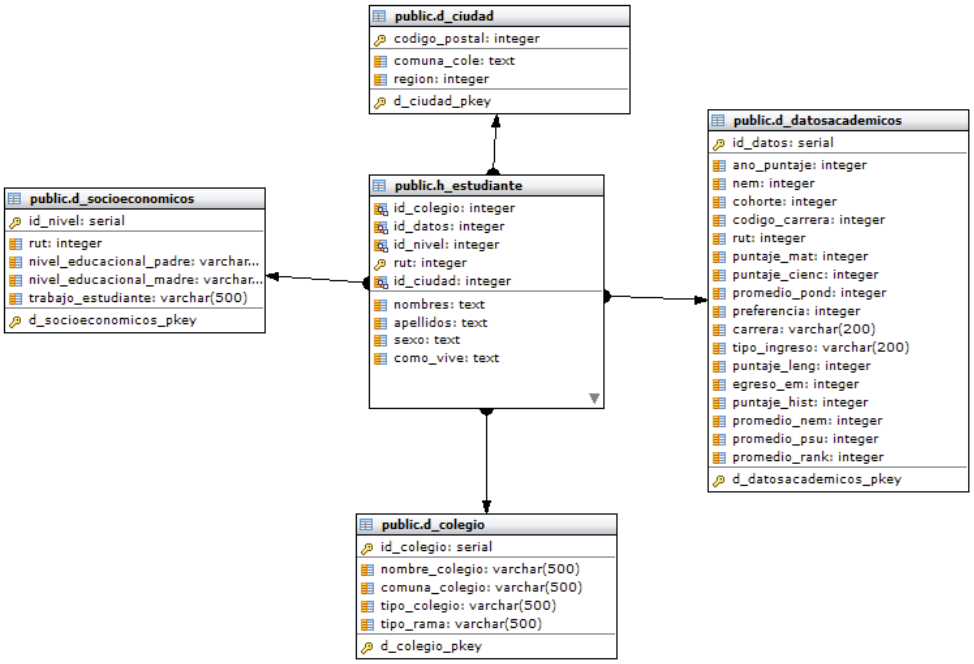


Figura 12: Estructura de datos basado en esquema estrella a utilizar dentro del DW

9.4.2. Diseño del ETL

Para realizar el proceso de ETL se debe diseñar en que consistirán los pasos a desarrollar en este proceso:

- **Extracción de los datos (Extract)**: Se recopilará toda la información disponible desde nuestras fuentes de datos provenientes desde la Casa Central de la PUCV, para poder obtener las estructuras de datos definidas anteriormente de base para el datawarehouse.
- **Limpieza de los datos (Clean)**: Durante este paso se comprobará la calidad de los datos, la eliminación de duplicados, corrección de valores erróneos y agregar valores vacíos para evitar errores de carga.
- **Transformación de los datos(Transform)**: Se tomará la estructura de datos base para extraer los campos que conformarán cada una de las tablas de dimensiones del datawarehouse. A cada uno de los registros de cada dimensión se le asociará un identificador clave, para luego construir la tabla de hechos en base a los campos claves de cada tabla de dimensión quedando como resultado la base de datos descrita en la sección 9.4.1.
- **Carga de los datos (Load)**: En este proceso se validará que los datos cargados en el datawarehouse sean consistentes con los formatos de éste y así, cargar y actualizar los nuevos datos al datawarehouse.

9.4.3. Selección de herramientas utilizadas para la solución

Un proceso importante es la selección de las herramientas utilizadas para desarrollar la solución del proyecto. Se ha de seleccionar Pentaho BI una suite de herramientas de código abierto comercial para BI que ofrece la arquitectura y la infraestructura necesaria para construir soluciones de inteligencia empresarial. Esta proporciona los servicios básicos, incluyendo la autenticación, registro, auditoría, flujo de trabajo, servicios Web y los motores de reglas. La plataforma también incluye un motor de solución que integra la presentación de informes, análisis, dashboards y componentes de minería de datos para formar una plataforma de BI sofisticada y completa.

Las herramientas para la solución BI a utilizar son:

- Spoon(ex Kettle), de Pentaho, para el proceso de ETL.
- Weka para el proceso de Minería de Datos.
- Pentaho User Console para la generación de reportes y la creación de los dashboards.

Terminada la etapa de diseño es tiempo de pasar a la etapa de construcción del proyecto.

9.5. Etapa de Construcción

Para realizar el proceso de ETL se ocupará la herramienta de integración de datos de Pentaho llamada Spoon, el cual permite hacer el proceso de transformación y carga de manera automatizada y rápida. El detalle de este proceso se encuentra detallado en el Anexo B.

Para el proceso de Data Mining, se utilizará la herramienta Weka, el cual servirá para clasificar a los estudiante, mediante el estado en que se encuentran en la universidad(Cursando o Desertor), utilizando los datos, presentados en la Figura 11, de los estudiantes de las generaciones 2015 y 2016, para el proceso de training y del cohorte 2017 para el proceso de testing. El detalle del proceso de Minería de Datos se encuentra explicado en el Anexo C.

Por último, se realizarán los reportes, análisis y dashboards necesarios mediante Pentaho User Console y se encuentra detallado en el Anexo D.

9.6. Etapa de Despliegue

Para la última etapa se analizarán los resultados obtenidos durantes las pruebas del proyecto. Comprobando si los modelos de datos, herramientas y técnicas utilizadas fueron útiles o si deben ser revisadas en caso de obtener resultados no esperados.

Las pruebas realizadas utilizaron:

- Los modos de entrenamiento Supplied Test Set y Porcentage Split, ambos detallados en la Sección C.2.
- Para cada tipo de entrenamiento la utilización de datos random, detallados en la sección B.1, y sin estos datos.

- Los algoritmos predictivos NaiveBayes, MLP y SMO.

Los datos utilizados para las pruebas fueron de estudiantes de la carrera de Ingeniería Civil Informática e Ingeniería en Ejecución Informática de las generaciones 2015, 2016 y 2017. Un resumen de la cantidad de datos se presenta en la Tabla 1.

	Matriculados en el año 2017	No Matriculados en el año 2017	% Alumnos Matriculados	% Alumnos no Matriculados	Total
Alumnos Cohorte 2015	164	40	80.39 %	19.61 %	204
Alumnos Cohorte 2016	152	52	74.51 %	25.49 %	204
Alumnos Cohorte 2017	200	0	100 %	0 %	200

Tabla 1: Tabla resumen matriculados año 2017

9.6.1. Prueba del modelo

Antes de realizar la predicción de los estudiantes del cohorte 2017 para el año 2018, es necesario comprobar si el modelo de predicción funciona, para esto se ha de realizar una prueba con las datos y matriculas reales de los estudiantes del cohorte 2015 y 2016, y comprobar si los resultados obtenidos del training y testing son similares a las matriculas reales del cohorte 2016 para el principio de este año 2017. Los resultados de estas pruebas de Clasificación están detallados en el Anexo E y un resumen de las pruebas se puede observar en la Tabla 2.

Resultados Porcentage Split (No Random)				Resultados Porcentage Split (Random)			
MLP	Número de	%	Errores	MLP	Número de	%	Errores
Matriculados	154	75,4902 %	3	Matriculados	154	75,4902 %	2
No Matriculados	50	24,5098 %		No Matriculados	50	24,5098 %	
SMO	Número de	%		SMO	Número de	%	
Matriculados	179	87,7451 %	34	Matriculados	170	83,3333 %	28
No Matriculados	25	12,2549 %		No Matriculados	34	16,6667 %	
NaivesBayes	Número de	%		NaivesBayes	Número de	%	
Matriculados	167	81,8627 %	45	Matriculados	171	83,8235 %	43
No Matriculados	37	18,1373 %		No Matriculados	33	16,1765 %	
Resultados Supplied test set (No Random)				Resultados Supplied test set (Random)			
MLP	Número de	%	Errores	MLP	Número de	%	Errores
Matriculados	155	75,9804 %	5	Matriculados	154	75,4902 %	2
No Matriculados	49	24,0196 %		No Matriculados	50	24,5098 %	
SMO	Número de	%		SMO	Número de	%	
Matriculados	178	87,2549 %	32	Matriculados	170	83,3333 %	23
No Matriculados	26	12,7451 %		No Matriculados	34	16,6667 %	
NaivesBayes	Número de	%		NaivesBayes	Número de	%	
Matriculados	168	82,3529 %	48	Matriculados	171	83,8235 %	43
No Matriculados	36	17,6471 %		No Matriculados	33	16,1765 %	

Tabla 2: Resultados prueba predicción de cohorte 2016 para el año 2017

La Tabla 2 muestra la cantidad de alumnos que el modelo predijo gracias a los distintos tipos de entrenamiento, algoritmos y datos utilizados. Los mejores resultados se encuentran al utilizar el algoritmo del Perceptrón Multicapa(MLP) al obtener un 75 % de alumnos con estado matriculado cercano a el valor real (74.51 %) y con una muy baja cantidad de predicciones del estado actual de los alumnos de forma errónea, tanto como para el Porcentage Split como para el Supplied Test Set, con sólo dos alumnos mal clasificados.

En conclusión, Supplied Test Set, con una configuración de datos random, viene a ser el mejor resultado de las pruebas. El detalle de ésta se encuentra detallado en el Anexo F.1, en las Tablas 4, 5, 6, 7, 8 y 9.

9.6.2. Predicción cohorte 2017

Al momento de realizar la clasificación, el algoritmo entrega una serie de indicadores, los cuales sirven para determinar la eficacia, relevancia, entre otros aspectos del clasificador, lo que ayuda al posterior análisis de resultados. Entre los indicadores arrojados por los algoritmos se encuentran:

- **F-score**: Es la medida de precisión que tiene un test, se considera una medida armónica que combina valores de precisión y exhaustividad.
- **TP Rate**: Tasa de verdaderos positivos.
- **FP Rate**: Tasa de falsos positivos.
- **Precisión**: Tasa de precisión del modelo.

9.6.3. Resultado proceso de clasificación

A continuación, en la Tabla 3, se resume los resultados del proceso de clasificación de estado de los estudiantes de la Escuela de Ingeniería Informática de la PUCV, del año 2017. Se ha de utilizar como método de entramiento Supplied Test Set, junto a los valores random ya que según las pruebas realizas en la Sección 9.6.1, son las que presentan una mayor cercanía a los valores reales del estado actual de los estudiantes de la Escuela de Ingeniería Informática.

Supplied Test Set - Random			
Aspectos	MLP	SMO	NaivesBayes
Instancias Correctas	76 %	84,5 %	57 %
Instancias Incorrectas	24 %	15.5 %	43 %
Verdaderos Positivos	0,760	0,845	0.570
Falsos Positivos	0,240	0,155	0,430
F-Score	0,864	0,916	0,726

Tabla 3: Resultados pruebas de predicción de deserción de los estudiantes del cohorte 2017

Como se puede observar en la Tabla 3, el mejor resultado lo obtiene MLP al tener un porcentaje de estudiantes matriculados dentro de los rangos que cada año va obteniendo la Escuela de Ingeniería Informática de la PUCV y un F-SCORE aceptable para medir la precisón y exhaustividad del modelo propuesto. Por otro lado, SMO presenta valores de predicción algo alejados a la realidad de la retención estudiantil de la Escuela, pero presenta un buen F-Score y su % de clasificación sigue la misma linea que las pruebas durante el año 2016. Por último, NaiveBayes muestra los valores mas deficientes, dentro de las pruebas realizadas, teniendo el valor de instancias correctamente predichas más bajo, junto a su F-Score.

Para ver el detalle de cada estudiante, y como lo clasificó el proceso de Minería de Datos, se debe observar las Tablas 10, 11, 12, 13, 14 y 15 del Anexo F.2. En éstas se encuentra el rut del estudiante, el estado inicial de cada uno, el estado luego del proceso de clasificación, si el algoritmo se equivocó y, por último, la probabilidad de que la instancia clasificada pertenezca realmente a la clase predicha.

10. Conclusión

En primera instancia, la investigación de los términos relacionados con BI permitió visualizar el espectro de lo que significa dicho concepto hoy en día y que su aplicación no sólo se utiliza en el mercado de retail y empresas con fines de lucro trascendiendo a cualquier tipo de organización, por lo que su uso representa muchos beneficios, así como grandes ventajas competitivas con los demás.

Dados los problemas que afectaban al sistema de indicadores de gestión del Navegador Académico, y centrándose en el problema de la deserción a nivel de los estudiantes de la Escuela de Ingeniería Informática. Se logró crear una propuesta de solución adecuada a los requerimientos de información que los directivos y jefes de docencia necesitan para identificar a estudiantes en riesgo de deserción y tomar decisiones para enfrentar dicho problema con una base de fundamentos clara y concisa.

A futuro se espera tener más datos de distintos periodos de alumnos que han pasado por la Escuela de Ingeniería Informática y realizar diversas pruebas, y así pulir más la solución a implementar.

Por último, se deja la reflexión de que si el uso de metodologías, herramientas gerenciales y técnicas de BI fueran más masivas dentro de las instituciones universitarias estas realmente podrían organizarse mejor, analizar más efectivamente la información que poseen, aprovechando las oportunidades de innovar y responder adecuadamente a los requerimientos del quehacer universitario que permitan dar cumplimiento a las metas institucionales y entregar una educación de mayor calidad para sus estudiantes.

11. Referencias

[1] T. Bäck. *Adaptive Business Intelligence Based on Evolution Strategies: Some Application Examples of Self-Adaptive Software*, pages 1–9. Springer London, London, 2005.

[2] J. Nader. *Sistema de Apoyo Gerencial Universitario*. Ts, Instituto Tecnológico de Buenos Aires, Buenos Aires - Universidad Politécnica, Madrid, 2004.

[3] Jing Luan. *Data Mining and Its Applications in Higher Education*. Toronto, 2002.

[4] Jobany José Heredia-Rico, Aida Georgina Rodríguez-Hernández, and José Alberto Vilalta-Alonso. El análisis de datos en apoyo a la gestión de la enseñanza en la carrera ingeniería industrial. *Ingeniería Industrial*, 33:19 – 30, 04 2012.

[5] J.L. Cano. *Business intelligence: competir con información*. Banesto, Fundación Cultur [i.e. Cultural], 2007.

[6] Luis Eduardo González and Daniel Uribe. Estimaciones sobre la repitencia y deserción en la educación superior chilena. consideraciones sobre sus implicaciones. *Revista Calidad en la educación*, pages 75–90, 2002.

[7] *Elevada deserción en educación superior*. `http://www.elmercurio.com/blogs/2015/02/22/29587/Elevada-desercion-en-educacion -superior.aspx`, 2015.

[8] *Chile pierde US$ 780 millones al año por universitarios que no terminan sus carreras*. `http://www.latercera.com/noticia/chile-pierde-us-780-millones-al-ano-por-universitarios-que-noterminan-sus-carreras/`, 2016.

[9] Universidad Nacional ICFES. *Estudio de la deserción estudiantil en la educación superior en Colombia*. Ts, 2002.

[10] Cheng-Che Shen, Ray-E Chang, Ching Jou Hsu, and I-Chiu Chang. How business intelligence maturity enabling hospital agility. *Telematics and Informatics*, 34(1):450 – 456, 2017.

[11] *Business Intelligence aplicada a la acción social: el ejemplo de Proniño*. `http://blogthinkbig.com/ business-intelligence-accion-social-pronino/`, Agosto 2012.

[12] *Business intelligence y las entidades deportivas*. `https://javier-sobrino.com/2013/10/15/business-intelligence-y-las-entidades-deportivas/`, Octubre 2013.

[13] Isabel Guitart Hormigo y Jordi Conesa i Caralt. Uso de analítica para dar soporte a la toma de decisiones docentes. *JENUI 2014. "XX Jornadas de Enseñanza Universitaria de la Informática". Oviedo: Universidad de Oviedo. Escuela de Ingeniería Informática*, pages 83–90, 2014.

[14] Yusnier Reyes Dixson y Lissette Nuñez Maturel. La inteligencia de negocio como apoyo a la toma de decisiones en el ámbito académico (business intelligence as decision support

system in academic environment). *GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología*, 3(2):63–73, 2015.

[15] Jonathan Narváez Triana, Camilo Monsalve Hernández, Alexander Bustamante Martínez, Ernesto Galvis Lista, and Luis Gómez Flórez. Business intelligence solution for managing educational resources and physical spaces in magdalena university. *AVANCES Investigación en Ingeniería*, 10, 2013.

[16] *Kr. Consulting.* http://dev.redodigital.com/clientes/kr-consulting/maqueta-2/, 2014.

[17] *DATAMART S.A.* http://www.datamart.cl/b3.htm, 2005.

[18] *DATAMART Cubix Olap Analyzer 3.1.* http://www.datamart.cl/pdf/cubix_olap_-analyzer.pdf, 2005.

[19] *MicroStrategy.* https://www.microstrategy.com/es, 2015.

[20] M. Flores. *Proyecto de Business Intelligence para la Universidad de La Serena: Sistema de Apoyo a la Gestión Institucional 2.0.* Ts, Departamento de Informática, Programa de Magíster en Tecnologías de la Información, Universidad Técnica Federico Santa María, 2012.

[21] Luis Fuentes Tapia and Ricardo Valdivia Pinto. Incorporación de elementos de inteligencia de negocios en el proceso de admisión y matrícula de una universidad chilena. *Ingeniare. Revista chilena de ingeniería*, 18:383 – 394, 12 2010.

[22] *Proyecto BI con Pentaho en la Universidad de Zaragoza.* http://www.bi-spain.com/articulo/73879/open-source-software-libre /educacion-y-formacion/proyecto-bi-con-pentaho-en-la-universidad-de-zaragoza -video-de-3-minutos, Enero 2015.

[23] *Universities: Competing through Business Intelligence and analytics.* http://www.yellowfinbi.com/YFCommunityNews-Universities-Competing-through -Business-Intelligence-and-analytics-147594, October 2013.

[24] *Business Intelligence in Education.* https://www.inetsoft.com/solutions /industry/education/, 2016.

[25] *EL SOFTWARE DE INTELIGENCIA DE NEGOCIO DE JASPERSOFT APORTA AL SECTOR EDUCATIVO UNA NUEVA MANERA DE MEDIR EL RENDIMIENTO.* https://www.jaspersoft.com/es/press/el-software-de-inteligencia-de -negocio-de-jaspersoft-aporta-al-sector-educativo-una-nueva-mane, Septiembre 2011.

[26] *Beneficios de IBM Cognos en escuelas primarias y secundarias (organizaciones K-12).* http://www.bi-spain.com/articulo/ 72934/ibm-/educacion-y-formacion/beneficios-de-ibm-cognos-en-escuelas-primarias-y-secundarias-organizaciones-k-12, Noviembre 2012.

[27] *Análisis de datos para un mejor reclutamiento y monitoreo del progreso estudiantil en los Campus con el BI de Oracle.* http://www.bi-spain.com/articulo0/71907/oracle/educacion-y-formacion/analisis-de-datos-para-un-mejor-reclutamiento-y-monitoreo-del-progreso-estudiantil-en-los-campus-con-el-bi-de-oracle-webinar -de-1-hora-en-ingles, Diciembre 2011.

[28] Christian Díaz Peralta. MODELO CONCEPTUAL PARA LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA CHILENA. *Estudios pedagógicos (Valdivia)*, 34:65 – 86, 00 2008.

[29] E. Fischer. *MODELO PARA LA AUTOMATIZACIÓN DEL PROCESO DE DETERMINACIÓN DE RIESGO DE DESERCIÓN EN ESTUDIANTES UNIVERSITARIOS.* Ts, FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN, UNIVERSIDAD DE CHILE, 2012.

[30] Josep Curto Díaz and Jordi Conesa Caralt. *Introducción al business intelligence.* UOC, Barcelona, 1a ed., 5a reimp. ed. edition, 2012.

[31] L.T. Moss and S. Atre. *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-support Applications.* Addison-Wesley information technology series. Addison-Wesley, 2003.

[32] Wayne W. Eckerson y Cindi Howson. *Enterprise Business Intelligence: Strategies and Technologies for Deploying BI on an Enterprise Scale.* https://tdwi.org/articles/2005/10/13/enterprise-business-intelligence-strategies-and-technologies-for-deploying-bi-on-an-enterprise-scale.aspx, October 2005.

[33] Bill Inmon. *Business Intelligence Network.* http://www.b-eye-network.com/, August 2006.

[34] Wayne Eckerson and Colin White. *Evaluating ETL and Data Integration Platforms.* http://download.101com.com /tdwi/research_report/2003ETLReport.pdf, 2003.

[35] W. H. Inmon. *Building the Data Warehouse.* QED Information Sciences, Inc., Wellesley, MA, USA, 1992.

[36] *Datawarehouse manager.* http://www.dataprix.com/datawarehouse-manager#x1-500003.4.5.1, 2009.

[37] E.F. Codd, S.B. Codd, and C.T. Salley. *Providing OLAP (On-line Analytical Processing) to User-analysts: An IT Mandate.* Codd & Associates, 1993.

[38] *MODELOS PREDICTIVOS Y DESCRIPTIVOS EN MINERÍA DE DATOS, UNIVERSIDAD POLITÉCNICA DE TLAXCALA .* https://es.slideshare.net/lalopg/mtodos-predictivos-y-descriptivos-minera-de-datos.

[39] Yvonne Gala García. *Algoritmos SVM para problemas sobre big data.* Te, Escuela Politécnica Superior, Departamento de Ingeniería Informática, Universidad Autónoma de Madrid, 2013.

[40] Rodolfo Iván Flores de la Torre. *Análisis Comparativo de Árboles de Decisión y Máquina de Vectores Soporte para conjuntos de datos de Diabetes y Hepatitis.* Te, Centro Universitario UAEM Zumpango, Universidad Autónoma del Estado de México, 2014.

[41] John C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING, 1998.

[42] Félix Hernán Castro Fuentes. *LS-SVM BASADO EN OPTIMIZACIÓN POR ENJAMBRES DE PARTÍCULAS PARA CLASIFICACIÓN DE ACCIDENTES DE TRÁNSITO* . Te, Escuela de Ingeniería Informática, Faculta de Ingeniería, Pontificia Universidad Católica de Valparaíso, 2010.

[43] Martha Artunduaga Murillo. *VARIABLES QUE INFLUYEN EN EL RENDIMIENTO ACADÉMICO EN LA UNIVERSIDAD.* Ts, Complutense de Madrid, 2008.

[44] *csv2arff: Online CSV - ARFF conversion tool.* `http://slavnik.fe.uni-lj.si/markot/csv2arff/csv2arff.php`.

[45] Diego García Morate. Manual de weka, 2006.

Anexos

A. Ejemplos de indicadores de Gestión de la PUCV

Cada indicador cuenta con distintos tipos de reportes, así como la opción de visualizarlos por unidad académica, carrera, el periodo lectivo actual o pasado, el año de la información y una forma de exportar dicho reporte mediante PDF, HTML y CVC.

A.1. Indicadores de ingreso

El indicador de ingreso cuenta con los siguientes reportes:

A.1.1. Distribución Ingresos Según Sexo (Anual)

Figura A.1: Indicador de Ingreso según sexo de la Carrera Ing. Civil Informática, año 2016.

A.1.2. Matrículas primer año Vía PSU (Anual)

Figura A.2: Indicador de matrículas vía PSU, año 2016, de la carrera Ing. Civil Informática

A.2. Indicadores de proceso formativo

El indicador de proceso formativo cuenta con los siguientes reportes:

A.2.1. Matrículas para Periodos en curso por año de Ingreso

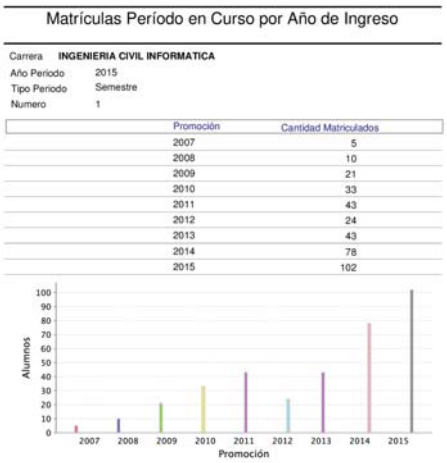


Figura A.3: Indicador de matrículas para periodos en curso, año 2015, Ing. Civil Informática

A.2.2. Matrículas Históricas (Anual)

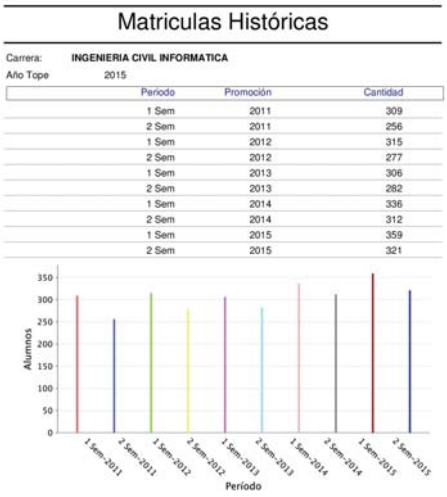
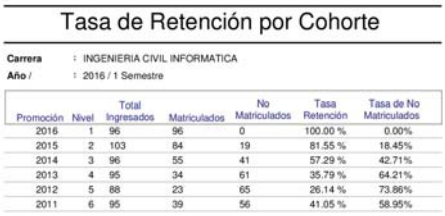


Figura A.4: Indicador de matrículas históricas hasta año 2015, Ing. Civil Informática

A.2.3. Tasa de Retención por Cohorte



Tasa de Retención por Cohorte

Carrera : INGENIERIA CIVIL INFORMATICA
Año / : 2016 / 1 Semestre

Promoción	Nivel	Total Ingresados	Matriculados	No Matriculados	Tasa Retención	Tasa de No Matriculados
2016	1	96	96	0	100.00 %	0.00%
2015	2	103	84	19	81.55 %	18.45%
2014	3	96	55	41	57.29 %	42.71%
2013	4	95	34	61	35.79 %	64.21%
2012	5	88	23	65	26.14 %	73.86%
2011	6	95	39	56	41.05 %	58.95%

Figura A.5: Indicador de retención por cohorte de la carrera de Ing. Civil Informática, año 2016

A.2.4. Ranking

RANKING POR PROMOCIÓN POR PROMEDIO SIMPLE

UNIDAD ACADEMICA: ESCUELA DE INGENIERÍA INFORMÁTICA
PROMOCION: 2011

RUT	NOMBRES		APELLIDO PATERNO	APELLIDO MATERNO	RANKING	PROMEDIO
DOCTORADO EN INGENIERIA INFORMATICA						
16485075	RODOLFO	ANDRES FERNANDO	INOSTROZA	CARVAJAL	1	6.71000
15070015	ENRIQUE	ANDRES	URRA	COLOMA	2	6.68000
22479908	CRISTHY	NATALY	JIMENEZ	GRANIZO	3	6.62500
15070527	DANIEL	ANDRES	CABRERA	PANIAGUA	4	6.50000

Figura A.6: Indicador de Ranking de los estudiantes de Doctorado.

A.2.5. Retiros

RETIROS PARCIALES ANUALES POR ASIGNATURA

Unidad Academica: ESCUELA DE INGENIERÍA INFORMÁTICA

Retiros al: 2011

CARRERA	ASIGNATURA	2007	2008	2009	2010	2011
INGENIERIA CIVIL INFORMATICA	ART-051 - EL ESPACIO EN LAS ARTES PLASTICAS					1
	ART-082 - LAS UTOPIAS: MANIFESTACION UTOPICA EN EL ARTE					1
	EFI-023 - AUTOCUIDADO Y VIDA SALUDABLE					2
	EII-405 - INVESTIGACION DE OPERACIONES					4
	FIS-017 - FISICA Y TAMAÑO					2
	FIS-019 - FISICA DEL SIGLO XX					1
	FIS-133 - FISICA GENERAL MECANICA					16
	FIS-331 - FISICA GENERAL ELECTROMAGNETISMO					2
	FIS-431 - FISICA GENERAL: ONDAS					7
	ICI-140 - INTRODUCCION A LA INGENIERIA INFORMATICA					2
	ICI-142 - FUNDAMENTOS DE PROGRAMACION				1	2
	ICI-241 - ESTRUCTURA DE DATOS					12
	ICI-343 - ORGANIZACION Y MANEJO DE ARCHIVOS					1
	ICI-441 - SISTEMAS DE INFORMACION 1					1
	ICI-447 - INGENIERIA ECONOMICA					10
	ICI-541 - FORMULACION Y EVALUACION DE PROYECTOS					3
	ICI-543 - TALLER DE DESARROLLO DE SISTEMAS					1
	ICI-569 - ENGLISH FOR COMPUTING WORK					1
	ICI-599 - REDES NEURONALES ARTIFICIALES					1
	ICI-641 - PROYECTO 1				1	2
	ICI-642 - PROYECTO 2				1	
	ICI-644 - INFORMATICA Y SOCIEDAD					1

Figura A.7: Indicador de Retiros parciales anuales de la carrera de Ing. Civil Informática, año 2011.

A.2.6. Sanciones por Promoción - Art 28 (Anual)

Sanciones por Carrera

Unidad Académica **ESCUELA DE INGENIERÍA INFORMÁTICA**

Año tope Sanción **2016**

Años		2013		2014		2015		2016		TOTAL
Carrera/Sanción		1 Sem	2 Sem	1 Sem	2 Sem	1 Sem	2 Sem	1 Sem	2 Sem	
INGENIERIA CIVIL INFORMATICA	Art 28	5	4	6	5	6	2	4	0	32
	Art 33	30	61	46	48	30	50	29	0	294
	Art 45	0	3	0	1	0	5	0	0	9
INGENIERIA DE EJECUCION	Art 28	18	18	15	7	27	7	34	0	126
	Art 33	36	52	62	68	22	46	44	0	330
	Art 45	0	4	0	1	0	13	0	0	18

Figura A.8: Indicador de sanciones por carrera Art 28, año 2013, 2014, 2015, 2016

A.2.7. Sanciones por Promoción - Art 33 (Anual)

Sanciones por Promoción

Unidad académica ESCUELA DE INGENIERÍA INFORMÁTICA
Año tope sanción 2016
Art 33

		2011		2012		2013		2014		2015		2016		Total
		1 Sem	2 Sem	1 Sem	2 Sem	1 Sem	2 Sem	1 Sem	2 Sem	1 Sem	2 Sem	1 Sem	2 Sem	
INGENIERIA CIVIL INFORMATICA	2006	0	1	1	1	0	0	0						3
	2007	8	3	0	2	0	2	0	0	0				15
	2008	1	2	4	5	1	4	0	0	0	0	0		17
	2009	3	5	3	5	0	1	0	1	0	1	0	0	19
	2010	1	5	5	12	5	13	2	1	1	1	0	0	46
	2011	0	20	25	21	8	10	8	9	3	7	2	0	113
	2012			1	33	16	9	10	7	1	1	0	0	78
	2013					0	22	25	17	9	10	1	0	84
	2014							0	13	16	18	10	0	57
	2015									0	12	15	0	27
	2016											1	0	1
	Total anual	13	36	39	79	30	61	45	48	30	50	29	0	460

Figura A.9: Indicador de sanciones por carrera Art 33, año 2011, 2012, 2013, 2014, 2015, 2016.

A.2.8. Tasa Aprobación por Carrera (Anual)

Tasa de Aprobación Promedio por Carrera

Unidad Académica : ESCUELA DE INGENIERÍA INFORMÁTICA
Año de Análisis : 2016

Periodo	Total Alumnos	Tasa de Aprobación
INGENIERIA CIVIL INFORMATICA		
2011	301	79,34 %
2012	308	62,98 %
2013	302	74,69 %
2014	336	75,57 %
2015	358	77,93 %
2016	354	43,69 %

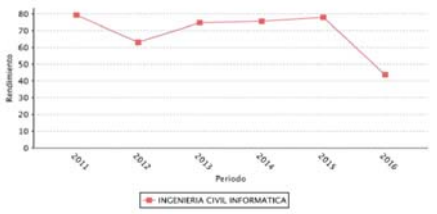


Figura A.10: Indicador de tasa de aprobación por carrera, Doctorado en Ingeniería Informática, año 2011 hasta 2016

A.2.9. Tasa Aprobación por Promoción (Anual)

Tasa de Aprobación Promedio por Promoción

Unidad Académica : ESCUELA DE INGENIERÍA INFORMÁTICA
Máxima Promoción : 2011

Promoción	Período	Total Alumnos	Tasa de Aprobación
DOCTORADO EN INGENIERIA INFORMATICA			
2011			
	2011	4	100,00 %
	2012	4	52,17 %
	2013	4	1,09 %
	2014	3	1,45 %
	2015	4	51,09 %
	2016	1	0,00 %

Figura A.11: Indicador de tasa de aprobación por promoción, Doctorado en Ingeniería Informática, año 2011 hasta 2016

A.2.10. Tasa Aprobación por Alumno (Anual)

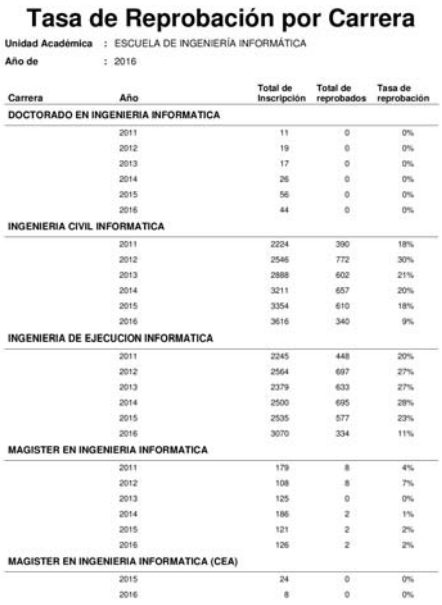
Tasa de Aprobación por Alumno

Unidad Académica : ESCUELA DE INGENIERÍA INFORMÁTICA
Año de Análisis : 2011

Nombre	Período	Total Créditos Inscritos	Total Créditos Aprobados	Tasa de aprobación
DOCTORADO EN INGENIERIA INFORMATICA				
CABRERA PANIAGUA, DANIEL ANDRES - 15070527-4				
	2 SEM 2011	24.0	24.0	100,00 %
	1 SEM 2012	72.0	6.0	8,00 %
	2 SEM 2012	66.0	0.0	0,00 %
	1 SEM 2013	66.0	0.0	0,00 %
	2 SEM 2013	66.0	0.0	0,00 %
	1 SEM 2014	72.0	6.0	8,00 %
	2 SEM 2014	66.0	0.0	0,00 %
	1 SEM 2015	66.0	66.0	100,00 %
			Desempeño:	20,48 %

Figura A.12: Indicador de tasa de aprobación por alumno, Doctorado en Ingeniería Informática, año 2011 hasta 2015

A.2.11. Tasa Reprobación por Carrera (Anual)



Tasa de Reprobación por Carrera

Unidad Académica : ESCUELA DE INGENIERÍA INFORMÁTICA
Año de : 2016

Carrera	Año	Total de Inscripción	Total de reprobados	Tasa de reprobación
DOCTORADO EN INGENIERIA INFORMATICA				
	2011	11	0	0%
	2012	19	0	0%
	2013	17	0	0%
	2014	26	0	0%
	2015	56	0	0%
	2016	44	0	0%
INGENIERIA CIVIL INFORMATICA				
	2011	2224	390	18%
	2012	2546	772	30%
	2013	2888	602	21%
	2014	3211	657	20%
	2015	3354	610	18%
	2016	3616	340	9%
INGENIERIA DE EJECUCION INFORMATICA				
	2011	2245	448	20%
	2012	2564	697	27%
	2013	2379	633	27%
	2014	2500	695	28%
	2015	2535	577	23%
	2016	3070	334	11%
MAGISTER EN INGENIERIA INFORMATICA				
	2011	179	8	4%
	2012	108	8	7%
	2013	125	0	0%
	2014	186	2	1%
	2015	121	2	2%
	2016	126	2	2%
MAGISTER EN INGENIERIA INFORMATICA (CEA)				
	2015	24	0	0%
	2016	8	0	0%

Figura A.13: Indicador de tasa de reprobación por carrera, año 2011 hasta 2016

A.2.12. Tasa Reprobación por Asignatura (Anual)

Unidad Académica : ESCUELA DE INGENIERÍA INFORMÁTICA
Año de Análisis : 2016
MII-784 : OPTIMIZACIÓN COMBINATORIAL

Sigla	Periodo	Total de alumnos	Total de reprobados	Tasa de reprobación
MII-784-1	2 SEM 2012	6	0	0.00 %
MII-784-1	1 SEM 2014	6	0	0.00 %
MII-784-1	1 SEM 2015	5	0	0.00 %
MII-784-1	2 SEM 2015	7	1	14.29 %
MII-784-2	2 SEM 2015	1	0	0.00 %
MII-784-1	2 SEM 2016	4	0	0.00 %

Figura A.14: Indicador de tasa de reprobación por asignatura, año 2012 hasta 2016

A.3. Indicador comparativo

El indicador comparativo cuenta con el siguiente reporte:

A.3.1. Tiempo medio de titulación vs egreso (Anual)

Tiempo Medio Titulación vs Egreso

Unidad Académica ESCUELA DE INGENIERÍA INFORMÁTICA

Máximo Año Titulación/Egreso 2015

	Período	TME	TMT
DOCTORADO EN INGENIERIA	2015		Años : 3.00 Nº Titulados: 1
INGENIERIA CIVIL INFORMATICA	2005	Años : 7.00 Nº Egresados: 27	Años : 8.06 Nº Titulados: 16
	2006	Años : 7.00 Nº Egresados: 19	Años : 8.28 Nº Titulados: 18
	2007	Años : 7.00 Nº Egresados: 37	Años : 8.91 Nº Titulados: 11
	2008	Años : 7.00 Nº Egresados: 26	Años : 9.30 Nº Titulados: 20
	2009	Años : 7.00 Nº Egresados: 30	Años : 8.26 Nº Titulados: 31
	2010	Años : 7.00 Nº Egresados: 26	Años : 9.63 Nº Titulados: 24
	2011	Años : 9.00 Nº Egresados: 7	Años : 9.43 Nº Titulados: 21
	2012	Años : 7.00 Nº Egresados: 39	Años : 8.76 Nº Titulados: 33
	2013	Años : 7.00 Nº Egresados: 14	Años : 9.72 Nº Titulados: 32
	2014	Años : 7.00 Nº Egresados: 24	Años : 9.35 Nº Titulados: 20
	2015	Años : 7.00 Nº Egresados: 39	Años : 8.94 Nº Titulados: 33
INGENIERIA DE EJECUCION INFORMATICA	2005	Años : 5.00 Nº Egresados: 54	Años : 7.06 Nº Titulados: 33
	2006	Años : 5.00 Nº Egresados: 41	Años : 7.11 Nº Titulados: 45
	2007	Años : 5.00 Nº Egresados: 35	Años : 8.00 Nº Titulados: 36
	2008	Años : 5.00 Nº Egresados: 47	Años : 7.46 Nº Titulados: 39
	2009	Años : 5.00 Nº Egresados: 42	Años : 8.11 Nº Titulados: 63
	2010	Años : 5.00 Nº Egresados: 41	Años : 8.18 Nº Titulados: 33
	2011	Años : 5.00 Nº Egresados: 15	Años : 7.84 Nº Titulados: 38
	2012	Años : 5.00 Nº Egresados: 79	Años : 8.30 Nº Titulados: 46
	2013	Años : 5.00 Nº Egresados: 22	Años : 6.48 Nº Titulados: 66
	2014	Años : 5.00 Nº Egresados: 70	Años : 6.07 Nº Titulados: 59

Figura A.15: Indicadores tiempo medio titulación vs egreso 2015

B. Implementación de proceso ETL

B.1. Pasos Previos

Antes de comenzar con el proceso ETL, es necesario revisar los datos a transformar obtenidos por parte de la casa central de la PUCV, en un formato de tabla Excel. En el caso del presente proyecto, al principio no todos los datos requeridos fueron obtenidos. Los datos faltantes fueron:

- Código Postal
- Nivel Educacional del Padre.
- Nivel Educacional de la Madre.
- Si el alumno trabaja o no.
- Si el alumno vive solo o no.
- Estado actual de matrícula.

Para completar estos valores, se utilizaron distintas formas de llenado. En el caso del Código Postal, se debió crear una columna en la tabla Excel inicial para luego en el proceso ETL extraer la información desde otra fuente de información. Para el caso del Nivel Educacional del Padre y la Madre, si el alumno trabaja y si vive solo, se le asignaron valores aleatorios, mediante funciones en Excel. Otro valor agregado fue el de obtener un promedio de los campos NEM, Promedio NEM, Promedio Ranking, las pruebas de Lenguaje, Matemáticas, Historia, Ciencias, Promedio PSU y Promedio Ponderado. Todos estos valores fueron agregados en la tabla Excel inicial para luego mediante el proceso ETL llenar valores vacíos y nulos.

B.2. Esquema general de proceso ETL

El proceso ETL utilizado durante este proyecto se divide en dos, el primer proceso, como se indica en la Figura B.1, consiste en la extracción de los datos de la tabla Excel y mediante distintas funciones del programa Spoon de Pentaho y así, transformar los datos y lograr obtener tres tablas Excel para su uso. La primera tabla con toda la información ya transformada, se utilizará como respaldo. La segunda tabla Excel contendrá todos sus valores en formatos numéricos y ya clasificados en rangos(algunos) para el proceso de minería de datos con la herramienta Weka. Por último, una tabla Excel con todos los valores transformados y sin clasificación de rangos, para su posterior carga al repositorio de datos.

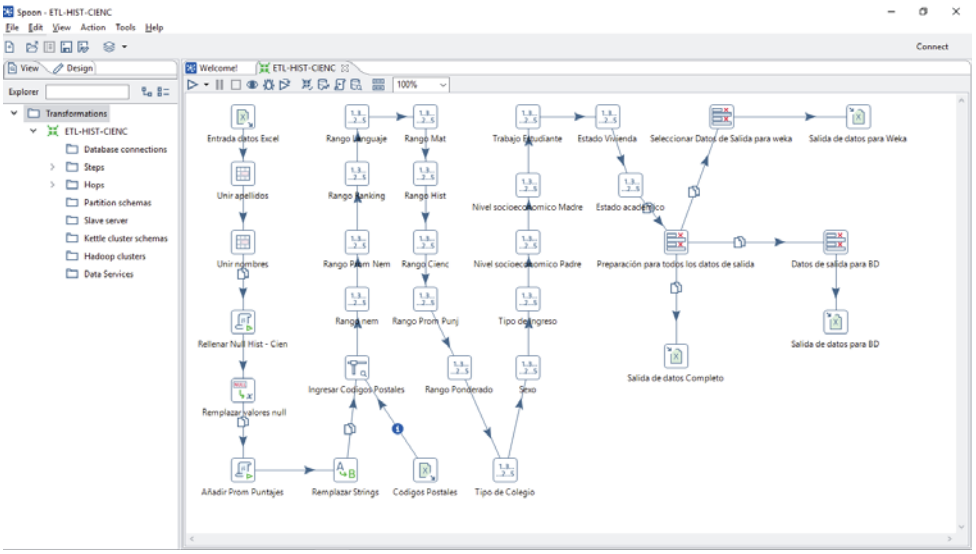


Figura B.1: Esquema general de los pasos del proceso de Extracción y Transformación en Spoon

El segundo proceso, como se puede apreciar en la Figura B.2, consiste en la carga de los datos al datawarehouse a utilizar en el esquema presentado en la Figura 12 en el Anexo A. Se comienza por llenar las tablas de dimensión para luego finalizar con la tabla de hecho Estudiante.

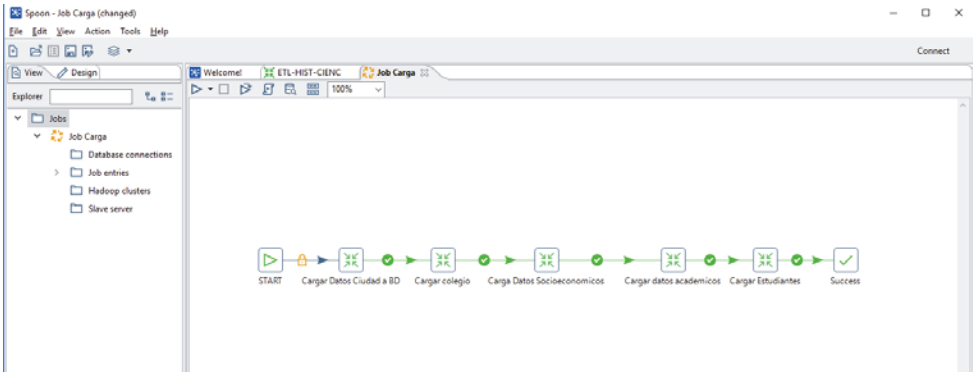


Figura B.2: Esquema general de los pasos del proceso carga en Spoon

B.3. Proceso de Extracción y Transformación

A continuación, se detallan los distintos pasos utilizados durante el proceso de extracción y transformación de los datos.

B.3.1. Unir apellidos y nombres

En este paso, gracias a la función Concant Fields, se toman las celdas NOMBRE1 Y NOMBRE2, para unirlas y generar la celda Nombres con el primer y segundo nombre. Lo mismo

sucede en el caso de los apellidos, con se indica en la Figura B.3, donde se unen las celdas AP PATERNO y AP MATERNO para formar la celda Apellidos.

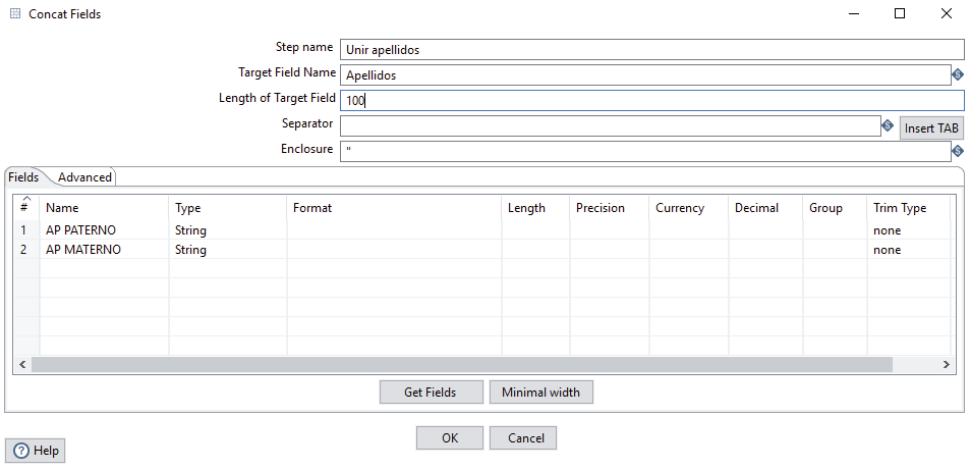


Figura B.3: Función Concat Fields, para unir los apellidos del estudiante

B.3.2. Rellenar valores NULL para puntajes Historia y Ciencias

Algunos estudiantes solo rindieron una prueba específica, Historia o Ciencias. El archivo de datos entregados por la casa central entrega valores NULL, para las pruebas no rendidas por los estudiantes. Es por esta razón, que se ha de agregar una función que permite agregar un JavaScript, llamada Script Value. Con ésta se remplaza el valor NULL con el puntaje de la prueba rendida por el estudiante, como se indica en la figura B.4.

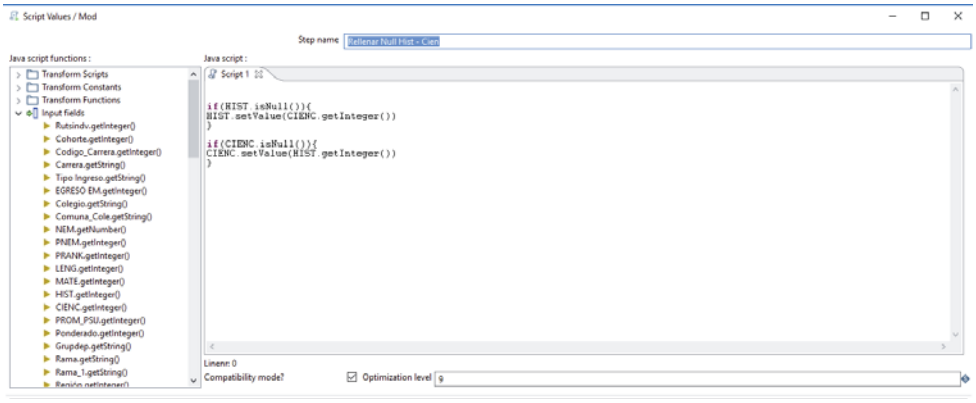


Figura B.4: Función Script Value utilizada para rellanar los valores NULL de las pruebas de Historia y Ciencias

B.3.3. Rellenar valores NULL

Otros valores NULL se encontraron dentro de los datos entregados por la casa central, éstos se encontraban en los campos de Preferencia, Año Puntaje, Nombre del Colegio, Comuna Colegio, Región, Categorización numérica del tipo de establecimiento, el tipo de rama del colegio, y el año de egreso de enseñanza media. Como indica la Figura B.5, gracias a la función Replace null Value, los valores NULL de dichos campos fueron cambiados por los valores que más se repiten dentro de los datos, para así no afectar el posterior análisis predictivo. El único valor, que no se rellenó con dicha representación, y se le otorgó un valor "Desconocido"fue el nombre del colegio ya que este no es un valor que se utilizará dentro del paso de Minería de datos.

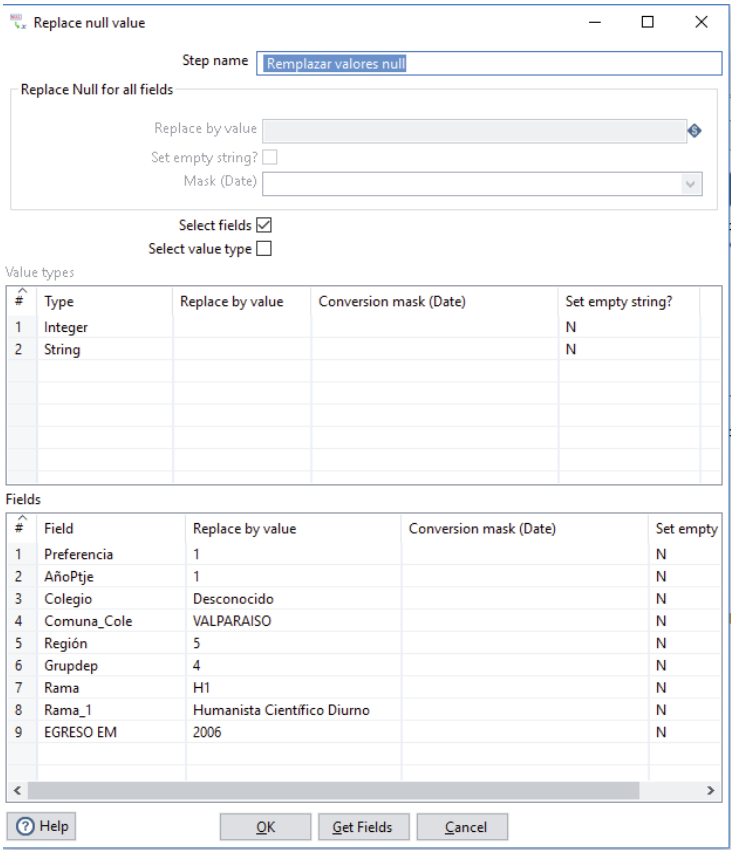


Figura B.5: Función Replace null Value utilizada

B.3.4. Añadir promedio a puntajes NULL

Como se indicó en la Sección B.1, Pasos Previos, se obtuvieron los promedios de los distintos puntajes logrados por los alumnos. La razón de esto es rellenar los valores NULL de estudiantes que no tenían dicha información en la información entregada por casa central y no sesgar los datos para su futuro análisis. Para esto se utiliza un Script Value, como se indica en la Figura B.6.

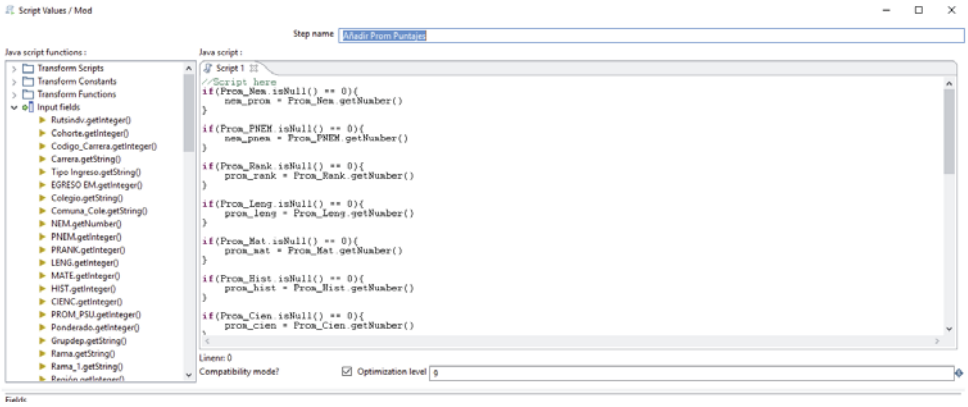
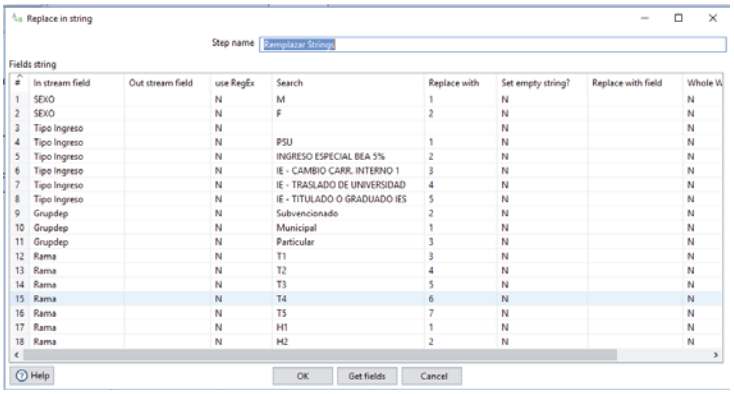


Figura B.6: Función Script Value y el código para añadir el promedio a los valores NULL

B.3.5. Remplazar Strings

La función Replace in String en Spoon, permite buscar dentro de los campos, valores indicados por el usuario y remplazarlos por otros valores, también, indicados por el usuario. Como se indica en la Figura B.7, se ha remplazar distintos valores String en valores numéricos para utilizarlos en las futuras pruebas de análisis de Minería de Datos.



#	In stream field	Out stream field	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole W
1	SEXO		N	M	1	N		N
2	SEXO		N	F	2	N		N
3	Tipo Ingreso		N			N		N
4	Tipo Ingreso		N	PSU	1	N		N
5	Tipo Ingreso		N	INGRESO ESPECIAL BEA 5%	2	N		N
6	Tipo Ingreso		N	IE - CAMBIO CARR. INTERNO 1	3	N		N
7	Tipo Ingreso		N	IE - TRASLADO DE UNIVERSIDAD	4	N		N
8	Tipo Ingreso		N	IE - TITULADO O GRADUADO IES	5	N		N
9	Grupdep		N	Subvencionado	2	N		N
10	Grupdep		N	Municipal	1	N		N
11	Grupdep		N	Particular	3	N		N
12	Rama		N	T1	3	N		N
13	Rama		N	T2	4	N		N
14	Rama		N	T3	5	N		N
15	Rama		N	T4	6	N		N
16	Rama		N	T5	7	N		N
17	Rama		N	H1	1	N		N
18	Rama		N	H2	2	N		N

Figura B.7: Función Replace in String utilizada para remplazar a valores numéricos

B.3.6. Ingresar códigos postales

Para este paso, es necesario obtener un archivo con los distintos códigos postales de las ciudades del país, en el caso del presente proyecto se ha de realizar uno en formato Excel.
Como se indica en la Figura B.8, con la función Excel Input se ha agregar la tabla Excel que contiene el nombre de la ciudad y su código postal a la función Stream Value Lookup, llamada Ingresar Códigos Postales en el esquema, la cual como indica la Figura B.9, compara los nombres de las ciudades y en caso de que coincidan agrega un nuevo campo a la tabla llamado Código Postal.

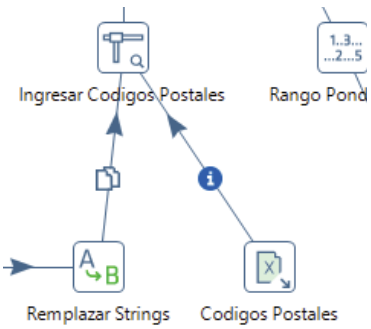


Figura B.8: Función Excel Input con los códigos postales de las ciudades y la función Stream Value Lookup, para comprar e ingresar los códigos postales a la tabla principal

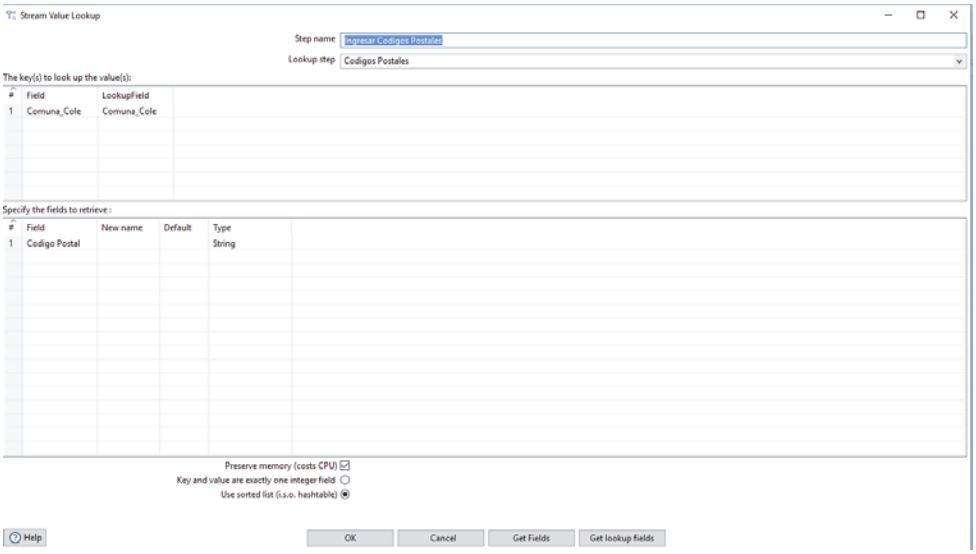


Figura B.9: Función Stream Value Lookup y los valores a comparar para ingresar el Código Postal a la tabla

B.3.7. Agregar Rangos a Puntajes

Para agregar rangos a distintos valores los campos de datos, es necesario utilizar la función Number Ranges en Spoon. Ingresando el valor mínimo y máximo (que no se toma en cuenta dentro del rango a evaluar) se puede agregar una nueva columna a la tabla con dicha clasificación. En la Figura B.10, se puede apreciar los rangos de valores del campo NEM y su clasificación numérica, para su posterior utilización durante el proceso de Minería de Datos.

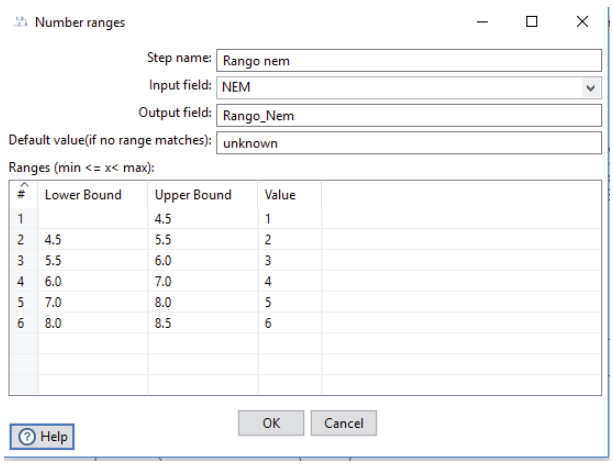


Figura B.10: Función Number Ranges para clasificar los valores de la columna NEM

Para el caso de los campos Promedio Nem, Promedio Ranking, Puntaje Lenguaje, Puntaje Matemáticas, Puntaje Historia, Puntaje Ciencias, Promedio Puntaje y Promedio Ponderado se ha de utilizar la misma asignación de rango, como se puede observar en la Figura B.11, para clasificar el campo Promedio Nem.

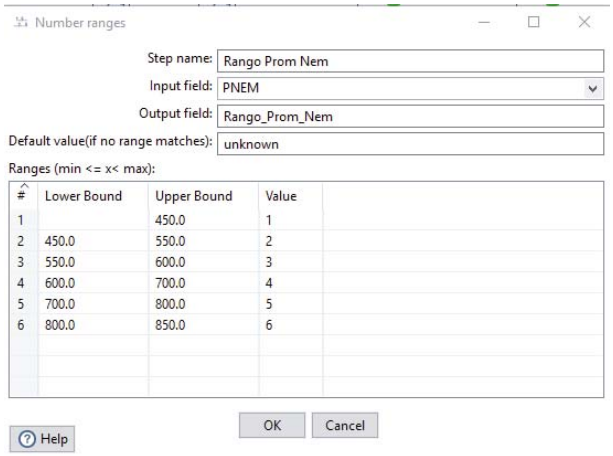


Figura B.11: Función Number Ranges para clasificar los valores de la columna PromNem

B.3.8. Agregar valores String

Para una correcta visualización de la información luego de cargarlos al Datawarehouse, es necesario identificar con texto los datos numéricos(que son utilizados para el proceso de Minería de Datos). Para esto se ha de utilizar nuevamente la función Number Ranges, para clasificar, por su valor mínimo, el dato que se quiere representar y agregarlo a la tabla de datos. A continuación, se presentan todos los datos agregados.

B.3.8.1. Clasificación de Tipo de Colegio

Para la clasificación de los valores del campo Tipo de Colegio, se ha de agregar cuatro valores, como se indica en la Figura B.12.

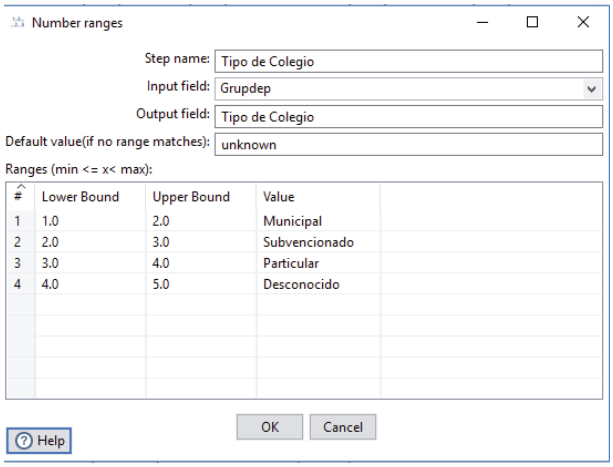


Figura B.12: Función Number Ranges para clasificar los valores de la columna Tipo de Colegio

B.3.8.2. Clasificación Sexo

Para la clasificación del Sexo de los estudiantes se ha clasificar como se indica en la Figura B.13; 1 es Masculino con la clasificación M y 2 es Femenino con la clasificación F.

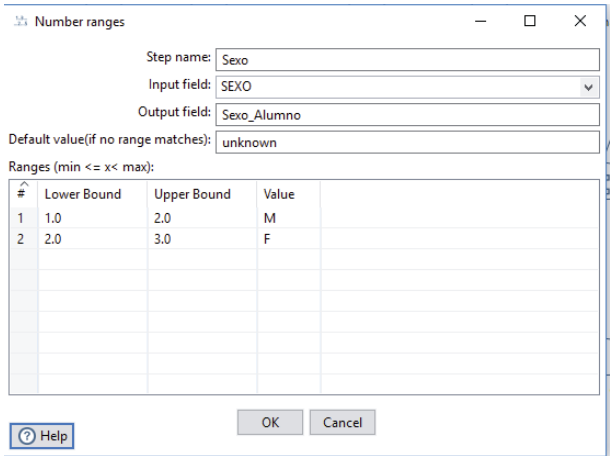


Figura B.13: Función Number Ranges para clasificar los valores de la columna Sexo

B.3.8.3. Clasificación Tipo de Ingreso

Para la clasificación del tipo de ingreso de los estudiantes a las carreras de Ingeniería Civil Informática e Ingenierá en Ejecución Informática , se ha dispuesto de cinco valores, los que se

pueden observar en la Figura B.14.

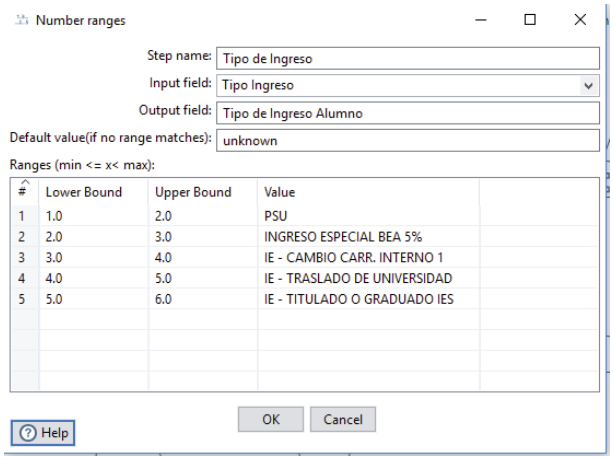


Figura B.14: Función Number Ranges para clasificar los valores de la columna Tipo de Ingreso

B.3.8.4. Nivel Educacional de los padres

La clasificación del nivel educacional de los padres de los estudiantes se ha establecido en ocho valores. Éstos se pueden observar en la Figura B.15, que muestra la clasificación del nivel educacional del padre. Para la madre del estudiante se ha de utilizar la misma asignación.

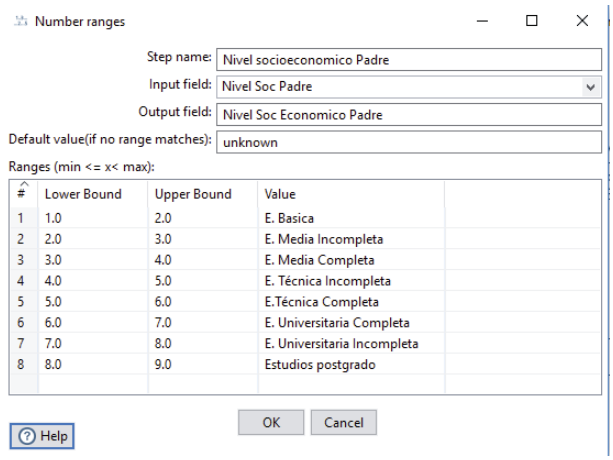


Figura B.15: Función Number Ranges para clasificar los valores de la columna Nivel Educacional del padre

B.3.8.5. Trabajo del Estudiante

Se ha de definir con dos valores, SI o No, si el estudiante tiene un trabajo fuera del horario de estudios. La clasificación se puede observar en la Figura B.16

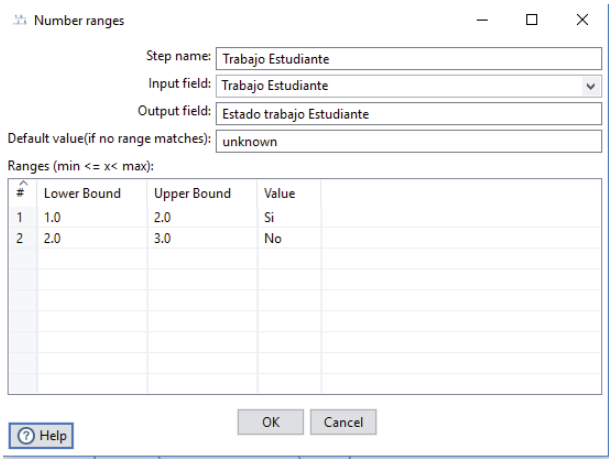


Figura B.16: Función Number Ranges para clasificar los valores de la columna Trabajo Estudiante

B.3.8.6. Estado de convivencia del Estudiante

Para clasificar si el alumno vive solo o con alguien más cercano, se ha clasificar en simples dos variables, "Solo" y "Acompañado", la Figura B.18 muestra la función Number Ranges que ha de hacer la clasificación.

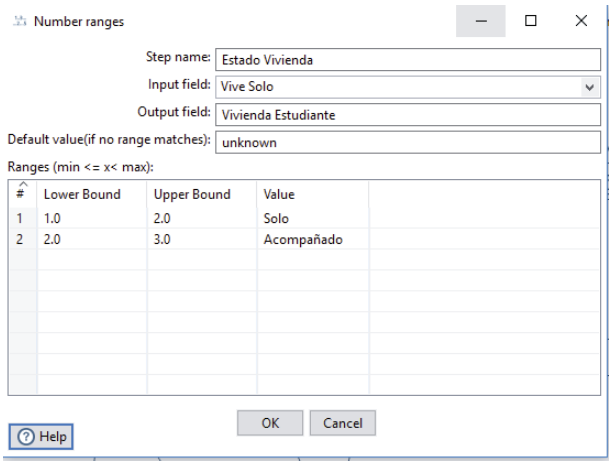


Figura B.17: Función Number Ranges para clasificar con quien vive el estudiante

B.3.8.7. Estado académico del estudiante

Por último, la clasificación más importante es para saber si el estudiante se encuentra cursando la carrera o ha desertado, esta se puede observar en la Figura B.18.

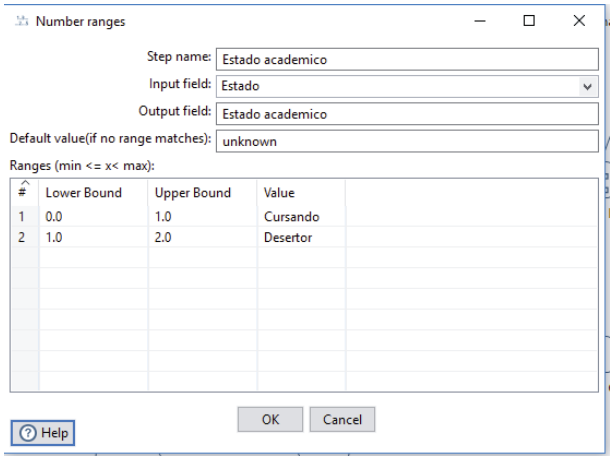


Figura B.18: Función Number Ranges para clasificar el Estado Académico del estudiante

B.3.9. Selección final de valores para las tablas de salida

Como último paso de la transformación se seleccionarán y ordenarán los datos necesarios para obtener un respaldo de los datos transformados, su uso en la herramienta Weka y su carga en el Datawarehouse para su posterior visualización y utilización en cubos ROLAP, análisis y reportes. Para seleccionar, modificar y eliminar valores de la tabla es necesario utilizar la función de Spoon Select/Rename values, en esta se puede ordenar las columnas, re-nombrarlas, eliminar columnas y modificar su formato.

B.3.9.1. Selección de datos para el respaldo de los datos

Para la selección de datos para el respaldo, se ha de ordenar todos los datos obtenidos, como se indica en la Figura B.19; tanto los valores de rangos numéricos y los valores específicos en String, de forma que se visualice de forma ordenada al ser exportado en la tabla Excel.

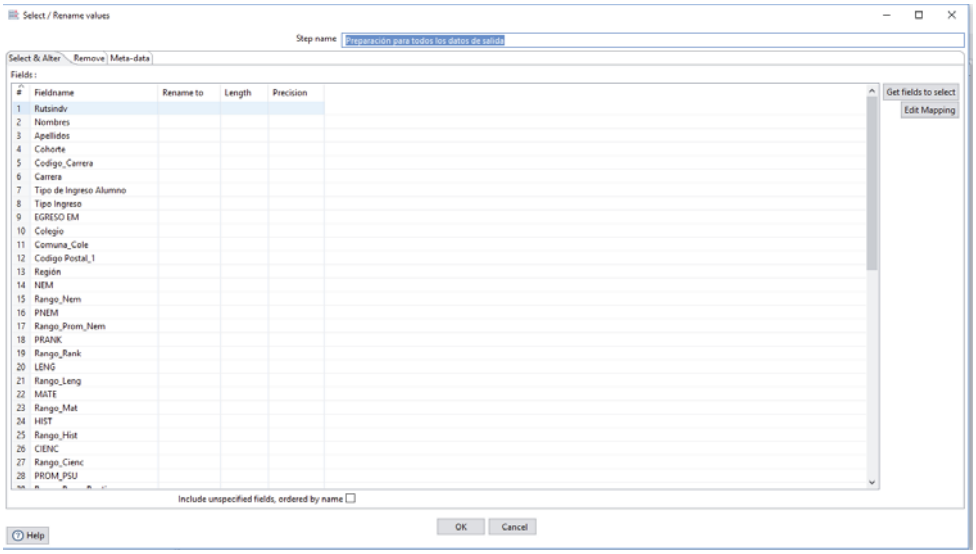


Figura B.19: Función Select/Rename values, en su opción Select & Alter, para ordenar los datos de salida

Además de ordenar, se debe de borrar las columnas utilizadas al comienzo para obtener los promedios de los puntajes y así llenar los valores NULL de estos, y la columna Código Postal sobrante del proceso de obtención de estos. La misma función Select/Rename values, puede remover estos valores como se indica en la Figura B.20.

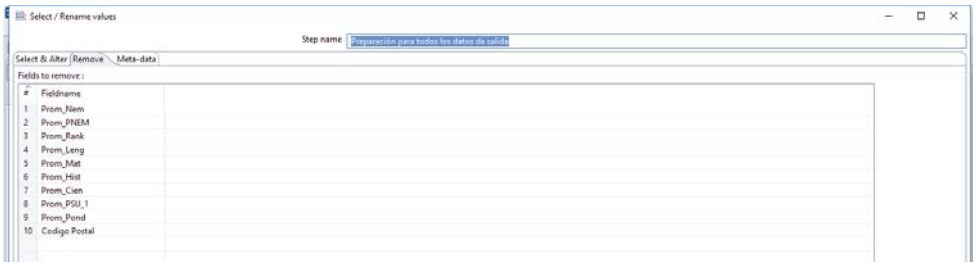


Figura B.20: Función Select/Rename values, en su opción Remove, para eliminar datos no necesarios

Por último, se realizan modificaciones a nombres y formatos de salida de algunos datos, como se indica en la Figura B.21.

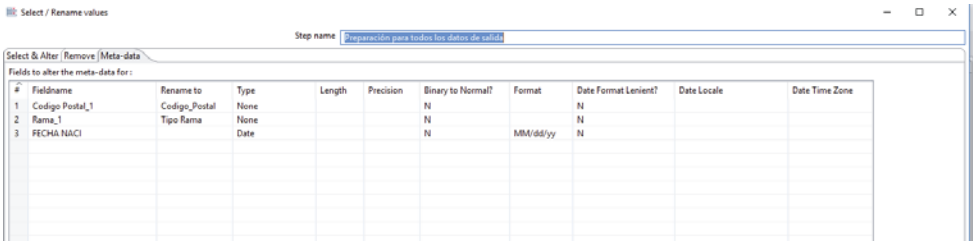


Figura B.21: Función Select/Rename values, en su opción Meta-data, para modificar el formato de los datos

B.3.9.2. Selección de datos para el análisis de Minería de Datos

Para la selección de datos a utilizar en el proceso de Minería de Datos, se ha de utilizar todos las representaciones numéricas de los datos a utilizar, estos son Rut, Cohorte, el Código de la carrera, el año de egreso de enseñanza media, el tipo de ingreso del estudiante, el código postal de la ciudad donde estudio, la región, el rango al cual pertenece el NEM que obtuvo, el rango del promedio del NEM, ranking, promedio del puntaje y puntaje ponderado y los rangos de los puntajes de las pruebas de lenguaje, matemáticas, historia y ciencias. Además, el tipo de establecimiento donde estudio, la rama educativa de su establecimiento, edad, preferencia de la postulación, año utilizado del puntaje PSU, los niveles educativos de los padres, si vive solo o acompañado, si trabaja en tiempos fuera de la universidad y por último la variable clasificadora de estado.

Todos los valores no numéricos, o no clasificados en rango, son eliminados como se puede apreciar en la Figura B.22.

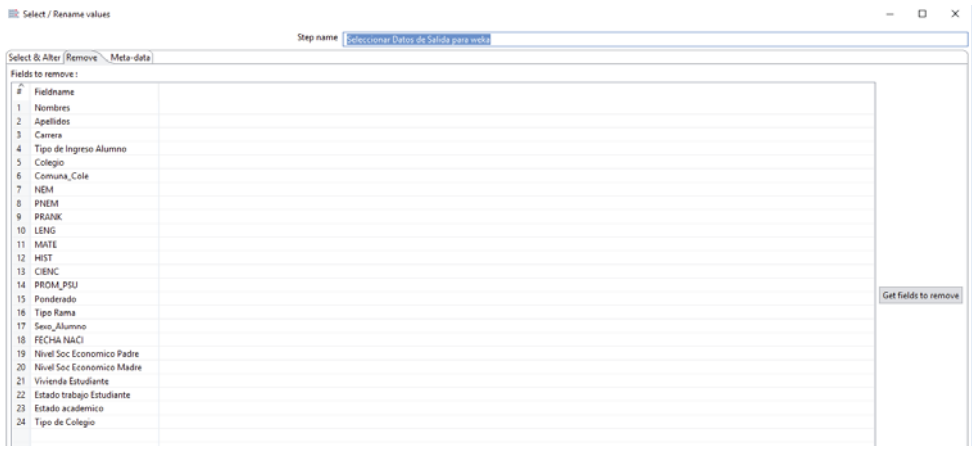


Figura B.22: Función Select/Rename values, en su opción Remove, para eliminar datos no necesarios para el proceso de Minería de Datos

Por último, se ha de cambiar el nombre de distintas variables para su mejor comprensión en los pasos futuros como se indica en la Figura B.23.

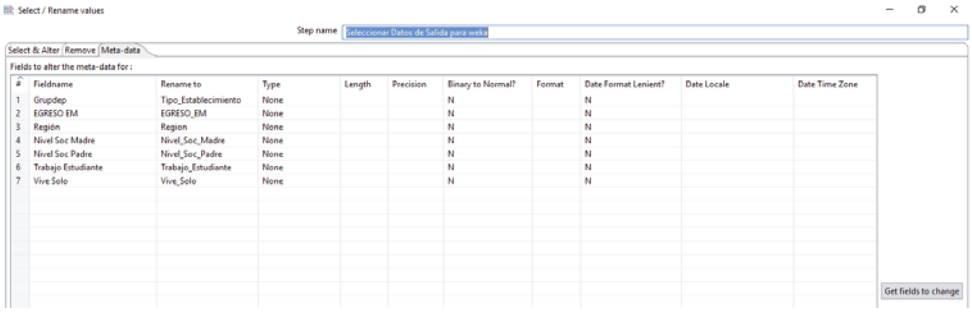


Figura B.23: Función Select/Rename values, en su opción Meta-data, para modificar el nombre de algunos valores para la métrica del proceso de Minería de Datos

B.3.9.3. Selección de datos para la carga en la BD

Al contrario de los pasos utilizados en la selección de datos para el proceso de Minería de Datos B.3.9.2, donde se utilizan los datos de carácter numérico, para el proceso de carga a la base de datos, es necesario utilizar los valores textuales de los datos. Es por esta razón, que son eliminados todos los valores numéricos para describir los datos y clasificarlos en rangos, como se puede apreciar en la Figura B.24.

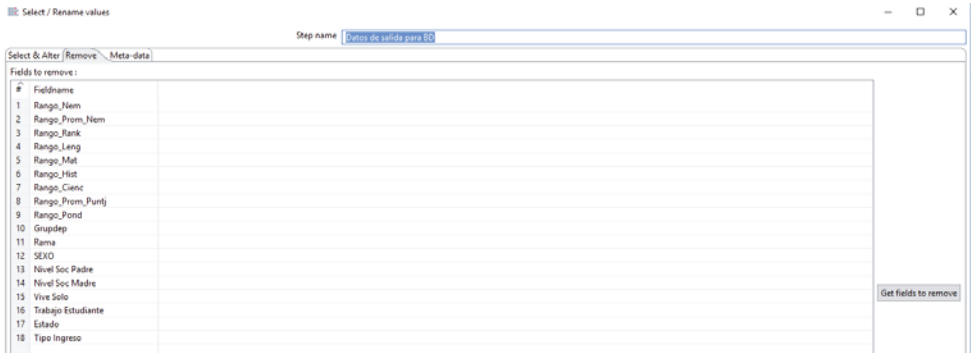


Figura B.24: Función Select/Rename values, en su opción Remove, para eliminar datos no necesarios para la carga en la Base de Datos

B.4. Proceso de Carga a la Base de Datos

B.4.1. Sistema de Gestión de Base de Datos utilizado

Para el presente proyecto se ha de utilizar PostgreSQL 9.3 como sistema de gestión de datos relacional y PgAdmin 4 como entorno visual para conectarse a la base de datos de PostgreSQL, se puede visualizar en la Figura B.25. Ésta facilita la gestión y administración de ésta mediante instrucciones SQL o ayuda de un entorno gráfico para consultar, manipular y gestionar datos.

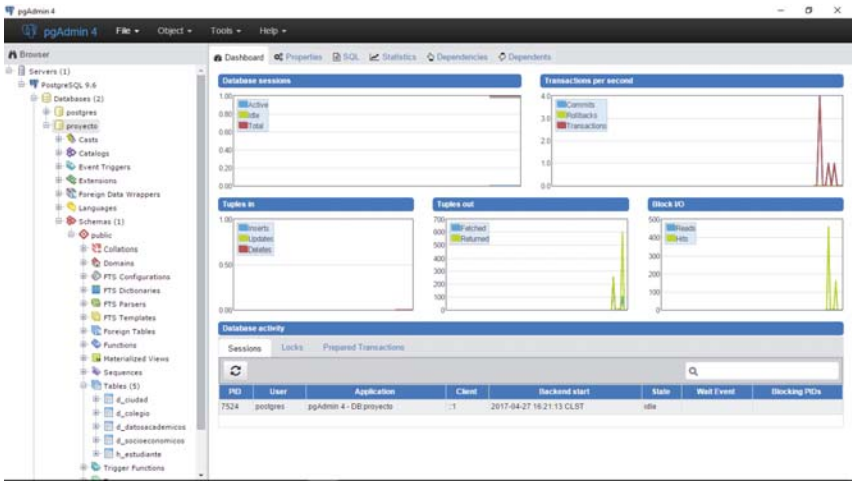


Figura B.25: Entorno gráfico PgAdmin en el cual se ha trabajado para gestionar el Datawarehouse

PgAdmin también ayuda a hacer más rápida la creación de las tablas, es por esta razón que la tabla de hechos estudiante, como se ejemplifica en la Figura B.26, y las tablas de dimensiones ciudad, datos socio económicos, datos académicos y el colegio fueron creados mediante este entorno visual, además de crear las Claves Foráneas correspondientes como se puede observar en la Figura B.27

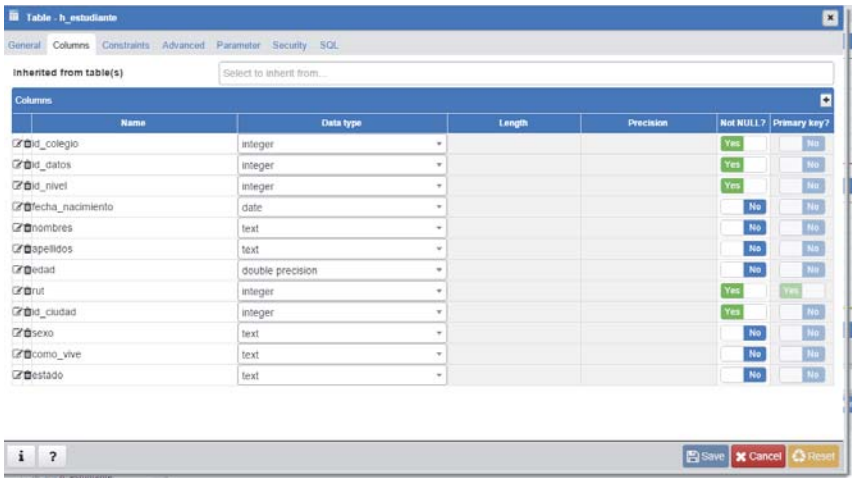


Figura B.26: Creación de tabla de hechos h_estudiante

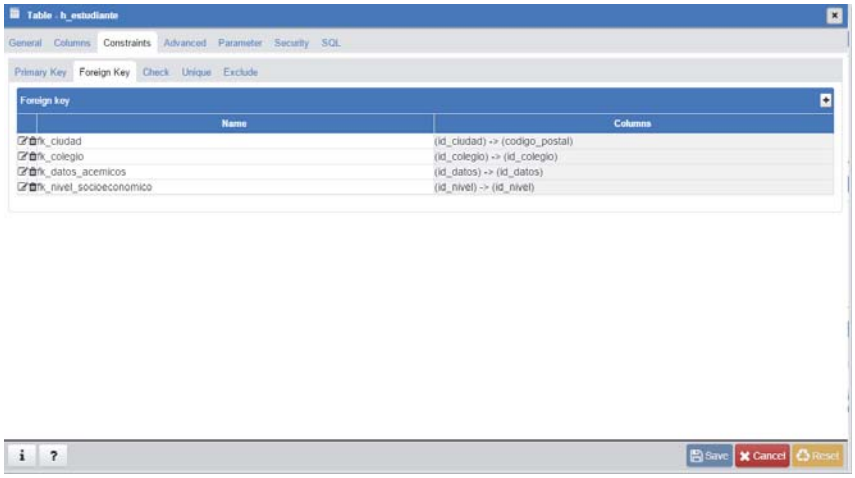


Figura B.27: Creación de Claves foráneas de la tabla h_estudiante

B.4.2. Conexión de la Base de Datos a la plataforma Spoon de Pentaho

Para conectar la base de datos PostgreSQL a Spoon, basta con agregar una nueva conexión dentro de las opciones de la aplicación. Luego, como se indica en la Figura B.28, se ha de buscar el tipo de conexión (PostgreSQL), llenar los campos de Host, el nombre de la base de datos a conectar, el número del puerto, el usuario y la contraseña de PostgreSQL. Para verificar si la conexión se realiza con éxito se ha de seleccionar la opción de Test, al comprobar que la conexión se realizó con éxito se confirma la conexión.

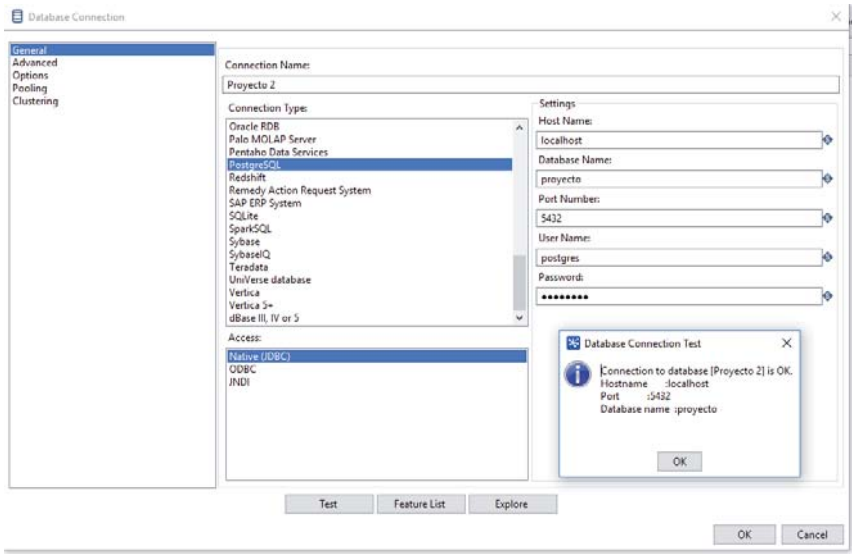


Figura B.28: Conexión de a la base de datos PostgreSQL con Spoon de Pentaho

B.4.3. Transformaciones para realizar la carga de Datos a la base de datos

B.4.3.1. Cargar datos a tabla d_ciudad

El esquema para realizar la carga de los datos a la tabla d_ciudad se indica en la Figura B.29, en la cual se utilizan cuatro funciones de la herramienta Spoon; Select/Rename Values, indicado en la Figura B.30, para eliminar los datos no utilizados para llenar la tabla d_ciudad, quedando los campos nombre_ciudad, codigo_postal y región.

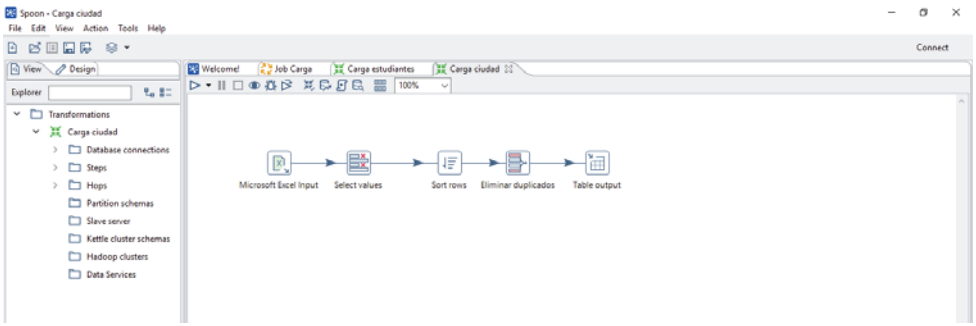


Figura B.29: Esquema de la transformación de carga de la tabla d_ciudad

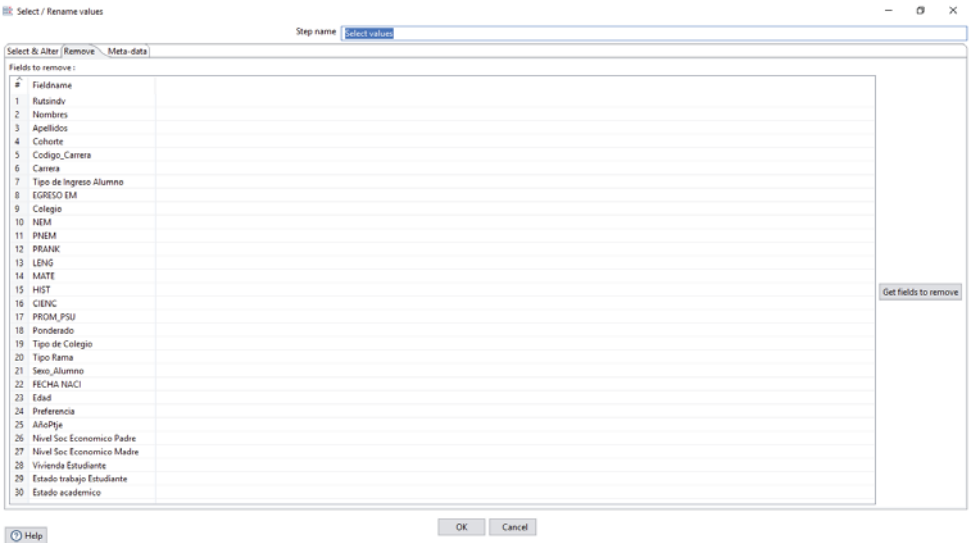


Figura B.30: Función Select/Rename Values para eliminar los datos no utilizados para la tabla d_ciudad

Luego se utiliza la función Sort Row, para ordenar de forma ascendente o descendente los datos seleccionados. Como se indica en la Figura B.31, se ordena de forma ascendente el campo codigo_postal, esto es un paso necesario para lograr utilizar de forma correcta la función Unique Rows, para encontrar valores duplicados y eliminarlos. En la Figura B.32 se indica como los campos nombre_comuna y codigo_postal son revisados y se ignora el campo región.

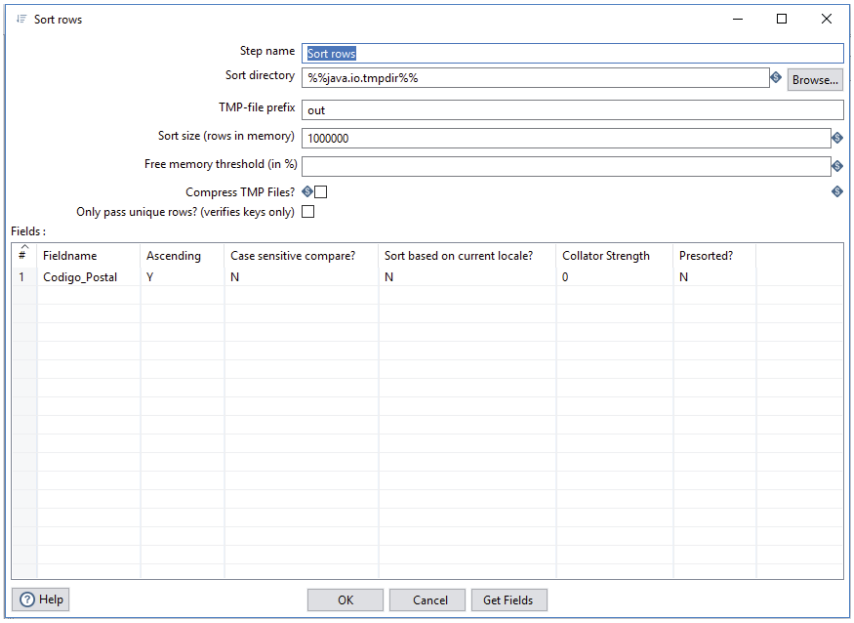


Figura B.31: Función Sort Row para ordenar el campo codigo_postal

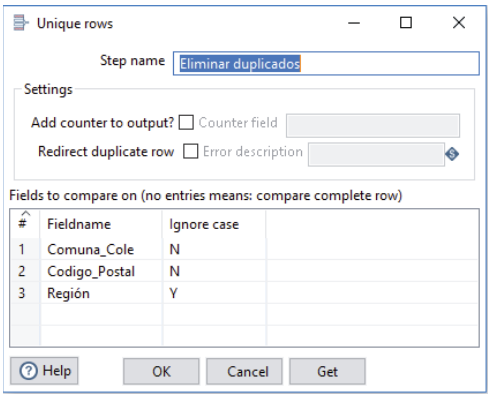


Figura B.32: Función Unique Rows para eliminar los valores duplicados para su posterior carga a la tabla d_ciudad

Para insertar los datos a la tabla se utiliza la función Table Input. En ésta, como indica la Figura B.33, se debe de especificar la conexión a la base de datos, en este caso “Proyecto 2” y la tabla objetivo “‘d_ciudad”.

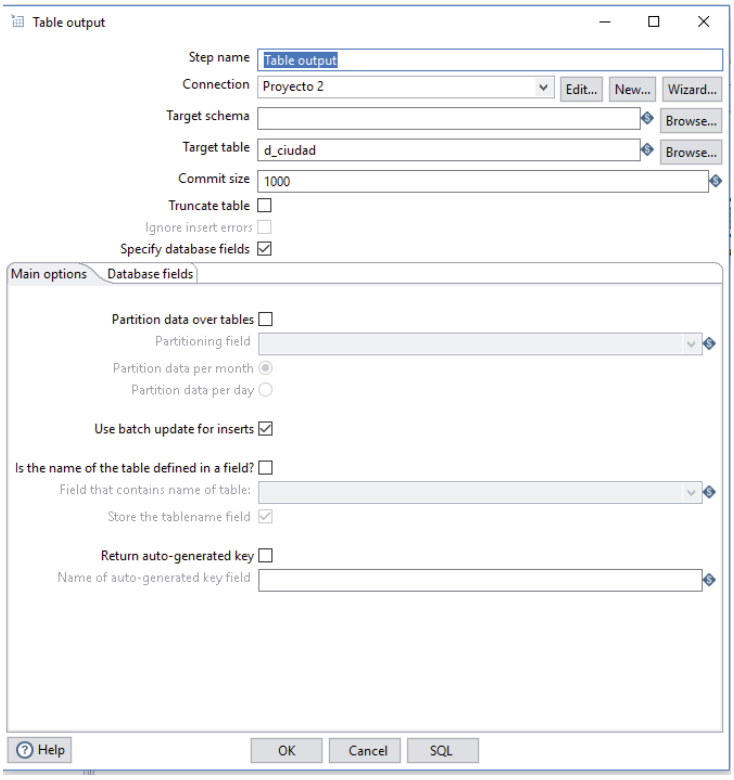


Figura B.33: Función Table Output, para la carga de datos a la tabla d_ciudad

En la Figura B.34, se indican los datos para insertar a la tabla d_ciudad. Se procede a apretar la opción SQL, para generar el código SQL, como se indica en la Figura B.35 y luego se selecciona la opción Excute(una sola vez).

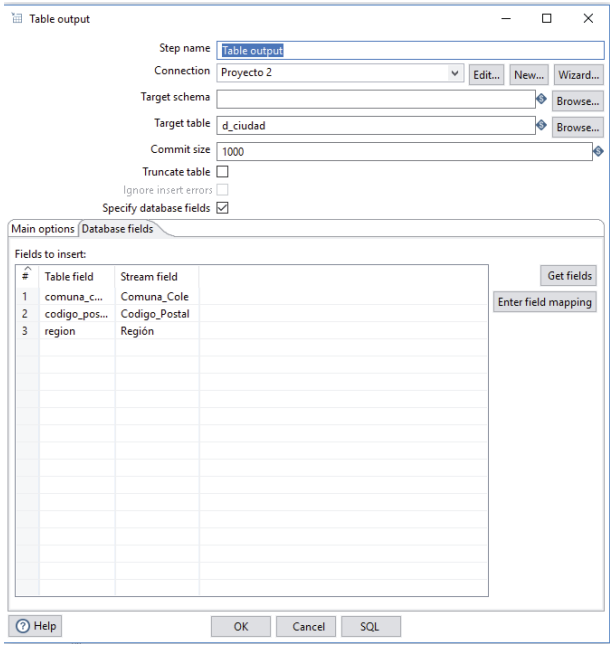


Figura B.34: Función Table Output, con los valores seleccionados para la tabla d_ciudad

Figura B.35: Script SQL generado para cargar la tabla d_ciudad

B.4.3.2. Cargar datos a tabla d_colegio

Para cargar los datos, a la tabla d_colegio, como se indica en la Figura B.36, se utilizaron las mismas funciones de Spoon para cargar la tabla d_ciudad B.4.3.1. En la Figura B.37 se especifican los valores eliminados, gracias a la función Select/Rename Values, que no se insertaran en la tabla d_colegio. La Figura B.38 muestra la función Sort Row, para ordenar de forma ascendente el campo Colegio, para posteriormente utilizar la función Unique Row, Figura B.39,

para eliminar los valores duplicados. Por último, la función Table output, Figuras B.40 y B.41, inserta los valores a la tabla d_colegio, especificando la conexión a la base de datos, la tabla objetivo, seleccionando los datos y marcando la opción de auto generar la id_colegio, de valor Serial.

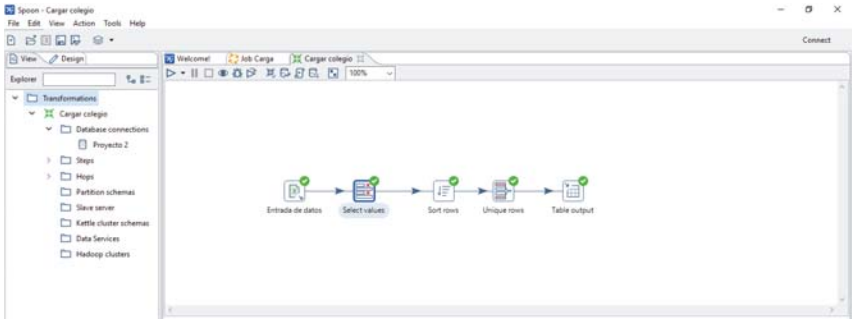


Figura B.36: Esquema de la transformación de carga de datos a la tabla d_colegio

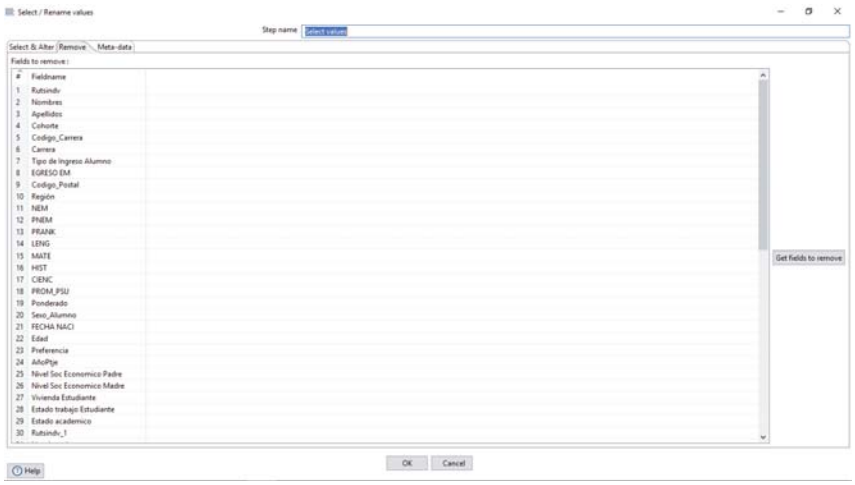


Figura B.37: Función Select/Rename Values para eliminar los datos no utilizados para la tabla d_colegio

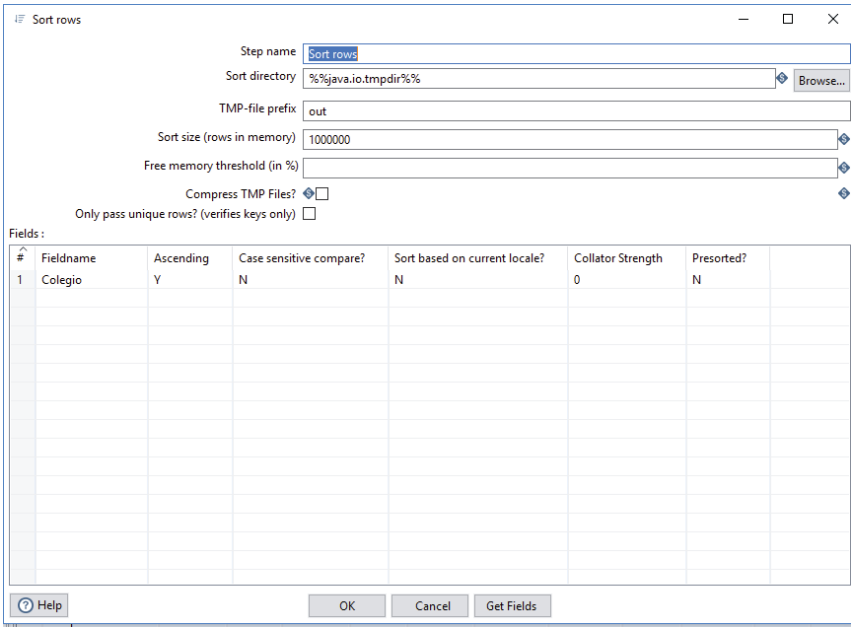


Figura B.38: Función Sort Row para ordenar el campo Colegio

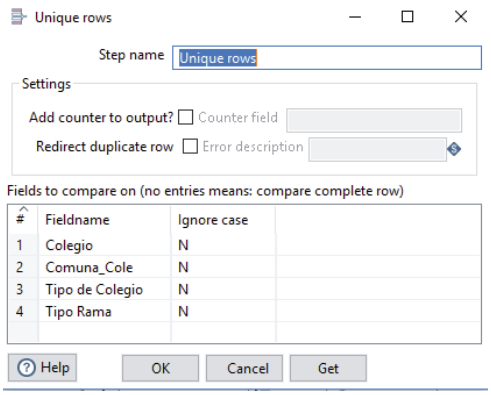


Figura B.39: Función Unique Rows para eliminar los valores duplicados para su posterior carga a la tabla d_colegio

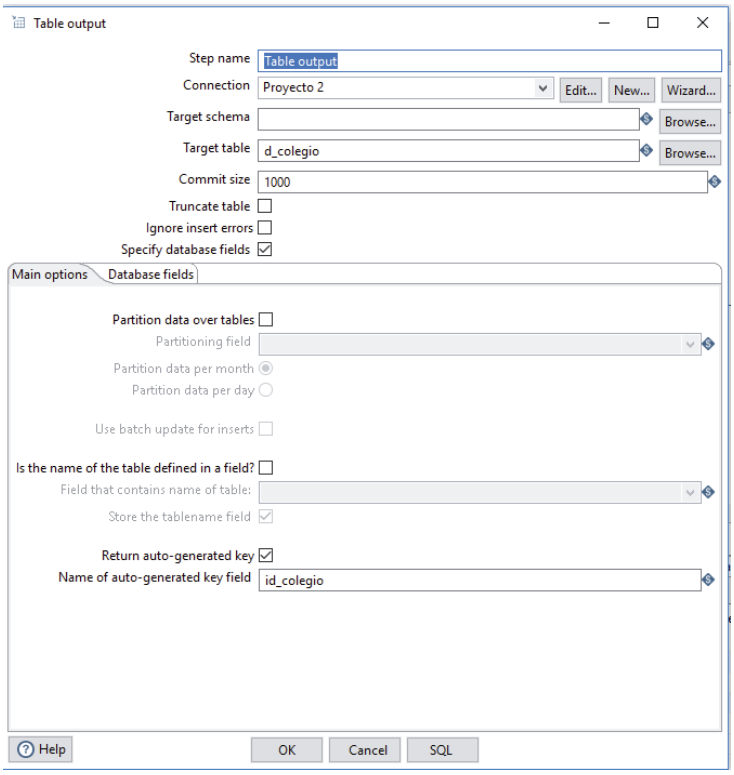


Figura B.40: Función Table Output, para la carga de datos a la tabla d_colegio

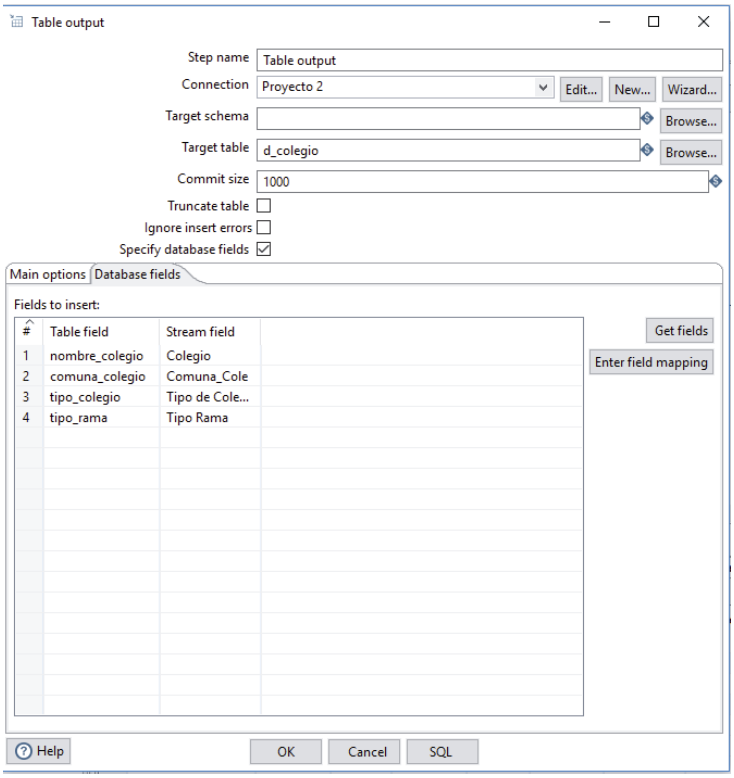


Figura B.41: Función Table Output, con los valores seleccionados para la tabla d_colegio

B.4.3.3. Cargar datos a tabla d_datosocioeconomicos

El esquema para la transformación de carga de datos a la tabla d_datosocioeconomicos está representada por la Figura B.42.

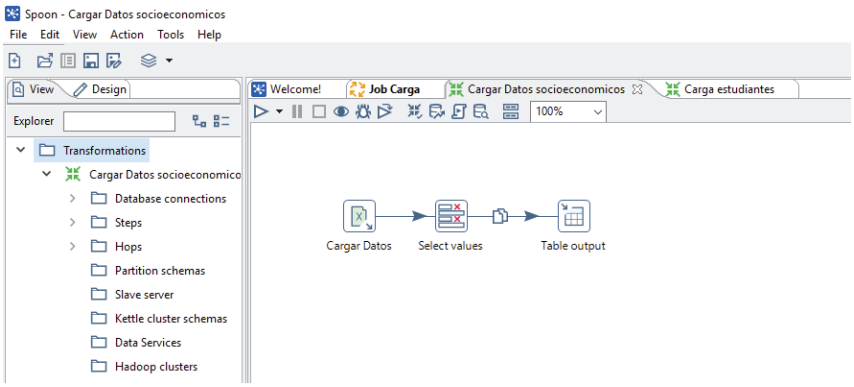


Figura B.42: Esquema de la transformación de carga de datos a la tabla d_datosocioeconomicos

La primera función es Select/Rename Values, para eliminar los valores que no se utilizarán, como se puede apreciar en la Figura B.43. Por último, las Figuras B.44 y B.45 indican los valores

a insertar en la tabla d_datosocioeconomicos, especificando la conexión a la base de datos, la tabla objetivo, seleccionando los datos y marcando la opción de auto generar la id_nivel, de valor Serial.

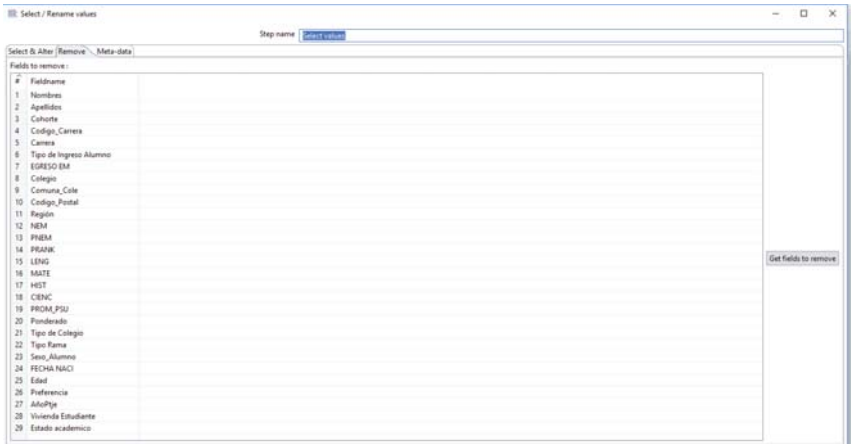


Figura B.43: Función Select/Rename Values para eliminar los datos no utilizados para la tabla d_datosocioeconomicos

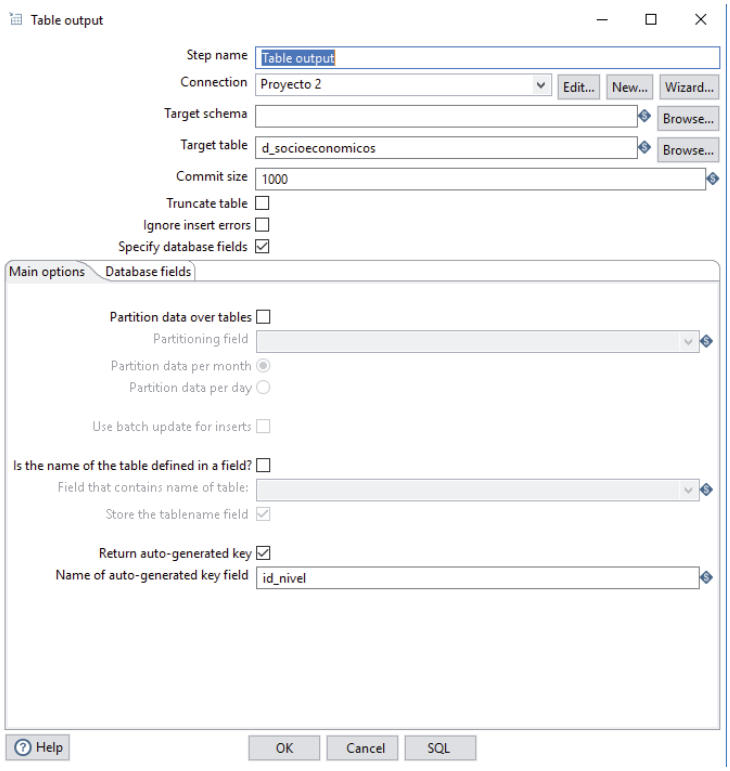


Figura B.44: Función Table Output, para la carga de datos a la tabla d_datosocioeconomicos

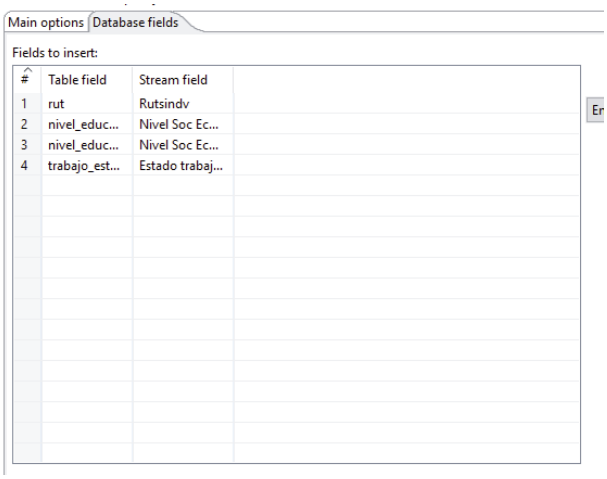


Figura B.45: Función Table Output, con los valores seleccionados para la tabla d_datosocioeconomicos

B.4.3.4. Cargar datos a tabla d_datosacademicos

El esquema para la transformación de carga de datos a la tabla d_datosacademicos se puede apreciar en la Figura B.46.

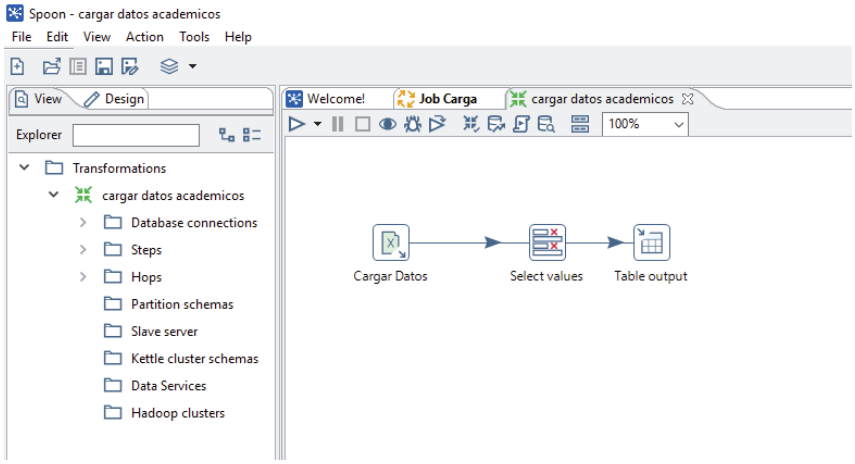


Figura B.46: Esquema de la transformación de carga de datos a la tabla d_datosocioeconomicos

La función es Select/Rename Values, elimina los datos que no se utilizaran, Figura B.47, por otro lado, las Figuras B.48 y B.49 indican los valores a insertar en la tabla d_datosacademicos, indicando la conexión a la base de datos, la tabla objetivo, seleccionando los datos y marcando la opción de auto generar la id_datos, de valor Serial.

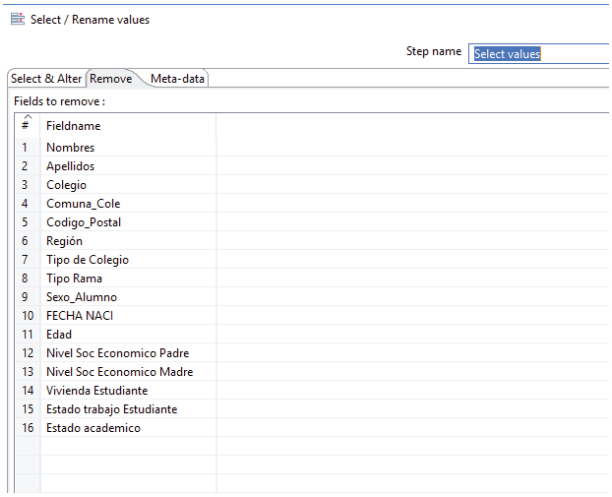


Figura B.47: Función Select/Rename Values para eliminar los datos no utilizados para la tabla d_datosacademicos

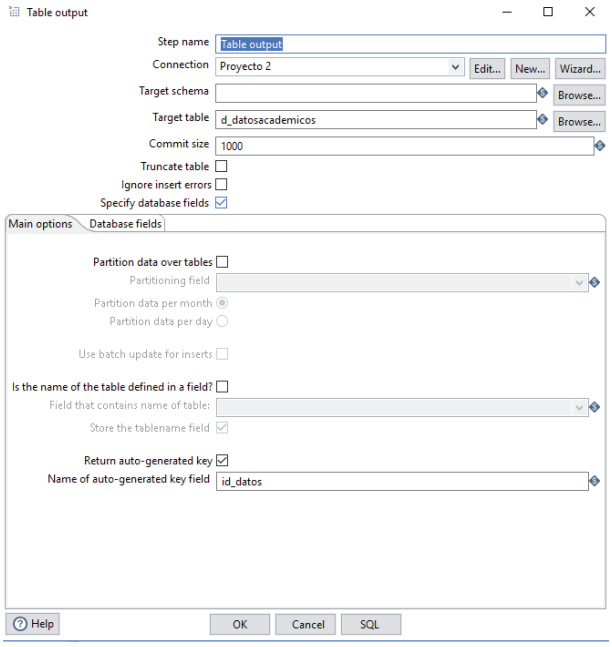
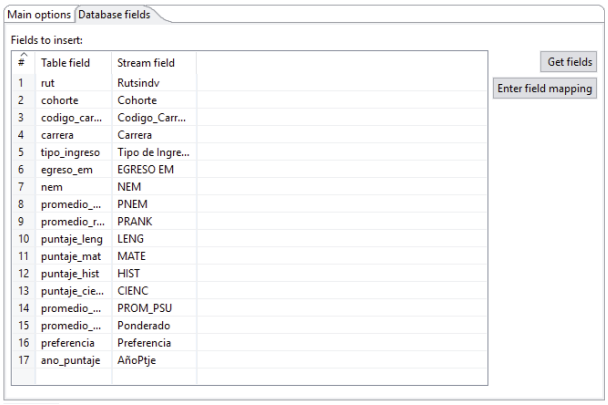


Figura B.48: Función Table Output, para la carga de datos a la tabla d_datosacademicos



#	Table field	Stream field
1	rut	Rutsindv
2	cohorte	Cohorte
3	codigo_car...	Codigo_Carr...
4	carrera	Carrera
5	tipo_ingreso	Tipo de Ingre...
6	egreso_em	EGRESO EM
7	nem	NEM
8	promedio_...	PNEM
9	promedio_r...	PRANK
10	puntaje_leng	LENG
11	puntaje_mat	MATE
12	puntaje_hist	HIST
13	puntaje_cie...	CIENC
14	promedio_...	PROM_PSU
15	promedio_...	Ponderado
16	preferencia	Preferencia
17	ano_puntaje	AñoPtje

Figura B.49: Función Table Output, con los valores seleccionados para la tabla d_datosacademicos

B.4.3.5. Cargar datos a tabla h_estudiante

El esquema representado en la Figura B.50, muestra el proceso de transformación de carga de datos a la tabla h_estudiante, esta contiene muchos pasos vistos anteriormente y necesita de las tablas anteriores previamente cargadas para poder obtener las claves foráneas correspondientes.

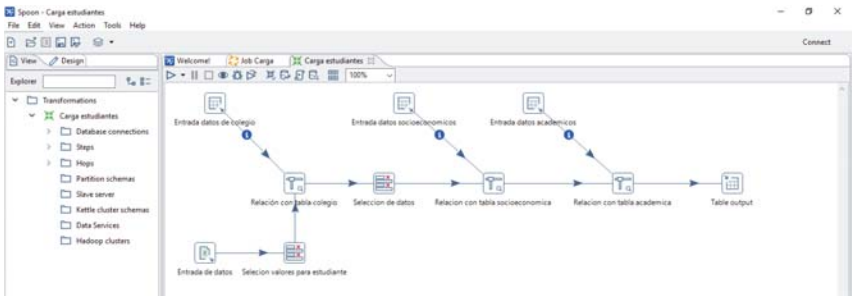


Figura B.50: Esquema de la transformación de carga de datos a la tabla d_datosocioeconomicos

Con la función Relación con tabla colegio (Stream Value),como se aprecia en la Figura B.51 se obtienen los datos de la tabla d_colegio y se compara el nombre del colegio con los registros del estudiante, así se le agrega una columna con el campo id_colegio.

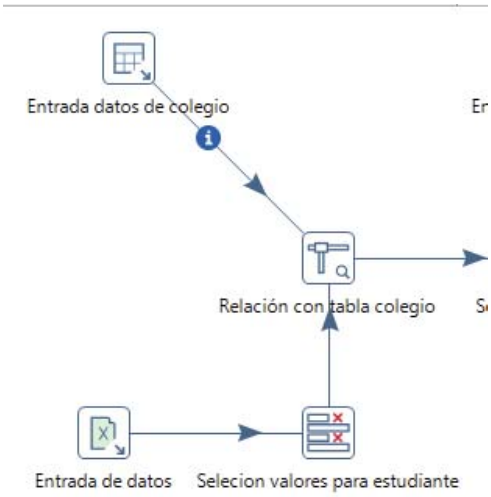


Figura B.51: Esquema de la entrada de datos desde la tabla d_colegio a la función Stream Value

Lo mismo se realiza para obtener el id de los datos socio económicos y académicos, como se indica en la Figura B.52, comparando el rut de los campos.

Figura B.52: Función Table Output, con los valores seleccionados para la tabla d_datosacademicos

En la Figura B.53, se puede apreciar los datos definitivos a ingresar a la tabla h_estudiante y así completar el proceso de carga de datos a la base de datos PostgreSQL. Con esto es posible utilizar los datos para generar Cubos ROLAP y así visualizar la información en Dashboards, análisis dinámicos y reportes.

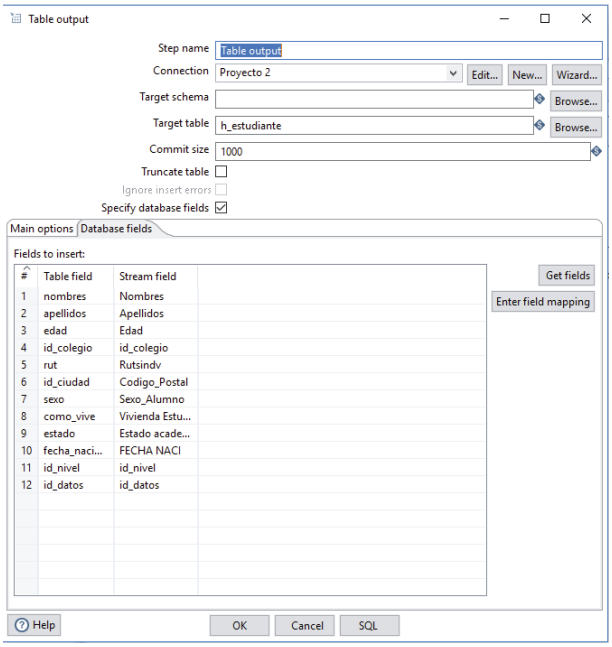


Figura B.53: Función Table Output, con los valores seleccionados para la tabla d_datosacademicos

C. Proceso de Minería de Datos

Luego de obtener las distintas tablas Excel, Figura C.1, desde el proceso ETL explicado en el Anexo B, con los datos necesarios para el análisis de Minería de datos, es necesario transformar este archivo a CSV, Figura C.2, para su posterior conversión a .arff, el archivo que utiliza el programa Weka.

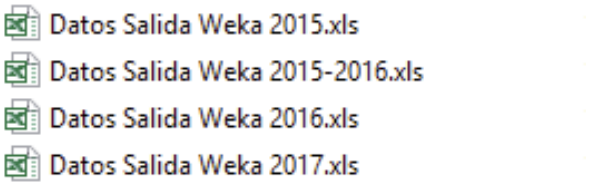


Figura C.1: Tabla Excel obtenida del proceso de ETL

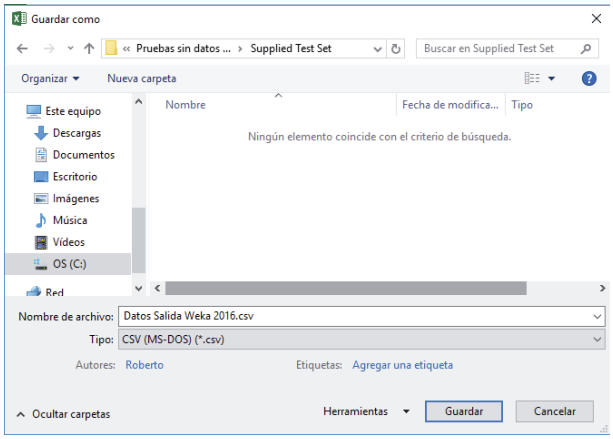


Figura C.2: Guardar tabla como archivo CSV

Para transformar el archivo .csv a .arff se utilizó una herramienta del programa Weka llamado ArffViewer, como se puede apreciar en la Figura C.3. En ésta, se abre el archivo csv, como se indica en la Figura C.4, para luego guardar el archivo en el formato .arff, Figura C.5, archivo con la extensión necesaria para la herramienta Weka.

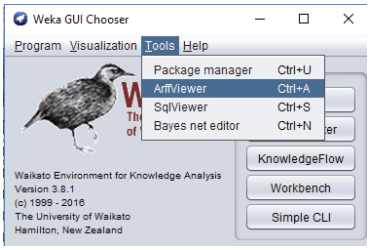


Figura C.3: Interfaz principal del programa Weka, selección de ArffViewer

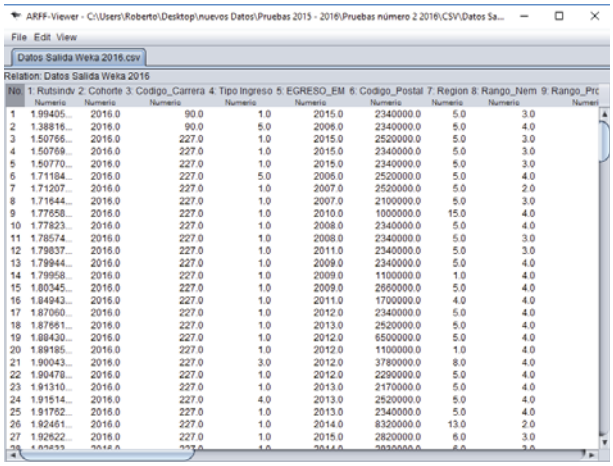


Figura C.4: Visualización del archivo csv abierto en ArffViewer

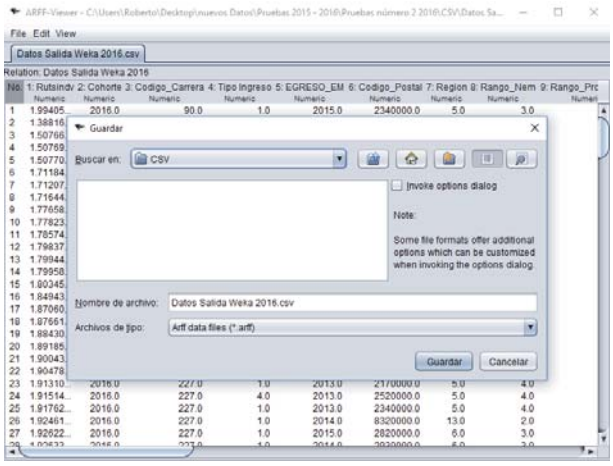


Figura C.5: Selección de dirección para guardar el archivo Arff

C.1. Preproceso

Weka permite aplicar una distintos tipos de filtros sobre los datos, permitiendo realizar transformaciones sobre éstos. Al pulsar el botón Choose dentro del recuadro Filter se despliega distintas opciones de filtros a escoger. Se ha de utilizar el filtro NumericToNominal, que transforma un conjunto de valores numéricos a valores nominales. Este filtrado transforma los datos del archivo arff con elementos numéricos en atributos nominales para que los datos sean cuantificables. En la Figura C.6 y C.7 se puede ver la diferencia antes y después del preproceso de los datos.

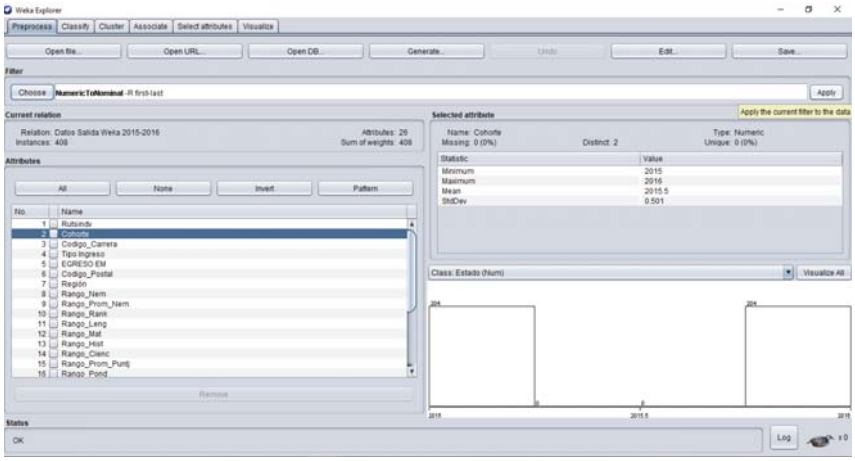


Figura C.6: Archivo Arff con datos numéricos

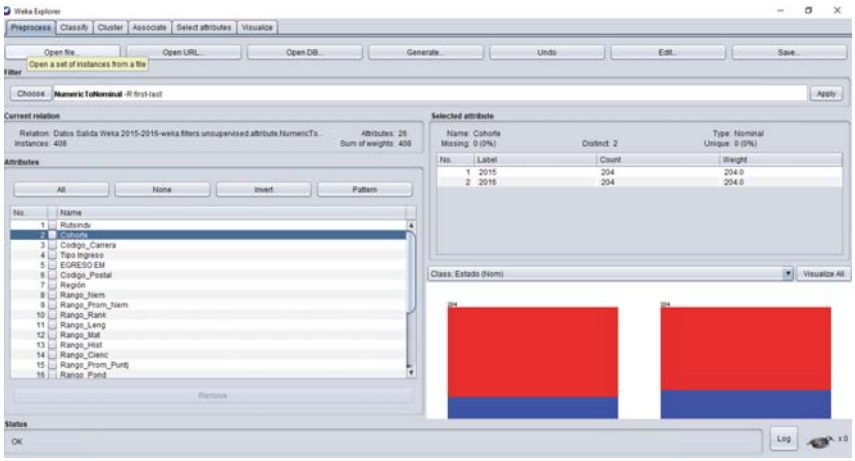


Figura C.7: Archivo Arff con datos nominales, luego del preproceso

Para la Clase predictiva, como se puede apreciar en la Figura C.8 muestra en color azul los alumnos que desertaron, mientras que el color rojo muestra a los alumnos que siguen matriculados.

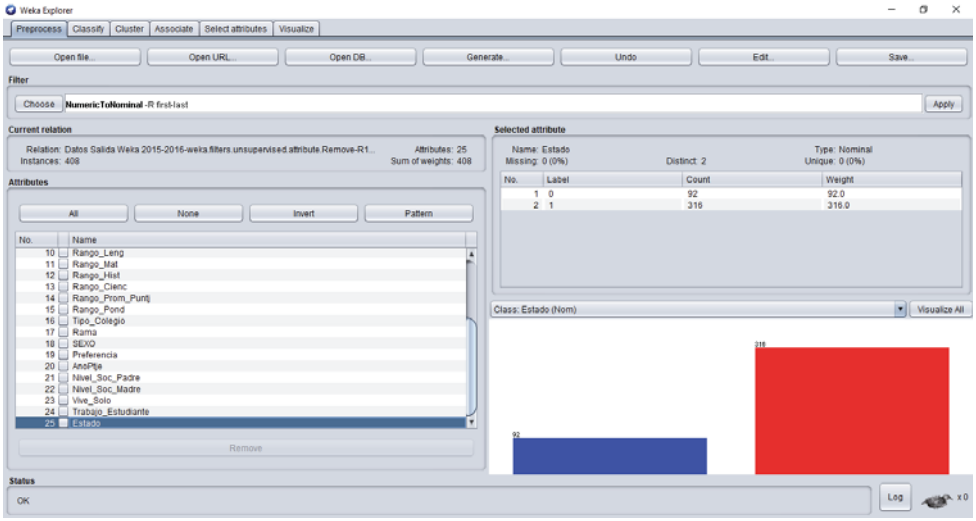


Figura C.8: Número de alumnos con estado desertor y matriculado

C.2. Clasificación

Para entrar en el modo clasificación, se ha de seleccionar la segunda pestaña en la zona superior del explorador de Weka.Para entrar en el modo clasificación, se ha de seleccionar la segunda pestaña en la zona superior del explorador de Weka. Luego se ha de elegir un modelo clasificador y configurarlo a gusto, para ello se pulsa sobre el botón Choose dentro del área Classifier, como se indica en la Figura C.9. Para el presente proyecto se han de utilizar los modelos clasificadores, NavesBayes, MLP y SMO que se detallan en la Sección 8.

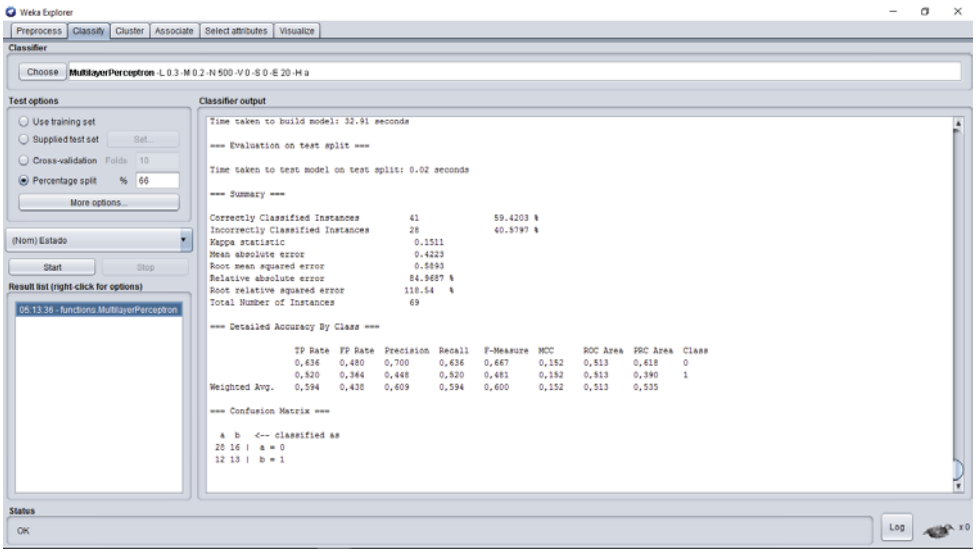


Figura C.9: Interfaz de clasificación de la herramienta Weka

Después de elegir el clasificador y sus características, se procede a seleccionar el modo de entrenamiento (Test Options). Weka tiene 4 modos de entrenamiento [45]:

- **Use training set**: Con esta opción Weka entrenará el método con todos los datos disponibles y luego lo aplicará otra vez sobre los mismos datos. Esta opción no sera utilizada.
- **Supplied test set**: Esta opción permite la oportunidad de seleccionar, pulsando el botón Set, un fichero de datos con el que se probará el clasificador obtenido con el método de clasificación usado y los datos iniciales. Esta opción sera utilizada para probar el modelo, utilizando los datos reales del cohorte 2016 y 2017.
- **Cross-validation**: Weka realizará una validación cruzada estratificada del número de particiones dado (Folds). Este método consiste en dado un número n se divide los datos en n partes y, por cada parte, se construye el clasificador con las n - 1 partes restantes y se prueba con ésta y así, por cada una de las n particiones. Esta opción no sera utilizada.
- **Percentage split**: Se define un porcentaje con el que se construirá el clasificador y con el porcentaje restante se probará la clasificación. Para las pruebas del cohorte 2016 se utilizó un 66.67 % de training, y para el cohorte 2017 un 67.1 %.

Los resultados de estas pruebas se han de detallar en el Anexo E.

D. Reporte y Dashboard

D.1. Pentaho User Console

El sistema de BI de Pentaho ofrece un servidor local donde poder trabajar las distintas funcionalidades de análisis, reporte y dashboarding. El servidor se aloja de forma local y es posible entrar como administrador o usuario común mediante una clave previamente asignada por el sistema, como se indica en la Figura D.1.

Figura D.1: Pantalla login del Pentaho User Console

Antes de comenzar cualquier función es necesario conectar la base de datos previamente cargada en el Anexo B, Sección B.4, para esto dentro de la pantalla principal del Pentaho User Console, en la Figura D.2, es necesario seleccionar New ->Data Source.

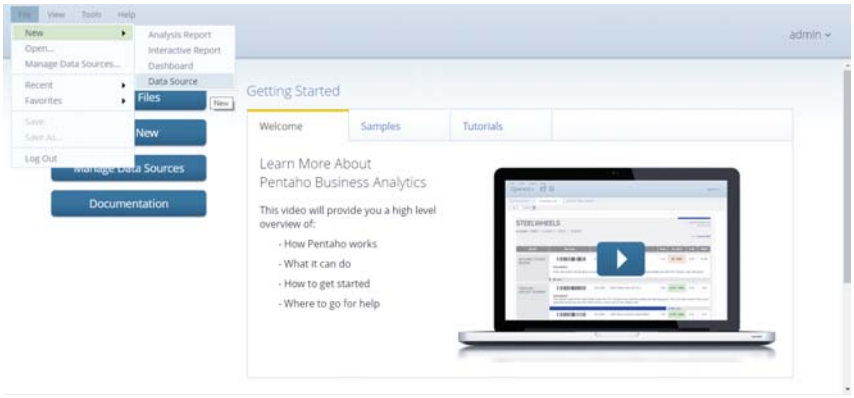


Figura D.2: Pantalla principal del Pentaho User Console, con la opción New Data Source seleccionada

Para conectar la base de datos, es necesario llenar su información como en la Figura B.28.

Luego de tener una exitosa conexión a la base de datos es necesario nombrar la fuente de datos, en este caso Proyecto 2 y seleccionarla para utilizarla en Reporting and Analysis, como se aprecia en la Figura D.3.

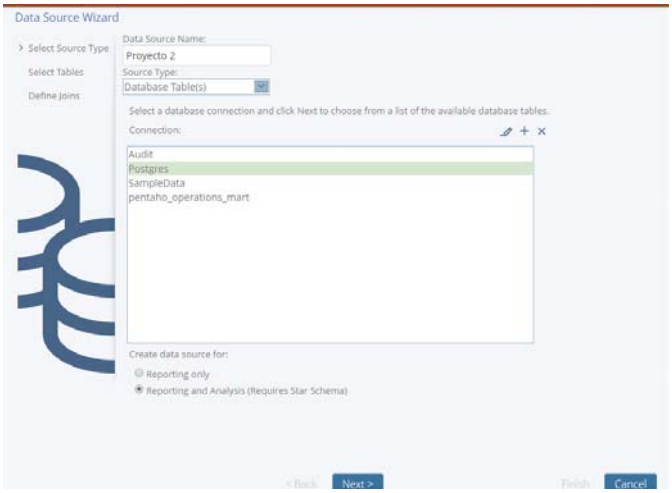


Figura D.3: La conexión Postgres ya completada y lista para utilizar

El siguiente paso es seleccionar las tablas a utilizar para el Modelo ROLAP, en este caso se seleccionan las cinco tablas de la Base de Datos creada con anterioridad. La tabla h_estudiante, d_ciudad, d_colegio, d_datosacademicos y d_nivelsocioeconomico. Por último, como se indica en la Figura D.4, se selecciona como la tabla de hecho (Fact Table) a la tabla h_estudiante.

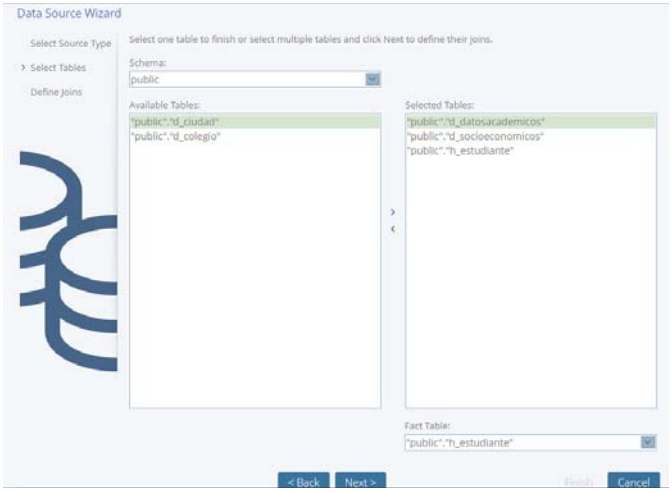


Figura D.4: Selección de las tablas a utilizar

A continuación, es necesario agregar las uniones de las claves foráneas de la tabla de hecho h_estudiante con cada una de sus dimensiones d_ciudad, d_colegio, d_datosacademicos y d_-nivelsocioeconomico, como se aprecia en la Figura D.5

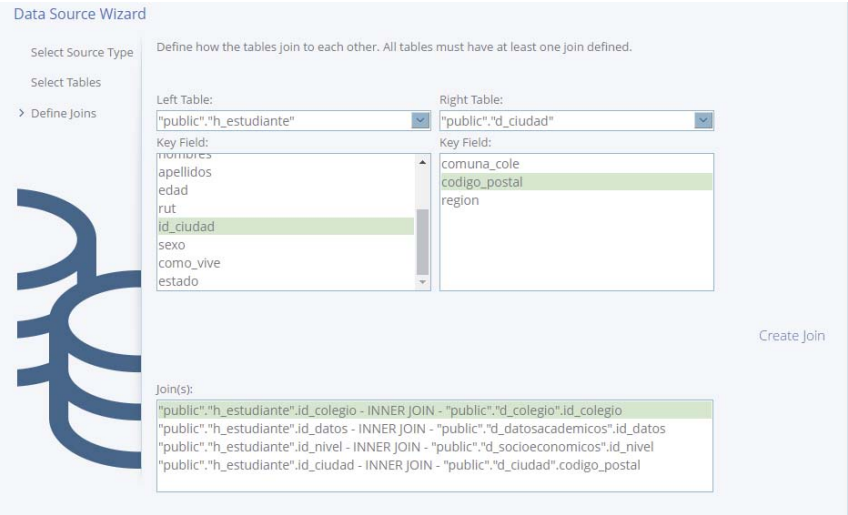


Figura D.5: Selección de uniones entre la tabla de hechos y sus dimensiones

Dentro de la pantalla principal de Pentaho User Console, es necesario administrar la fuente de datos para su utilización en los procesos de análisis, reporte y dashboarding. Primero se debe de seleccionar la fuente de datos, como se indica en la Figura D.6.

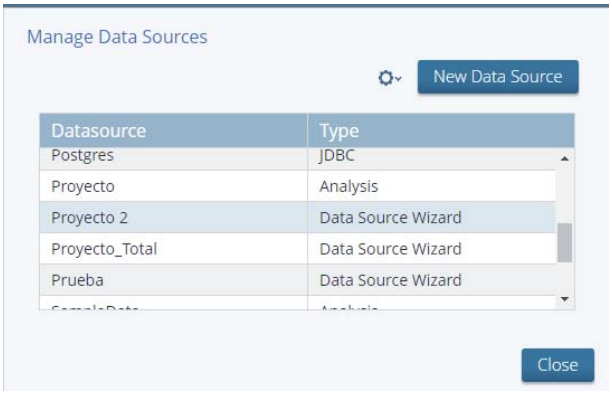


Figura D.6: Selección de la fuente de datos a gestionar

Luego es necesario indicar la métrica a utilizar para la funcionalidad del análisis, en este caso se utilizará la función de cuenta del número de alumnos por rut y las distintas dimensiones a medir por los valores de las tablas de dimensión como se puede apreciar en la Figura D.7.

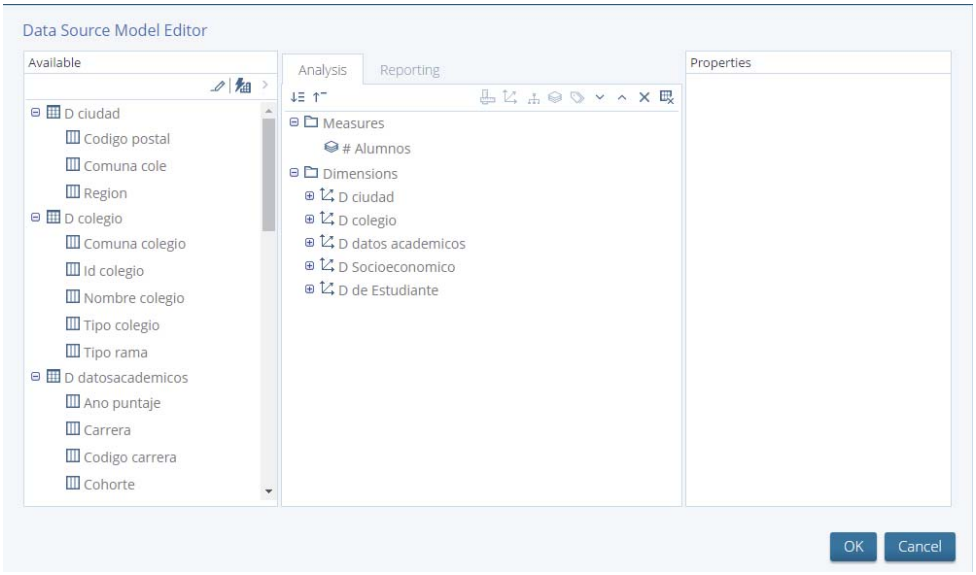


Figura D.7: Selección de las dimensiones y métricas a utilizar en los reportes de análisis

D.2. Analysis Report

Ya con la fuente de datos cargada y configurada se puede iniciar un nuevo Analysis Report, como se puede observar en la Figura D.8.

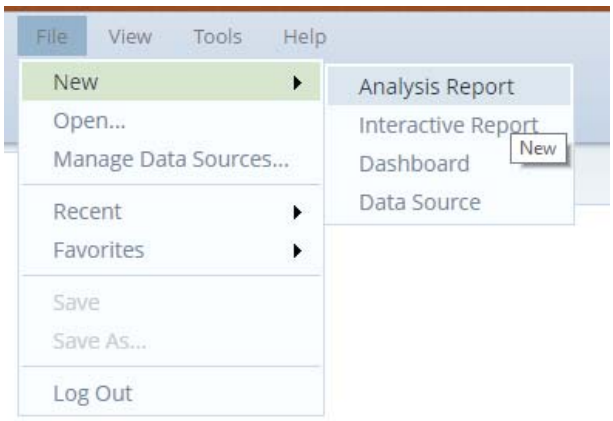


Figura D.8: Selección de un nuevo Analysis Report

Se selecciona la fuente de datos a utilizar, Figura D.9, para luego en la pantalla de Analysis Report comenzar a medir a los estudiantes según distintas métricas, un ejemplo de esto es la Figura D.10 donde se mide la cantidad de alumnos con distintos puntajes en la prueba de

lenguaje por región. Estos reportes pueden ser exportados a distintos formatos para que otros usuarios puedan visualizarlos.

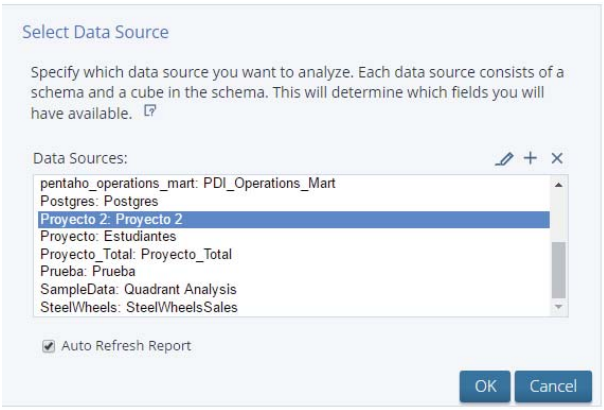


Figura D.9: Selección de fuente de datos a utilizar

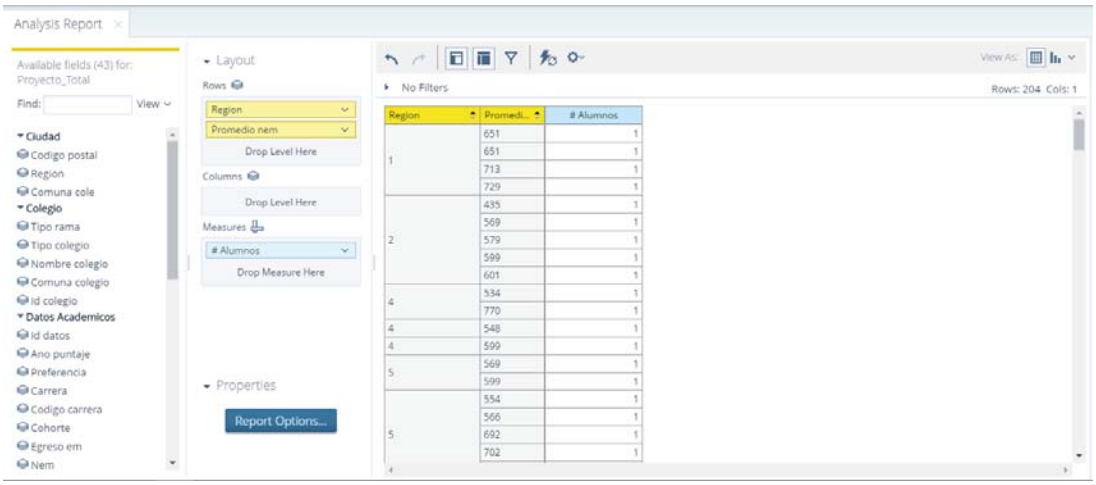


Figura D.10: Pantalla principal de Analysis Report

D.3. Interactive Report

Para iniciar un Interactive Report es necesario generar un nuevo archivo Interactive Report, Figura D.11.

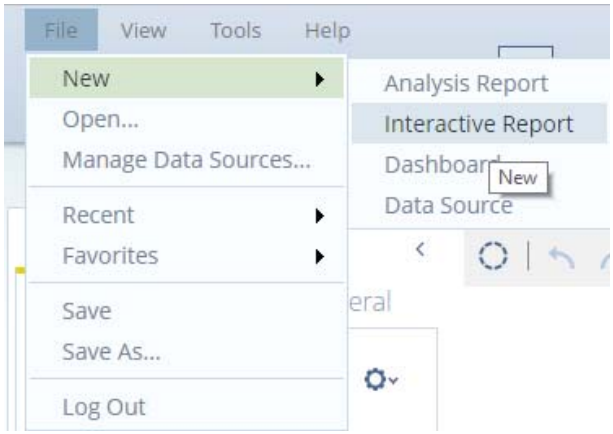


Figura D.11: Selección de un nuevo Interactive Report

Luego se ha de seleccionar la fuente de datos a utilizar, Figura D.12, para finalmente realizar distintos reportes interactivos con la información que se necesite, como se aprecia en la Figura D.13 con capacidad de exportar a PDF, Excel, entre otros.

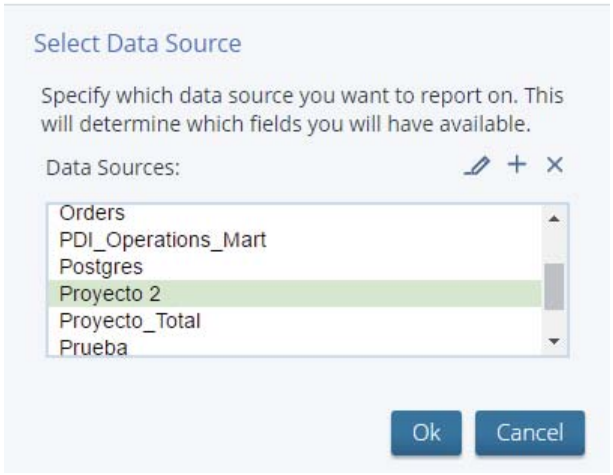


Figura D.12: Selección de la fuente de datos a utilizar

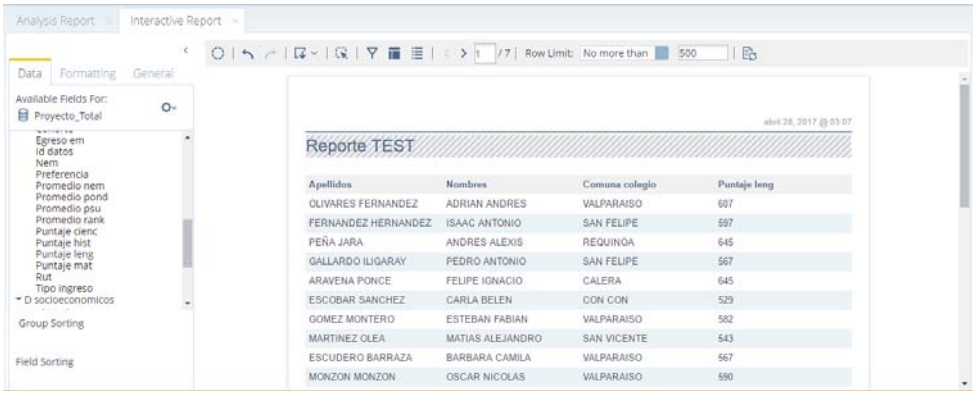


Figura D.13: Pantalla principal de trabajo de Interactive Report

D.4. Dashboard

Por último, Pentaho User Console trae una herramienta de Dashboard para visualizar distintos análisis y reportes previos y generar nuevos análisis para su uso, un ejemplo de esto es la Figura D.14

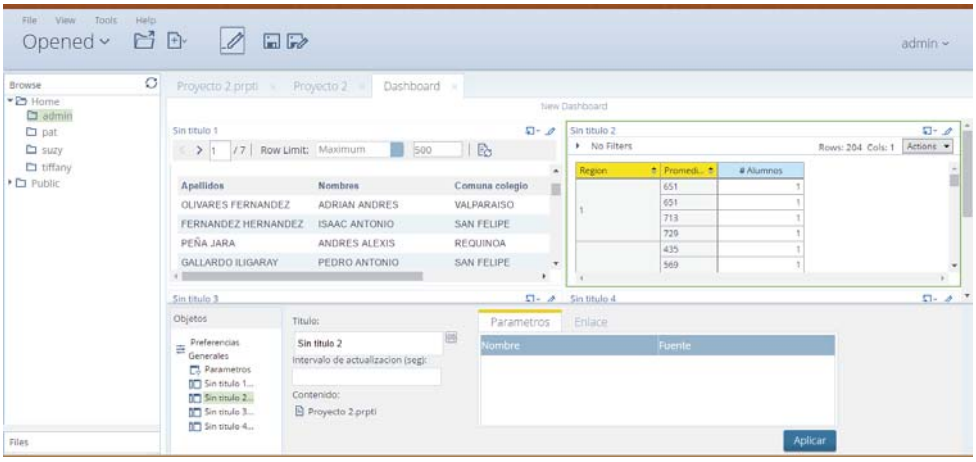


Figura D.14: Pantalla principal Dashboard

E. Pruebas Clasificación para el año 2016

Para cada uno de los modelos descriptivos utilizados (MLP, SMO y NavesBayes) se han de utilizar los modos de entrenamiento Supplied Test Set y Percentage Split, para las pruebas. Además, de realizar pruebas con los valores ingresados de forma random (Nivel educacional del padre, nivel educacional de la madre, si el alumno trabaja y si el alumno vive solo) y sin estos valores, para las pruebas del cohorte 2016. Los datos utilizados para el training son de 408 alumnos (204 cada cohorte, 2015 y 2016, con un 66.67 % de training para el modo Percentage Split) con sus datos reales de matricula y para el testing, el cohorte 2016 con su estado de matriculado para todos los estudiantes de dicha generación.

E.1. Datos Random

E.1.1. Supplied Test Set y Porcentage Split

Las pruebas realizadas con Supplied Test Set y Porcentage Split, al utilizar datos random entregaron los mismos resultados con los algoritmos probados, los resultados están en las tablas E.1, E.1 y E.1.

E.1.1.1. MLP

```
=== Summary ===

Correctly Classified Instances         154               75.4902 %
Incorrectly Classified Instances        50               24.5098 %
Kappa statistic                          0
Mean absolute error                      0.2462
Root mean squared error                  0.4829
Relative absolute error                108.531  %
Root relative squared error            212.8949 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,245    0,000      0,000    0,000      0,000    ?         ?         0
               0,755    0,000    1,000      0,755    0,860      0,000    ?         1,000     1
Weighted Avg.  0,755    0,000    1,000      0,755    0,860      0,000    0,000     1,000
```

Figura E.1: Resultados MLP Random usando Supplied Test Set

E.1.1.2. SMO

```
=== Summary ===

Correctly Classified Instances         170               83.3333 %
Incorrectly Classified Instances        34               16.6667 %
Kappa statistic                          0
Mean absolute error                      0.1667
Root mean squared error                  0.4082
Relative absolute error                 73.4767 %
Root relative squared error            179.9804 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,167    0,000      0,000    0,000      0,000    ?         ?         0
               0,833    0,000    1,000      0,833    0,909      0,000    ?         1,000     1
Weighted Avg.  0,833    0,000    1,000      0,833    0,909      0,000    0,000     1,000
```

Figura E.1: Resultados SMO Random usando Supplied Test Set

E.1.1.3. NavesBayes

```
=== Summary ===

Correctly Classified Instances         171               83.8235 %
Incorrectly Classified Instances        33               16.1765 %
Kappa statistic                          0
Mean absolute error                      0.2506
Root mean squared error                  0.3651
Relative absolute error                110.4592 %
Root relative squared error            160.9799 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,162    0,000      0,000    0,000      0,000    ?         ?         0
               0,838    0,000    1,000      0,838    0,912      0,000    ?         1,000     1
Weighted Avg.  0,838    0,000    1,000      0,838    0,912      0,000    0,000     1,000
```

Figura E.1: Resultados NavesBayes No Random usando Supplied Test Set

E.2. Datos No Random

E.2.1. Supplied Test Set

E.2.1.1. MLP

```
=== Summary ===

Correctly Classified Instances         155               75.9804 %
Incorrectly Classified Instances        49               24.0196 %
Kappa statistic                          0
Mean absolute error                      0.243
Root mean squared error                  0.4739
Relative absolute error                107.122  %
Root relative squared error            208.9334 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,240    0,000      0,000    0,000      0,000    ?         ?         0
               0,760    0,000    1,000      0,760    0,864      0,000    ?         1,000     1
Weighted Avg.  0,760    0,000    1,000      0,760    0,864      0,000    0,000     1,000

=== Confusion Matrix ===

   a   b   <-- classified as
   0   0 |   a = 0
  49 155 |   b = 1
```

Figura E.1: Resultados MLP No Random usando Supplied Test Set

E.2.1.2. SMO

```
=== Summary ===

Correctly Classified Instances         178               87.2549 %
Incorrectly Classified Instances        26               12.7451 %
Kappa statistic                          0
Mean absolute error                      0.1275
Root mean squared error                  0.357
Relative absolute error                 56.1881 %
Root relative squared error            157.3883 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,000    0,127    0,000      0,000    0,000      0,000    ?         ?         0
               0,873    0,000    1,000      0,873    0,932      0,000    ?         1,000     1
Weighted Avg.  0,873    0,000    1,000      0,873    0,932      0,000    0,000     1,000

=== Confusion Matrix ===

   a   b   <-- classified as
   0   0 |   a = 0
  26 178 |   b = 1
```

Figura E.1: Resultados SMO No Random usando Supplied Test Set

E.2.1.3. NavesBayes

```
=== Summary ===

Correctly Classified Instances         168               82.3529 %
Incorrectly Classified Instances        36               17.6471 %
Kappa statistic                          0
Mean absolute error                      0.2559
Root mean squared error                  0.3671
Relative absolute error                112.8189 %
Root relative squared error            161.8486 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,176    0,000      0,000    0,000      0,000    ?         ?         0
                 0,824    0,000    1,000      0,824    0,903      0,000    ?         1,000     1
Weighted Avg.    0,824    0,000    1,000      0,824    0,903      0,000    0,000     1,000

=== Confusion Matrix ===

   a   b   <-- classified as
   0   0 |   a = 0
  36 168 |   b = 1
```

Figura E.1: Resultados NavesBayes No Random usando Supplied Test Set

E.2.2. Porcentage Split

E.2.2.1. MLP

```
=== Summary ===

Correctly Classified Instances         154               75.4902 %
Incorrectly Classified Instances        50               24.5098 %
Kappa statistic                          0
Mean absolute error                      0.2646
Root mean squared error                  0.4824
Relative absolute error                116.6603 %
Root relative squared error            212.6701 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,245    0,000      0,000    0,000      0,000    ?         ?         0
                 0,755    0,000    1,000      0,755    0,860      0,000    ?         1,000     1
Weighted Avg.    0,755    0,000    1,000      0,755    0,860      0,000    0,000     1,000

=== Confusion Matrix ===

   a   b   <-- classified as
   0   0 |   a = 0
  50 154 |   b = 1
```

Figura E.1: Resultados MLP No Random usando Porcentage Spli

E.2.2.2. SMO

```
=== Summary ===

Correctly Classified Instances         179               87.7451 %
Incorrectly Classified Instances        25               12.2549 %
Kappa statistic                          0
Mean absolute error                      0.1225
Root mean squared error                  0.3501
Relative absolute error                 54.027  %
Root relative squared error            154.3319 %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,123    0,000      0,000    0,000      0,000    ?         ?         0
                 0,877    0,000    1,000      0,877    0,935      0,000    ?         1,000     1
Weighted Avg.    0,877    0,000    1,000      0,877    0,935      0,000    0,000     1,000

=== Confusion Matrix ===

   a   b   <-- classified as
   0   0 |   a = 0
  25 179 |   b = 1
```

Figura E.1: Resultados SMO No Random usando Porcentage Spli

E.2.2.3. NavesBayes

```
=== Summary ===

Correctly Classified Instances         167               81.8627 %
Incorrectly Classified Instances        37               18.1373 %
Kappa statistic                          0
Mean absolute error                      0.255
Root mean squared error                  0.3665
Relative absolute error                112.4348 %
Root relative squared error            161.589  %
Total Number of Instances              204

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,181    0,000      0,000    0,000      0,000    ?         ?         0
                 0,819    0,000    1,000      0,819    0,900      0,000    ?         1,000     1
Weighted Avg.    0,819    0,000    1,000      0,819    0,900      0,000    0,000     1,000

=== Confusion Matrix ===

   a   b   <-- classified as
   0   0 |   a = 0
  37 167 |   b = 1
```

Figura E.1: Resultados NavesBayes No Random usando Porcentage Spli

F. Resultados Clasificación

F.1. Resultados clasificación cohorte 2016

Rut	E. Original	MLP	SMO	Nayves
19940592	1	2:1	2:1	2:1
13881638	1	2:1	2:1	2:1
15076636	1	2:1	2:1	1:0
15076951	0	1:0	2:1	2:1
15077054	1	2:1	2:1	2:1
17118499	1	2:1	2:1	2:1
17120763	1	2:1	2:1	2:1
17164469	0	1:0	1:0	1:0
17765875	1	2:1	2:1	2:1
17782372	1	2:1	2:1	1:0
17857442	1	2:1	2:1	1:0
17983727	1	2:1	2:1	2:1
17994485	0	1:0	1:0	2:1
17995842	1	2:1	2:1	2:1
18034519	1	2:1	2:1	2:1
18494371	1	2:1	2:1	2:1
18706047	1	2:1	2:1	2:1
18766110	1	2:1	2:1	2:1
18843055	1	2:1	2:1	2:1
18918552	1	2:1	2:1	2:1
19004315	0	1:0	1:0	1:0
19047814	1	2:1	2:1	2:1
19131049	1	2:1	2:1	2:1
19151474	1	2:1	2:1	2:1
19176215	1	2:1	2:1	2:1
19246151	1	2:1	2:1	2:1
19262224	1	2:1	2:1	2:1
19263355	1	2:1	2:1	2:1
19274163	1	2:1	2:1	2:1
19275699	1	2:1	2:1	2:1
19326298	0	1:0	1:0	1:0
19339344	1	2:1	1:0	2:1
19351288	1	2:1	2:1	2:1
19357813	0	1:0	2:1	1:0
19390087	0	1:0	1:0	1:0
19394443	0	1:0	1:0	1:0
19394805	1	2:1	1:0	2:1

Tabla 4: Resultados de las mejores pruebas para los estudiantes cohorte 2016

Rut	E. Original	MLP	SMO	Nayves
19437068	1	2:1	2:1	2:1
19448699	1	2:1	1:0	1:0
19467668	0	1:0	1:0	2:1
19469054	0	1:0	2:1	2:1
19475437	1	2:1	2:1	2:1
19484636	0	1:0	1:0	1:0
19488769	0	1:0	1:0	1:0
19489138	0	1:0	1:0	2:1
19490616	1	2:1	2:1	2:1
19497799	1	2:1	2:1	1:0
19525992	1	2:1	2:1	2:1
19527109	1	2:1	2:1	2:1
19552185	1	2:1	1:0	1:0
19569842	1	2:1	2:1	2:1
19580045	1	2:1	2:1	2:1
19581364	1	2:1	2:1	2:1
19582606	1	2:1	2:1	1:0
19589083	1	2:1	2:1	2:1
19604430	1	2:1	2:1	1:0
19604560	0	1:0	2:1	2:1
19617044	1	2:1	2:1	2:1
19617083	1	2:1	2:1	2:1
19617258	1	2:1	2:1	2:1
19617478	1	2:1	2:1	2:1
19618759	1	2:1	2:1	2:1
19619730	1	2:1	2:1	2:1
19620381	0	1:0	2:1	2:1
19660686	1	2:1	2:1	2:1
19671124	1	2:1	2:1	1:0
19679368	1	2:1	1:0	2:1
19679822	1	2:1	2:1	2:1
19688674	1	2:1	2:1	2:1
19711565	1	2:1	2:1	2:1
19727480	1	2:1	2:1	2:1
19729221	1	2:1	2:1	2:1
19736911	1	2:1	2:1	2:1
19738285	1	2:1	2:1	2:1
19750034	1	2:1	2:1	2:1
19757223	1	2:1	2:1	2:1

Tabla 5: Resultados de las mejores pruebas para los estudiantes cohorte 2016

Rut	E. Original	MLP	SMO	Nayves
19786754	1	2:1	2:1	2:1
19786784	1	2:1	2:1	2:1
19787113	0	1:0	1:0	1:0
19787201	1	2:1	2:1	1:0
19790924	1	2:1	2:1	2:1
19800459	1	2:1	2:1	2:1
19819424	1	2:1	2:1	2:1
19821715	0	1:0	1:0	1:0
19824937	0	1:0	1:0	1:0
19850701	1	2:1	2:1	2:1
19857674	1	2:1	2:1	1:0
19868640	1	2:1	2:1	2:1
19884554	0	1:0	2:1	2:1
19887617	0	1:0	1:0	2:1
19887900	0	1:0	1:0	2:1
19903106	1	2:1	2:1	2:1
19910293	0	1:0	1:0	1:0
19940101	1	2:1	2:1	2:1
19940563	1	2:1	2:1	2:1
19980741	1	2:1	2:1	2:1
23994308	1	2:1	2:1	2:1
15076423	1	2:1	2:1	2:1
15076507	1	2:1	2:1	2:1
15076595	0	1:0	2:1	2:1
15076628	1	2:1	2:1	2:1
17119248	1	2:1	2:1	2:1
17716963	0	1:0	1:0	1:0
17976811	0	2:1	1:0	2:1
18033978	1	2:1	2:1	2:1
18106159	1	2:1	2:1	2:1
18123218	0	1:0	1:0	1:0
18162154	0	1:0	1:0	1:0
18329249	1	2:1	2:1	2:1
18583011	1	2:1	2:1	2:1
18618307	1	2:1	2:1	2:1
18649838	1	2:1	2:1	2:1
18654678	1	2:1	2:1	2:1
18683758	1	2:1	2:1	2:1
18704276	0	1:0	2:1	2:1

Tabla 6: Resultados de las mejores pruebas para los estudiantes cohorte 2016

Rut	E. Original	MLP	SMO	Nayves
18704576	0	1:0	1:0	1:0
18705168	1	2:1	2:1	2:1
18750446	0	1:0	2:1	2:1
18758163	0	1:0	1:0	1:0
18760669	1	2:1	2:1	2:1
18848707	0	1:0	2:1	1:0
18879274	0	1:0	1:0	1:0
18890316	1	2:1	2:1	2:1
18907502	1	2:1	2:1	2:1
18914962	0	1:0	1:0	1:0
18917423	1	2:1	2:1	2:1
18949969	1	2:1	2:1	2:1
18996909	1	2:1	2:1	2:1
18999474	0	1:0	2:1	2:1
19014084	0	1:0	2:1	2:1
19040413	0	1:0	2:1	2:1
19152796	1	2:1	2:1	2:1
19208530	1	2:1	2:1	2:1
19209363	0	2:1	2:1	2:1
19243769	1	2:1	2:1	2:1
19254780	0	1:0	1:0	2:1
19264177	1	2:1	2:1	2:1
19291386	1	2:1	2:1	2:1
19291790	1	2:1	2:1	2:1
19337926	1	2:1	2:1	2:1
19338114	1	2:1	2:1	2:1
19339975	1	2:1	2:1	2:1
19363085	1	2:1	2:1	2:1
19388282	0	1:0	1:0	2:1
19432978	1	2:1	2:1	2:1
19443838	1	2:1	2:1	2:1
19461777	0	1:0	2:1	2:1
19471040	1	2:1	2:1	2:1
19471789	0	1:0	2:1	2:1
19490522	0	1:0	2:1	2:1
19498195	1	2:1	2:1	2:1
19499028	1	2:1	2:1	2:1
19500433	1	2:1	2:1	2:1
19541455	1	2:1	2:1	2:1

Tabla 7: Resultados de las mejores pruebas para los estudiantes cohorte 2016

Rut	E. Original	MLP	SMO	Nayves
19566529	1	2:1	2:1	2:1
19579955	1	2:1	2:1	2:1
19594086	0	1:0	2:1	2:1
19610623	1	2:1	2:1	2:1
19612937	1	2:1	2:1	2:1
19613265	1	2:1	2:1	2:1
19614791	1	2:1	2:1	2:1
19616787	1	2:1	2:1	2:1
19617101	1	2:1	2:1	2:1
19617142	1	2:1	2:1	2:1
19617161	1	2:1	2:1	2:1
19617676	1	2:1	2:1	2:1
19617743	1	2:1	2:1	2:1
19617878	1	2:1	2:1	2:1
19618478	1	2:1	2:1	2:1
19619062	1	2:1	2:1	2:1
19619104	1	2:1	2:1	2:1
19619390	1	2:1	2:1	2:1
19620321	1	2:1	2:1	2:1
19631869	0	1:0	1:0	2:1
19642826	0	1:0	1:0	2:1
19668433	0	1:0	2:1	1:0
19728077	1	2:1	2:1	1:0
19728422	1	2:1	2:1	2:1
19735794	1	2:1	2:1	2:1
19736853	1	2:1	2:1	2:1
19756644	0	1:0	2:1	2:1
19758188	1	2:1	2:1	2:1
19760706	0	1:0	2:1	2:1
19772251	1	2:1	2:1	2:1
19772587	1	2:1	2:1	2:1
19773278	1	2:1	2:1	2:1
19773884	0	1:0	2:1	2:1
19774094	1	2:1	2:1	2:1
19774382	1	2:1	2:1	2:1
19787055	1	2:1	2:1	2:1
19791106	1	2:1	2:1	2:1
19791304	1	2:1	2:1	2:1
19828774	1	2:1	2:1	2:1

Tabla 8: Resultados de las mejores pruebas para los estudiantes cohorte 2016

Rut	E. Original	MLP	SMO	Nayves
19830601	1	2:1	2:1	2:1
19832358	1	2:1	2:1	2:1
19843104	1	2:1	2:1	2:1
19851503	1	2:1	2:1	2:1
19883713	0	1:0	2:1	2:1
19898838	1	2:1	2:1	2:1
19898956	1	2:1	2:1	2:1
19940753	0	1:0	1:0	2:1
19980727	1	2:1	2:1	2:1
19980793	1	2:1	2:1	2:1
21700188	1	2:1	2:1	2:1

Tabla 9: Resultados de las mejores pruebas para los estudiantes cohorte 2016

F.2. Resultados clasificación cohorte 2017

Rut/Instancia	actual	predicted	error	prediction
14736503	2:1	2:1		1
16701618	2:1	2:1		0.983
17177231	2:1	2:1		1
17618472	2:1	2:1		1
17984211	2:1	2:1		1
18102700	2:1	2:1		0.986
18114162	2:1	2:1		1
18265098	2:1	2:1		1
18268256	2:1	2:1		1
18396257	2:1	1:0	+	0.996
18396551	2:1	2:1		1
18503848	2:1	2:1		1
18592651	2:1	2:1		1
18619068	2:1	2:1		1
18619430	2:1	2:1		1
18650263	2:1	2:1		1
18659337	2:1	2:1		1
18689453	2:1	2:1		1
18703462	2:1	2:1		1
18719567	2:1	1:0	+	0.959
18751756	2:1	1:0	+	1
18781990	2:1	2:1		1
18783264	2:1	1:0	+	1
18784223	2:1	2:1		1
18786864	2:1	1:0	+	1
18908038	2:1	2:1		1
18915158	2:1	2:1		1
18986470	2:1	1:0	+	0.999
19015771	2:1	2:1		1
19041272	2:1	2:1		0.986
19047391	2:1	1:0	+	0.858
19049258	2:1	2:1		1
19049311	2:1	2:1		0.988
19083344	2:1	2:1		1
19101195	2:1	1:0	+	1
19150907	2:1	1:0	+	1
19152198	2:1	2:1		1
19153797	2:1	1:0	+	1
19154059	2:1	2:1		1

Tabla 10: Resultados de la clasificación de los estudiantes cohorte 2017, mediante MLP

Rut/Instancia	actual	predicted	error	prediction
19154113	2:1	2:1		0.992
19154617	2:1	2:1		1
19176789	2:1	1:0	+	1
19176887	2:1	2:1		1
19191807	2:1	2:1		1
19234064	2:1	2:1		1
19267473	2:1	2:1		0.987
19282756	2:1	2:1		1
19287985	2:1	2:1		1
19289616	2:1	2:1		1
19311783	2:1	2:1		0.988
19319550	2:1	2:1		1
19329961	2:1	1:0	+	0.974
19341715	2:1	2:1		1
19348740	2:1	1:0	+	1
19373803	2:1	2:1		1
19379257	2:1	2:1		1
19394268	2:1	1:0	+	1
19394923	2:1	2:1		0.997
19403722	2:1	1:0	+	1
19463640	2:1	2:1		1
19487832	2:1	2:1		1
19489044	2:1	2:1		1
19489428	2:1	1:0	+	1
19537505	2:1	2:1		1
19540715	2:1	2:1		1
19579793	2:1	2:1		0.995
19581562	2:1	2:1		1
19590633	2:1	2:1		1
19592209	2:1	2:1		1
19593420	2:1	2:1		1
19614178	2:1	1:0	+	0.974
19615044	2:1	2:1		1
19615063	2:1	1:0	+	0.695
19618166	2:1	2:1		0.982
19618282	2:1	2:1		1
19618529	2:1	2:1		0.988
19619674	2:1	2:1		1
19639662	2:1	2:1		1

Tabla 11: Resultados de la clasificación de los estudiantes cohorte 2017, mediante MLP

Rut/Instancia	actual	predicted	error	prediction
19640199	2:1	2:1		1
19658923	2:1	1:0	+	0.51
19662269	2:1	1:0	+	0.628
19664793	2:1	2:1		1
19664987	2:1	2:1		1
19665014	2:1	2:1		1
19665080	2:1	2:1		1
19665725	2:1	2:1		0.988
19671919	2:1	2:1		1
19673296	2:1	2:1		1
19678963	2:1	1:0	+	1
19679546	2:1	2:1		1
19695297	2:1	2:1		0.986
19699058	2:1	2:1		1
19699137	2:1	2:1		1
19700975	2:1	2:1		0.986
19727366	2:1	2:1		0.986
19728588	2:1	2:1		1
19728843	2:1	1:0	+	0.986
19736820	2:1	1:0	+	0.974
19740399	2:1	2:1		1
19744280	2:1	2:1		1
19749870	2:1	2:1		1
19773760	2:1	2:1		1
19774486	2:1	2:1		1
19776509	2:1	2:1		1
19776599	2:1	2:1		0.986
19776636	2:1	2:1		1
19776680	2:1	1:0	+	1
19776788	2:1	2:1		1
19776968	2:1	2:1		1
19776980	2:1	1:0	+	1
19792556	2:1	2:1		1
19818656	2:1	2:1		1
19825332	2:1	1:0	+	1
19829578	2:1	2:1		1
19851980	2:1	1:0	+	1
19854304	2:1	1:0	+	0.999
19864174	2:1	2:1		1

Tabla 12: Resultados de la clasificación de los estudiantes cohorte 2017, mediante MLP

Rut/Instancia	actual	predicted	error	prediction
19873045	2:1	2:1		1
19880521	2:1	2:1		1
19899192	2:1	2:1		1
19903732	2:1	1:0	+	0.972
19911712	2:1	2:1		1
19919145	2:1	2:1		1
19924656	2:1	2:1		1
19924662	2:1	2:1		1
19925203	2:1	1:0	+	0.985
19926787	2:1	2:1		1
19928303	2:1	1:0	+	1
19931170	2:1	2:1		1
19939690	2:1	2:1		0.762
19939761	2:1	2:1		1
19939961	2:1	1:0	+	0.994
19940440	2:1	2:1		1
19940632	2:1	2:1		1
19940734	2:1	2:1		1
19940840	2:1	2:1		1
19940860	2:1	2:1		1
19942831	2:1	2:1		0.988
19946019	2:1	2:1		1
19959281	2:1	2:1		0.988
19968186	2:1	2:1		1
19971629	2:1	1:0	+	1
19972190	2:1	2:1		1
19973815	2:1	2:1		1
19974843	2:1	2:1		1
19976706	2:1	2:1		1
19977071	2:1	2:1		1
19977202	2:1	1:0	+	0.985
19980696	2:1	1:0	+	1
19981438	2:1	2:1		1
19981487	2:1	2:1		0.988
19981584	2:1	1:0	+	0.989
19981730	2:1	2:1		0.993
19982043	2:1	1:0	+	0.974
19982274	2:1	2:1		0.972
19982432	2:1	2:1		0.986

Tabla 13: Resultados de la clasificación de los estudiantes cohorte 2017, mediante MLP

Rut/Instancia	actual	predicted	error	prediction
19996715	2:1	2:1		1
19997394	2:1	2:1		1
20010651	2:1	2:1		1
20011461	2:1	1:0	+	0.78
20011492	2:1	2:1		1
20011949	2:1	2:1		0.989
20011973	2:1	2:1		1
20012495	2:1	2:1		1
20012919	2:1	2:1		1
20026704	2:1	2:1		1
20038932	2:1	1:0	+	1
20043750	2:1	2:1		1
20058162	2:1	2:1		1
20067266	2:1	1:0	+	1
20067339	2:1	2:1		1
20067354	2:1	2:1		1
20067462	2:1	2:1		1
20080519	2:1	2:1		1
20081839	2:1	2:1		1
20082211	2:1	2:1		1
20084043	2:1	2:1		0.999
20088634	2:1	2:1		1
20091921	2:1	2:1		1
20113160	2:1	2:1		1
20117846	2:1	2:1		1
20123526	2:1	2:1		1
20158372	2:1	2:1		1
20171531	2:1	2:1		1
20171656	2:1	1:0	+	1
20172375	2:1	1:0	+	1
20181368	2:1	2:1		1
20182101	2:1	2:1		1
20182315	2:1	2:1		1
20182490	2:1	1:0	+	1
20182559	2:1	2:1		1
20182785	2:1	2:1		0.888
20183151	2:1	1:0	+	1
20211095	2:1	1:0	+	1
20220417	2:1	1:0	+	0.959

Tabla 14: Resultados de la clasificación de los estudiantes cohorte 2017, mediante MLP

Rut/Instancia	actual	predicted	error	prediction
20239105	2:1	2:1		1
20239328	2:1	2:1		1
21441139	2:1	1:0	+	1
22091525	2:1	2:1		1
25612549	2:1	1:0	+	1

Tabla 15: Resultados de la clasificación de los estudiantes cohorte 2017, mediante MLP