UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE INFORMÁTICA

TESINA DE LICENCIATURA

Título: Aplicación de técnicas y estrategias de Inteligencia de Negocio (BI) para analizar/integrar información de los alumnos de la Facultad de Informática de la UNLP.

Autores: Claudia Yanina Kruzylko Ostojic
Director: Lic. Javier Díaz
Codirector: Lic. Ana Paola Amadeo y Lic. María Alejandra Osorio

Carrera: Licenciatura en Informática – Plan 1990

Resumen

La Facultad de Informática – UNLP cuenta con varios sistemas informáticos que almacenan información de sus estudiantes y graduados. El análisis de estos datos, puede facilitar a las autoridades la toma de decisiones. La cantidad de datos almacenados en las diferentes bases de datos, dificulta el análisis de los mismos de manera manual.
En esta tesina de grado, se aplican tareas y técnicas de minería de datos, a datos obtenidos de las bases de datos de diferentes sistemas utilizados por la Facultad de Informática – UNLP. Se aplican las tareas de clasificación y agrupamiento, teniendo en cuenta que son las dos tareas más usadas en el campo de la minería de datos aplicadas a la educación. El objetivo principal es la obtención de patrones y modelos de los datos obtenidos, que permitan definir un perfil de los estudiantes. También se desarrolló un sistema Ad-Hoc que permite la aplicación de las técnicas utilizadas y la observación de los resultados a usuarios no familiarizados con herramientas de minería de datos.
Este trabajo de grado se basa en el uso, en la medida de lo posible, de sistemas de software libre; tanto aquellos sistemas fuentes de obtención de los datos, como todas las herramientas utilizadas para el desarrollo de los análisis.

Palabras Claves

Business Intelligence, Proceso de Conocimiento en Bases de Datos (KDD), Minería de Datos, Minería de Datos Aplicada a la Educación (EDM), Preprocesamiento de Datos, Clasificación, Agrupamiento, Perfiles de Alumnos, Correlación de Datos de Diferentes Sistemas, Análisis Socio Demográfico, Análisis de Participación en las Cátedras, Redes Sociales, WEKA.

Trabajos Realizados

Se correlacionaron datos de diferentes sistemas usados por la Facultad de Informática; se aplicaron tareas y técnicas de minería de datos sobre los mismos, para determinar si existen patrones de comportamiento en ellos, que ayuden a identificar un perfil de los estudiantes.
Se creo una aplicación Ad-Hoc que permite a usuarios no relacionados con la minería de datos, observar los resultados obtenidos.
Se realizó también, una primera aproximación al análisis de datos obtenidos de las redes sociales; teniendo en cuenta con un caso específico.

Conclusiones

En este trabajo se experimentó el desafío de correlacionar datos de alumnos de la Facultad de Informática; obtenidos de su interacción con diferentes sistemas, basados en software libre, utilizados en esta unidad académica.
Se unificaron datos de tres sistemas fuentes diferentes; cada uno con su base de datos propia y representaciones distintas de algunos datos en común. El análisis realizado, permitió obtener modelos y patrones de comportamiento, que se pueden cotejar utilizando la herramienta desarrollada para tal fin en el marco de esta tesina.

Trabajos Futuros

Correlacionar, a los datos ya usados, datos de otros sistemas utilizados también por la Facultad de Informática – UNLP, que quedaron fuera del alcance de este trabajo de grado, para poder ampliar el perfil obtenido de los estudiantes.
Obtener datos referentes al uso que los estudiantes hacen de las redes sociales para poder distinguir la participación de los estudiantes en de los sistemas internos a la facultad de los sistemas externos.

Fecha de la presentación: NOVIEMBRE 2013

ÍNDICE DE CONTENIDOS

PREFACIO

Introducción

La analítica de datos, un concepto que evoluciona de la inteligencia de negocios, consiste en un conjunto de herramientas y técnicas informáticas, utilizadas para obtener información útil, y tal vez oculta, de una gran cantidad de datos almacenados en bases de datos masivas, para simplificar el análisis, sumarización y visualización de grandes volúmenes de datos.

La minería de datos es una rama de la inteligencia de negocios, que consiste en un conjunto de herramientas y técnicas, utilizadas para la extracción de conocimiento útil de grandes cantidades de datos históricamente almacenados. La aplicación de estas herramientas y técnicas tiene como objetivo la generación de modelos y patrones, que indiquen el comportamiento actual de los datos para poder tener un entendimiento mejor de los mismos y oportunamente poder predecir comportamiento futuro de datos actualmente desconocidos. En el mercado se puede encontrar una amplia cantidad de herramientas que facilitan la realización del análisis de los datos; muchas de ellas se distribuyen bajo licencia de Software Libre, lo que facilita la realización de este trabajo que se centrará en el uso de herramientas de Software Libre, tanto para la extracción de datos como para el análisis de los mismos.

La minería es parte de un proceso completo de extracción de conocimiento, formado por las siguientes etapas:

- Selección de los datos desde los sistemas fuentes.
- Pre-procesamiento y transformación para obtener una vista minable, que contiene los datos aplanados listos para ser analizados.
- Modelado con la aplicación de las técnicas de minería de datos.
- Interpretación y evaluación de los resultados.

Motivación

Los sistemas informáticos, relacionados a la Facultad, que los alumnos pueden utilizar desde que ingresan a la Facultad hasta que obtienen el diploma, son entre otros, sistemas de ingreso, sistema de alumnos, becas, libreta sanitaria, sistemas de biblioteca, entornos virtuales. El uso de estos sistemas genera numerosos registros, que contienen datos de los estudiantes, lo que permitiría el "tracking de los alumnos", un seguimiento del comportamiento de los estudiantes a lo largo de la carrera.

No contar con un proceso para el análisis de estos datos, dificulta muchas veces la toma de decisiones. Poder identificar los diferentes perfiles de los estudiantes y egresados,

ayudaría a comprender mejor su comportamiento e implementar propuestas educativas que los ayuden a completar sus estudios, evitando así la deserción de la carrera.

Objetivos

El objetivo de este trabajo es, experimentar un proceso de integración de datos de diferentes sistemas basados en software libre, que se utilizan habitualmente en la Facultad, para poder hacer un análisis de los mismos. Aplicar diferentes estrategias, técnicas y algoritmos que ofrecen las herramientas de minería de datos, para obtener información no trivial y detectar la existencia de patrones de comportamiento, en los datos integrados de los alumnos de la Facultad de Informática de la UNLP.
Se propone seguir los pasos del proceso de extracción de conocimiento mencionado anteriormente. Extraer estos datos de diferentes sistemas basados es software libre, utilizados en la Facultad. Preprocesar y transformar estos datos en la medida que sea necesario para luego aplicar diferentes algoritmos de minería de datos evaluando con expertos los resultados que cada uno de ellos arroja. Se propondrá integrar a los datos obtenidos, otros datos de sistemas externos a la Facultad, como es la participación en diferentes redes sociales tan ampliamente utilizadas actualmente por la sociedad.
El grado de aprovechamiento de la información minada por parte del usuario final, depende en gran medida de una correcta visualización y una interfaz amigable de interacción. El trabajo contempla el desarrollo de un aplicativo que facilite al usuario final el uso de estas técnicas y la interpretación de los resultados obtenidos.

Problemática

Al analizar detalladamente las diferentes tareas a realizar, nos encontramos con los siguientes problemas a resolver en las diferentes etapas de este trabajo de grado.

Obtención de los datos. Los datos requeridos para la realización de este trabajo, son privados de la Facultad de Informática de la UNLP, y el acceso a los mismos no será posible sin una autorización exclusiva de las autoridades pertinentes.

Análisis, interpretación y selección de los datos. Las estructuras de las bases de datos con los que se trabajarán son desconocidas, lo que requiere un análisis e interpretación de las mismas, para poder seleccionar correctamente el o los subconjuntos de datos relevantes con los que se trabajará. Teniendo en cuenta también, los problemas que afectan la calidad de los mismos, como por ejemplo datos anómalos, valores nulos, entre otros.

Integración de los datos. Los datos se seleccionarán de diferentes bases de datos, cada fuente de datos puede usar diferentes formatos de registro, diferentes grados de agregación de los datos, diferentes claves primarias, diferentes valores para la

representación de un mismo dato; lo que hay que hacer entonces, es integrar estos datos creando un almacén de datos, conocido como vista minable, para hacerlos accesibles para el análisis y la toma de decisiones.

Tareas o técnicas de minería de datos a utilizar. La selección de las tareas o técnicas correctas de minería de datos a utilizar, es importante para obtener un análisis óptimo de los datos. Para poder tener un análisis confiable, un objetivo es lograr diseñar estos modelos o patrones, con calidad de predicción o clasificación que supere el 60%[1].

Organización del documento

Con el propósito de facilitar la lectura del trabajo; se estructura el mismo en tres partes bien diferenciadas.

Parte I - Investigación teórica – conceptual.

En esta primera parte, se realiza la investigación teórica sobre minería de datos, su relación con otras disciplinas y la aplicación actual en el campo de la educación.
Se encuentra dividida en los siguientes dos capítulos.
Capítulo 1. Introducción – BI, KDD y DM.

En este capítulo se tratan diferentes definiciones y características de BI y DM; está pensado para aclarar los diferentes términos y saber qué lugar ocupa cada uno en el proceso de adquisición de conocimiento.
Capítulo 2. Minería de datos y estilos de aprendizaje.

En este capítulo, se desarrollan más en profundidad los temas relacionados con la minería de datos. Se realiza un desarrollo más detallado de las tareas, técnicas u métodos existentes actualmente para el desarrollo de minería de datos sobre un conjunto de datos.
Capítulo 3. Minería de datos educativa.

La minería de datos aplicada a la educación, comenzó a crecer como área de investigación. En este capítulo se presenta el estado del arte de la minería de datos aplicada a datos relacionados con el sistema educativo, haciendo una investigación sobre las tareas y técnicas de minería de datos más usadas en este campo.

Parte II - Herramientas y Datos.

Esta parte está dedicada a la investigación y descripción de los principales sistemas y herramientas de minería de datos que se encuentran actualmente en el mercado.

[1] Este valor es aceptable en otros trabajos sobre EDM, como [Dekker, Pechenizkiy & Vleeshouwers 2009], [Antunes 2010].

Se menciona también, los diferentes sistemas que se utilizan en la Facultad de Informática de la UNLP, para el almacenamiento de los datos de sus alumnos y egresados.
Se encuentra dividida en los siguientes dos capítulos.
Capítulo 4. Sistemas y Herramientas de minería de datos.

En este capítulo se investigan y describen las principales herramientas de minería de datos presentes en el mercado en la actualidad.
Capítulo 5. Dominios de investigación.

Este capítulo está dedicado a la investigación y descripción de los diferentes dominios de datos que serán la fuente de información para la extracción del conocimiento esperado.

Parte III - Solución propuesta.

Esta tercera y última parte, se detalla la solución propuesta, y el trabajo realizado para alcanzar el objetivo mencionado.
Se encuentra dividida en los siguientes dos capítulos.
Capítulo 6. Presentación del problema.

En este capítulo se presenta el problema a resolver junto con las decisiones tomadas acerca de lo datos utilizados, las tareas y algoritmos de minería de datos aplicados y la herramienta de minería seleccionada para la realización de los modelos.
Capítulo 7. Análisis de los datos.

Este capítulo describe los pasos de selección y pre-procesamiento de los datos necesarios para la realización de este trabajo. Se corresponde con los dos primeros pasos del proceso de KDD.
Capítulo 8. Modelado y Evaluación.

En este capítulo se muestra la aplicación de los diferentes algoritmos de minería de datos seleccionados sobre los datos obtenidos junto con los resultados obtenidos por cada uno.
Capítulo 9. Aplicativo.

Este capítulo es una guía de uso del aplicativo que se presenta en este trabajo de grado; el mismo permite la realización de los modelos presentados en el capítulo 8.
Capítulo 10. Extensión del trabajo.

Se presenta una extensión del presente trabajo, agregando información sobre el uso que los estudiantes hacen de las redes sociales Twitter y Facebook.

PARTE I

INVESTIGACIÓN TEÓRICA – CONCEPTUAL

En esta parte, se introduce al concepto de Business Intelligence, Knowledge Discovery Data Base y Minería de datos. Luego de desarrolla con más detalle los conceptos de minería de datos y por último se realiza una investigación sobre las diferentes aplicaciones de Minería de Datos en el ambiente educativo.

CAPÍTULOS

1. BI, KDD Y DM
2. MINERÍA DE DATOS Y ESTILOS DE APRENDIZAJES
3. MINERÍA DE DATOS EDUCATIVA

Capítulo 1
BI, KDD Y DM

1.1 Datos, Información y conocimiento

Se puede decir que "*Vivimos en la era de los datos*" [White, 2009 capítulo 1]; la explosión de datos, comenzó a suceder a todo nivel en los dispositivos electrónicos. Aplicaciones, individuos y organizaciones; guardan datos a un ritmo que crece de manera sorprendente. Los datos almacenados crecen tanto en el número de registros u objetos almacenados, como en el número de atributos ó propiedades. Las bases de datos con grandes volúmenes de registros, se volvieron cada vez más comunes, provocando que la cantidad de datos disponibles sea mucho mayor de los que se puedan analizar.

Actualmente nos encontramos en una transición hacia "*La era del conocimiento*" [Laurinen 2006; Maimon, Rokach 2010], los datos almacenados contienen la información útil y necesaria para la toma de decisiones, pero muchas veces acceder a esos datos para obtener información de ellos no es una tarea trivial; no solamente por la gran cantidad de datos almacenados sino también porque muchas veces los datos son almacenados en diferentes áreas y en diferentes dispositivos de almacenamiento.

Comienza a existir una brecha entre la cantidad de datos y la comprensión de ellos. A medida que la cantidad de datos aumenta, disminuye la proporción de ellos que podemos comprender. Oculto en todos estos datos hay información, potencialmente útil, que muchas veces no se hace explícita o aprovechable.
La necesidad de obtener información útil de los datos almacenados en grandes almacenes de datos, generó investigaciones en el área de informática, para la extracción de conocimientos de los datos almacenados en dispositivos electrónicos. De estas investigaciones, surgen nuevos términos y definiciones como Business Intelligence, Knowledge Discovery in Database, Data Warehouse, Data Mart, Analítica de Datos y Data Mining.
El uso de estos términos y otros relacionados, se escuchan cada vez con más frecuencia, provocando en muchos casos confusiones sobre qué se refiere cada uno.
Con el objetivo de aclarar ciertas confusiones; se presenta a continuación, una introducción sobre los principales términos relacionados con la obtención de conocimiento a partir de datos digitalizados.

1.2 Business Intelligence

El término Inteligencia de Negocios – BI por sus siglas del inglés Business Intelligence – engloba el uso de procesos y la aplicación de herramientas informáticas, con el objetivo de extraer información útil y tal vez oculta, de grandes cantidades de datos almacenados en diferentes dispositivos. A continuación se presentan algunas definiciones.

[bi-argentina] define BI como "*la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios*" y al asociarlo con las tecnologías de la información, lo define como "*el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la organización) en información estructurada, para su explotación directa (reporting, análisis OLTP / OLAP, alertas…) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.*"

[Gartner Group IT Glossary] por su lado, dice que BI "*es un término genérico que incluye las aplicaciones, la infraestructura, las herramientas y las mejores prácticas que permitan el acceso y análisis de la información para mejorar y optimizar las decisiones y el rendimiento*".

[Howson, 2007] dice sobre el BI que "*es un conjunto de tecnologías y procedimientos que permite a las personas de los diferentes niveles de una organización acceder a los datos, analizarlos e interpretarlos*".

[Schepes, 2008] define a la inteligencia de negocio como "*cualquier actividad, herramienta o proceso usados para obtener la mejor información que ayuden al proceso de la toma de decisiones*".

[Barbenabeu, 2010] *menciona una frase muy popular acerca de BI, que dice: "Inteligencia de Negocios es el proceso de convertir datos en conocimiento y el conocimiento en acción, para la toma de decisiones".*

Las definiciones anteriores, coinciden en que BI se trata de un grupo de tecnologías usadas en conjunto, para recolectar y utilizar efectivamente la información, con el objetivo de obtener conocimiento de los datos y mejorar las operaciones de una organización; ya que al contar con la información exacta y en tiempo real, es posible, identificar y corregir situaciones que se podrían convertir en problemas, pudiendo conseguir nuevas oportunidades o readaptarse frente a nuevos sucesos.

Las tecnologías o componentes del BI se diferencian de los sistemas operacionales en que están optimizadas para preguntar y dar conocimiento sobre los datos, ejecutando consultas de alto rendimiento. Estas herramientas tienen que garantizar el acceso de los usuarios a los datos, con independencia de la procedencia de estos y buscan ofrecer una presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que le permitan seleccionar y manipular sólo los datos de interés.

La Figura 1.1 – Tecnologías asociadas al proceso de inteligencia de negocios; muestra esta relación de tecnologías.

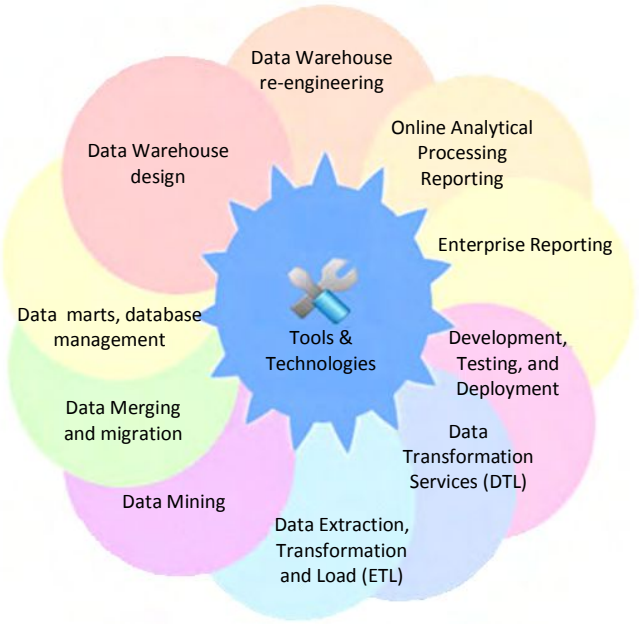


Figura 1.1 – Tecnologías asociadas al proceso de inteligencia de negocios.

- **Data Management**: Generalmente, los datos de los sistemas operacionales, necesitan transformaciones para que las herramientas de explotación se apliquen correctamente. Esta disciplina, se encarga del diseño, optimización y mantenimiento de estructuras de datos, tanto operacionales como analíticos.
- **Data Warehouse:** Es un entorno orientado a la consulta de datos, separado del operacional, que se nutre de la información de éste y otras fuentes externas, con el fin de utilizar dicha información como fuente única para la obtención de conocimiento.
- **Data Mart:** Se trata de un subconjunto de un Data Warehouse especializado en el almacenamiento de datos de un área específica de la organización.
- **OLAP:** Es una disciplina de generación de conocimiento deductivo o para obtención de información como verificación de hipótesis. Son herramientas que presentan los datos al usuario con una visión multidimensional, de manera rápida e interactiva.
- **Knowledge Discovery in Databases - KDD:** Es una disciplina de generación de conocimiento a partir de los propios datos. Es el proceso no trivial de extracción de

información implícita, previamente desconocida y potencialmente útil de los datos existentes en una base de datos.

- **Data Mining:** Es un paso en el proceso de KDD que consiste en la aplicación algoritmos de descubrimiento, con el objetivo de descubrir patrones y modelos entre los datos.

1.3 BI y Analítica de Datos

Entre las funciones que proveen las tecnologías relacionadas al BI, se encuentran, entre otras, presentación de informes (reporting), minería de datos, minería de textos y analítica de datos.

La analítica de datos (DA, por sus siglas del inglés Data Analitycs) [SDM, AllAnalitics, online-behavior, technopedia], se define como la ciencia de examinar los datos con el fin de sacar conclusiones sobre los mismos. Se refiere a las técnicas y procesos utilizados para medir, recopilar, analizar y presentar los datos a los efectos de entender y mejorar la productividad y ganancia. Es usada en muchas industrias para permitir a las compañías y organizaciones a tomar mejores decisiones de negocio y en la ciencia para verificar o refutar modelos y teorías existentes.

La analítica de datos generalmente se divide en dos grandes ramas: El análisis de exploración datos (EDA – Exploratoty Data Analysis), que se encarga del descubrimiento de nuevas características de los datos y el análisis de confirmación de datos (CAD – Confirmatory Data Analysis) que se encarga de las pruebas de las hipótesis existentes.

Se distingue de la minería de datos, hacemos esta comparación debido a que este trabajo de grado se centra en la investigación y aplicación de herramientas de minería de datos, en el alcance, propósito y el enfoque del análisis. En la minería de datos, se utiliza software sofisticado para identificar y establecer patrones ocultos de grandes cantidades de datos. La analítica de datos se centra en la inferencia, el proceso de obtención de una conclusión, basado exclusivamente en lo que ya es conocido por el investigador. Se puede ver a la minería de datos como una forma de analítica con una intención más exploratoria (o quizás expedicionaria), realizada generalmente en grandes cantidades de datos.

1.4 Knowledge Discovery in Database - KDD

El descubrimiento de conocimiento en bases de datos – o KDD, por sus siglas del inglés Knowledge Discovery in Database –, surge como una nueva rama de investigación en el campo de informática, a partir de la necesidad herramientas capaces de automatizar la extracción de información útil, del creciente volumen de datos almacenados en grandes repositorios. Algunas definiciones de KDD son las siguientes.

[Fayyad et al. 1996] lo define como "*el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles en los datos*".

[Friedman 1997] considera al proceso de KDD como un "*análisis automático y exploratorio de datos de grandes bases de datos*".

[Hand 1998] lo ve como un "*análisis secundario de datos de grandes bases de datos*". El término "secundario" hace hincapié en el hecho de que el propósito principal de la base de datos no fue el análisis de datos.

[Frawley et.al, 1992] dice que KDD es "*la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de datos*".

[Kopanakis et.al, 2003] define al término KDD como "*El proceso de descubrimiento de conocimiento sobre repositorios de datos complejos mediante la extracción oculta y potencialmente útil en forma de patrones globales y relaciones estructurales implícitas entre datos*".

Otras definición es la de [Witten et. al 2011] que dicen que KDD "*Es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos*".

[Maimon & Rokach 2010] dicen que KDD "Es un modelado y análisis automático y exploratorio de grandes repositorios de datos. KDD es el proceso organizado para la identificación de patrones válidos, novedosos, útiles y comprensibles a partir de conjuntos de datos grandes y complejos".

Por su parte, [Pazzani et.al, 1997], relacionan al KDD con otras áreas y lo definen como "*un campo cuyo objetivo es extraer conocimiento útil a partir de una colección de datos. Se basa en los métodos de la estadística, reconocimiento de patrones, la teoría de la información, aprendizaje de máquinas y redes neuronales para producir modelos que permiten conocer los datos.*"

En el contexto de KDD, se denomina *conocimiento*, a un patrón que tiene un cierto grado de certeza, es comprensible y es lo suficientemente interesante (útil y novedoso) para el usuario final. Sólo los patrones interesantes son conocimiento. Un patrón es interesante si es nuevo, útil y no trivial. [Frawley et.al, 1992]

1.4.1 Pasos del proceso de KDD

Investigadores, tales como [Brachman & Anand, 1994], [Fayyad et al., 1996], [Maimon & Last, 2000] y [Reinartz, 2002] proponen diferentes maneras de dividir el proceso de KDD

en fases o etapas iterativas. Adoptando un híbrido de estas propuestas, presentamos el proceso de KDD dividido en 5 etapas. Estas etapas se pueden observar en la Figura 1.2 – Pasos del proceso de KDD.

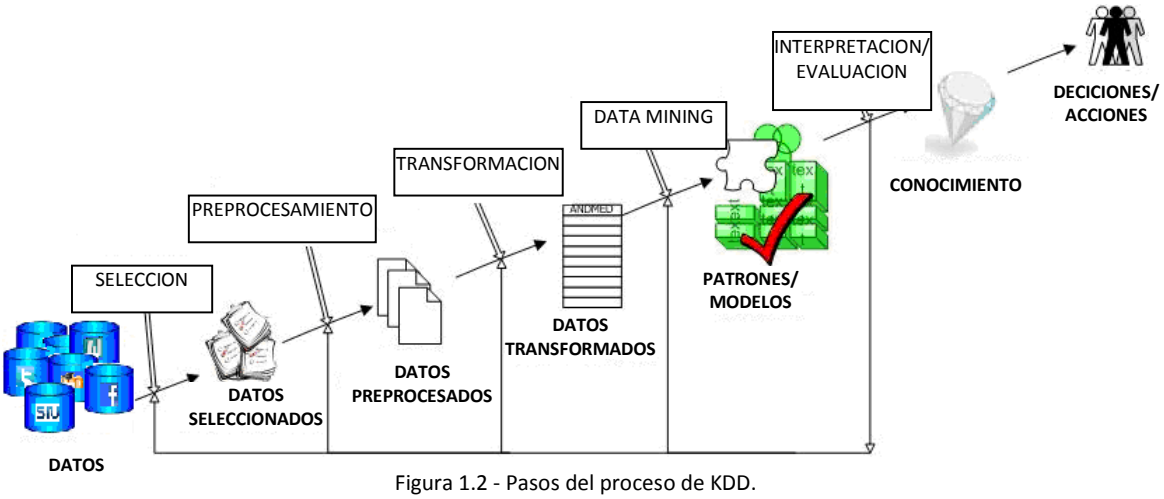


Figura 1.2 - Pasos del proceso de KDD.

- **Selección**. En esta etapa, se crea el conjunto de datos del cual se obtendrá el conocimiento. Se buscan los datos disponibles, datos adicionales, se integran todos los datos, se determinan los atributos que serán necesarios en el proceso.
- **Preprocesamiento.** Esta etapa, consiste en hacer una limpieza de los datos seleccionados; las operaciones básicas son eliminación del ruido, manejo de datos nulos y atípicos, entre otros.
- **Transformación**. Esta etapa tiene como objetivo mejorar los datos para la aplicación de los métodos de minería de datos. Se trata de reducción y proyección de los datos.

El resultado de aplicar estas primeras etapas del proceso de KDD, es la creación de una vista minable, un archivo de datos –tabla de base de datos o archivo específico –, que contiene todos los atributos relevantes para el proceso de minería de datos.

Figura 1.3 – Creación de una vista minable.

- **Minería de Datos.** En esta etapa se decide la tarea de minería de datos que se aplicará; se selecciona el método ó técnica específico que se usará para la realización de la tarea y por último se aplica el algoritmo seleccionado sobre el conjunto de datos para la obtención de los resultados.
- **Evaluación.** En esta etapa, se visualizan y evalúan los resultados obtenidos.

El paso de Minería de Datos en el proceso de KDD, es considerado el más importante [Maimon & Rokach 2010; Fayyad, 1997], tanto que muchas veces se usa el término de Minería de Datos como sinónimo del proceso completo de KDD.
Teniendo en cuenta que minería de datos es un tema principal de este trabajo de grado, y la importancia en el proceso de KDD, se dedicará un capítulo para describir este tema. En el capítulo siguiente, se detallan las diferentes tareas, técnicas o métodos y algoritmos para la resolución de problemas de minería de datos.

Capítulo 2
MINERÍA DE DATOS Y ESTILOS DE APRENDIZAJE

En este capítulo se describe la disciplina de minería de datos. Los conceptos y procedimientos descriptos en este capítulo, se utilizarán luego para la realización de la parte práctica de este trabajo de grado.

2.1 Minería de Datos - DM

"*Los datos históricos ayudan a explicar el pasado para conocer el presente y predecir información futura*". Los datos almacenados, generalmente corresponden a datos históricos. La toma de decisiones se basa también en estos datos. La extracción de conocimiento manual se vuelve impracticable si la cantidad de datos es muy alta.
La minería de datos – DM por sus siglas en inglés de Data Mining –, es un conjunto de herramientas y técnicas para soportar la extracción de conocimiento útil de los datos. Se pueden destacar las siguientes definiciones:

[Gartner Group IT Glossary] dice que "*es el proceso de descubrir nuevas correlaciones significativas, patrones y tendencias a través de la exploración de grandes cantidades de datos almacenados en repositorios, usando tecnología de reconocimiento de patrones así como técnicas estadísticas y matemáticas*".

[Hand et. al. 2001] indican que "*es el análisis de –generalmente grandes – conjuntos de datos observados para encontrar las relaciones insospechadas y resumir los datos de una manera novedosa para que sean tanto entendibles como útiles para el dueño de los mismos*".

Evangelos Simoudis [Cabena et al., 1998], la define como "*un campo interdisciplinario que reúne las técnicas de aprendizaje automático, reconocimiento de patrones, estadísticas, bases de datos y visualización para abordar el tema de la extracción de información de grandes bases de datos*".

[Han et. al., 2012] adopta una visión amplia de la funcionalidad y la definen como "*el proceso de descubrimiento de patrones interesantes y conocimiento de grandes*

cantidades de datos. Las fuentes de datos pueden incluir bases de datos, data Warehouse, la WEB, otros repositorios de información, o datos que se transmiten en el sistema dinámicamente".

Al igual que la BI, la Minería de Datos es un campo multidisciplinar. Algunas de las disciplinas con las que se relaciona la minería de datos se pueden observar en la Figura 2.1 – Disciplinas relacionadas a la Minería de Datos.

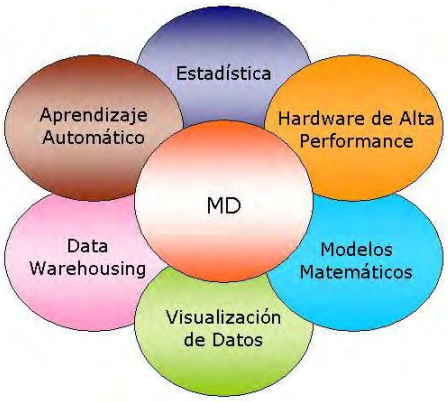


Figura 2.1 – Disciplinas relacionadas a la Minería de Datos.

2.2 Clasificación de minería de datos

Diferentes autores mencionan distintas clasificaciones de minería de datos, con diferencias no muy amplias [Han, et al 2012; Maimon & Rokach 2010, capítulo 1]. Se opta por la clasificación mostrada en la Figura 2.2 – Clasificación de Minería de Datos; ya que es la que expone, con claridad, la separación entra las diferentes tareas y técnicas de minería de datos.

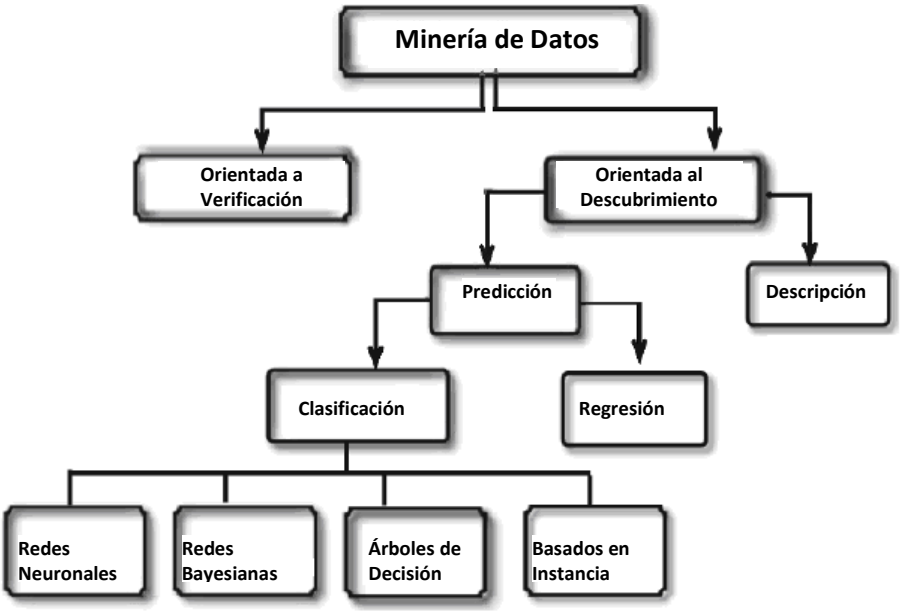


Figura 2.2 - Clasificación de Minería de Datos.

Se pueden destacar dos clasificaciones principales: la minería de datos orientada a la verificación y la minería de datos orientada al descubrimiento.

- **Minería de datos orientada a la verificación**: el sistema evalúa una hipótesis propuesta. Muchas veces, la verificación no se asocia con la minería de datos, ya que la mayoría de los problemas de minería de datos, se enfocan en el descubrimiento de una hipótesis en vez de testear una ya existente.
- **Minería de datos orientada al descubrimiento**. El sistema encuentra nuevas reglas y patrones en los datos de manera automática. Se subdivide en minería de datos predictiva y minería de datos descriptiva.
    - **Minería de datos descriptiva.** Está orientada a la interpretación de los datos, se enfoca en la comprensión de la relación entre ellos. Busca patrones que expliquen o resuman los datos; exploran las propiedades de los datos ya existentes para presentarlos al usuario de una manera más comprensible, no predicen nuevos datos.
    - **Minería de datos predictiva.** Se construye un modelo de comportamiento mediante el uso de datos ya conocidos. Este modelo, tiene que ser capaz de predecir el comportamiento futuro de datos que no se usaron para la construcción del modelo. Estas predicciones es lo que se denomina conocimiento.

2.3 Pasos para el desarrollo de la fase de minería de datos

El proceso de minería de datos, representado en la Figura 2.2 – Pasos del proceso de minería de datos, consta de tres decisiones principales, a tener en cuenta en el momento de comenzar la extracción de conocimiento de un conjunto de datos.

Figura 2.2 – Pasos del proceso de minería de datos.

- Determinar la tarea de minería de datos más apropiada para la resolución del problema. Por ejemplo si el problema es determinar qué clientes pueden retirarse como tales de una compañía, se podría usar una tarea de clasificación.
- Una vez seleccionado el tipo de tarea, se debe seleccionar el método o la técnica de minería de datos que resuelva dicha tarea. Por ejemplo, para una tarea de clasificación, se puede usar la técnica de árbol de decisión.
- Por último, se elige el algoritmo de minería de datos que resuelva la tarea y obtenga el modelo que está buscando. Existen muchos métodos o algoritmos para la construcción de un mismo modelo. Por ejemplo para construir de un árbol de decisión, se pueden usar los algoritmos CART, C4.5, C5.0, entre otros.

En el problema: “Determinar el aprendizaje de los alumnos de una asignatura, como óptimo, bueno, regular o malo".
El tipo de tarea es la clasificación; y para resolverla se puede seleccionar el método de árboles de decisión utilizando el algoritmo C4.5.

Generalmente, se selecciona más de una técnica para la resolución de una misma tarea, y luego se comparan los resultados obtenidos para determinar cuál se adapta mejor a la solución esperada.

2.4 Tareas de minería de datos

Una tarea de minería de datos, es un problema a ser resuelto por un algoritmo de minería de datos. Las tareas que pueden realizarse a través de las distintas técnicas, tienen la siguiente clasificación:

- **Descripción**. Las técnicas que resuelven este tipo de tareas, buscan derivar descripciones concisas de características de los datos. De esta manera, es posible disponer de un modelo de la información existente. Se subclasifican en agrupamiento y asociación.
  - **Agrupamiento o Segmentación**. Las técnicas que resuelven la tarea de agrupamiento, dividen los datos en grupos con características similares, según un criterio de comparación entre los valores de los atributos de las instancias. El objetivo es maximizar la similitud de las instancias pertenecientes a un mismo grupo y minimizar la similitud con las instancias que se encuentran fuera del mismo.
  - **Asociación o correlación**. Identifica relaciones no explícitas entre los atributos. No siempre implica una relación causa-efecto. Por ejemplo, se puede determinar que los alumnos que desaprueban la materia X, entonces también desaprueban la materia Y.
- **Predicción**. La meta es inducir un modelo para poder predecir o estimar, dados los valores de los atributos, el valor de salida correspondiente. Por ejemplo, podría predecirse el resultado de las calificaciones de un alumno, en base a lo observado en otros alumnos con características similares. Se subclasifican en clasificación y regresión.
  - **Clasificación**. Trata de encontrar las fronteras que mejor dividen a los datos en las diferentes clases. Permite predecir la categoría o clase de nuevas instancias, en función de una serie de atributos. Por ejemplo se puede clasificar a un alumno en “Aprobado” o “Desaprobado” dependiendo de los datos académicos que se tengan del mismo a lo largo de la cursada de una materia. Un algoritmo clasificador se “entrena” y arma un modelo a partir de un conjunto de datos conocidos –se conoce la clase de los mismos –. Este modelo se usa luego para predecir la clase de datos desconocidos. El objetivo es conseguir un modelo lo más preciso posible. La precisión se calcula entre la cantidad de instancias predichas correctamente sobre el total de predicciones.

- **Regresión o estimación**. Difiere a la clasificación, en que predice un valor numérico. Consiste en aprender una función, que asigna a cada instancia un valor real. El objetivo es minimizar el error entre el valor predicho y el valor real. Se puede utilizar, por ejemplo, para predecir las respuestas de alumnos, en base a la actuación del alumno en evaluaciones previas, a la complejidad de las preguntas y al contenido de su aprendizaje.

2.5 Métodos ó técnicas de minería de datos

Cada una de las tareas mencionadas anteriormente, requiere de la aplicación de métodos ó técnicas y algoritmos de minería de datos para ser resueltas.
Una técnica o método constituye el enfoque conceptual para extraer la información de los datos. Es preciso un entendimiento de alto nivel de los algoritmos para saber cual es la técnica más apropiada para cada problema. Se debe entender los parámetros y las características de los algoritmos para preparar los datos a analizar.
Una tarea puede tener muchos métodos diferentes que la resuelvan y el mismo método, puede utilizarse para resolver más de una tarea.
A continuación se listan los métodos más usados en el campo de la minería de datos junto con los algoritmos más populares que las resuelven.

- **Agrupamiento basado partición.** Es el método por excelencia para la resolución de problemas de agrupamiento o clustering. Determina el comportamiento de una nueva instancia dependiendo del comportamiento de las instancias anteriores "vecinas". Individuos similares –o cercanos –, deben pertenecer al mismo grupo, normalmente, se calcula la distancia euclediana para obtener la cercanía.

El algoritmo más popular que resuelve esta técnica es el K medias –ó k-means – [MacQueen 1967]. Sitúa en el espacio, un número prefijado de centros; luego, cada ejemplo es comparado con estos centros y asociado a aquel que sea más próximo; k es el valor inicial de centros de grupos. Una vez seleccionado el valor inicial k, el algoritmo calcula, para cada una de las instancias, el centro más cercano. Al finalizar, cada centro tendrá un conjunto de ejemplos que representa. Luego se calcula un nuevo centro para cada uno de los grupos, y se mueve el centro original al nuevo calculado. Se repite el procedimiento hasta que no haya ningún movimiento de centros en una iteración.
Este algoritmo depende de la inicialización aleatoria de los centros, que puede provocar que no se obtenga un agrupamiento óptimo, e influir en la cantidad de iteraciones necesarias para alcanzar la solución. Una alternativa es el algoritmo k-means++ [Arthur & Vassilvitskii 2007], el cual se basa en mejorar la inicialización de los centros aunque mantiene una parte de aleatoriedad.

- **Agrupamiento jerárquico.** Este método resuelve el problema de definir a priori, cuántos grupos puede haber en los datos. Construye un árbol denominado dendrograma, donde la raíz contiene el conjunto completo de datos, las hojas son nodos de datos individuales y cada uno de los nodos internos, son subconjuntos de los datos. Los diferentes algoritmos que implementan este método, difieren de la manera en que se calcula la distancia de enlace entre los diferentes nodos del árbol o grupos.
Con el árbol completo, se puede seleccionar el nivel en el que la diferencia entre los grupos sea más clara.
Entre los algoritmos más populares que implementa la metodología de agrupamiento jerárquico se encuentran el COBWEB [Fisher 1987] usado generalmente para atributos nominales y el Classit [Gennari et al. 1990] que trabaja con atributos numéricos.

- **Reglas de Asociación:** Las reglas de asociación determinan correlaciones entre los atributos de un conjunto de datos. Una regla de asociación tiene la forma "*Si X ENTONCES Y*". Las medidas de soporte y confianza, permiten estimar su calidad.
El Soporte o Cobertura, es el número de instancias que la regla predice correctamente y se expresa en proporción a la cantidad de ejemplos totales disponibles.
La Confianza o Precisión, es el porcentaje de veces que la regla se cumple cuando se puede aplicar. Se mide sobre el subconjunto de instancias para el cual la regla se puede aplicar.

> Supongamos que la regla "Si los estudiantes obtienen 'incorrecto en X' ENTONCES también obtienen 'incorrecto en Y'; posee 40% de soporte y una confianza del 66%.
> La regla predice correctamente al 40% del total de los alumnos y además, el 66% de los estudiantes que obtienen "incorrecto en X" también obtienen "incorrecto en Y", es decir que el 66% de las veces que la regla se puede aplicar, se aplica correctamente.

Un algoritmo para la obtención de reglas de asociación es el algoritmo "A Priori" [Agrawal, Imielinski & Swami 1993]. Se caracteriza por encontrar los ítems frecuentes, sobre los cuales después generará las reglas.

Un caso especial de reglas de asociación, son las reglas de asociación secuenciales; se utilizan para determinar patrones secuenciales en los datos, que se dan en instantes cercanos en el tiempo. Las relaciones entre los datos se basan en el tiempo. El objetivo es encontrar relaciones del tipo "El 40% de los alumnos que acceden a buscar información en la página WEB X, visitarán en los próximos dos días la página WEB Y". Este tipo de aprendizaje se basa en encontrar las secuencias más comunes; una secuencia se define como una lista de ítems de un mismo cliente ordenada por el tiempo. Los algoritmos más populares para el aprendizaje de estas reglas, son el algoritmo AprioriAll [Agrawall & Srikant 1995] que cuenta básicamente con una fase de ordenación de los datos antes de la

obtención del conjunto de ítems más frecuentes y la generación de las reglas. El algoritmo Generalized Sequential Patterns (GSP) [Agrawall & Srikant 1996] que permite considerar restricciones temporales, es decir, secuencias de una determinada duración.

- **Árboles de Decisión o de Clasificación.** Es la técnica de resolución de la clasificación por excelencia. Conforman una serie de condiciones organizadas en forma jerárquica a modo de árbol, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Los nodos internos del árbol contienen una condición sobre un atributo en particular y los nodos hojas contienen el valor de una clase con la cual clasifican todas las instancias que llegan a la hoja. Para clasificar una instancia nueva, se recorre el árbol desde la raíz evaluando las condiciones de cada nodo y siguiendo la rama del nodo determinada por los resultados de cada una de ellas, cuando se llega a una hoja, la instancia se clasifica con el valor de clase que la hoja contenga.

En la Figura 2.3 – Ejemplo de árbol de clasificación, se muestra un ejemplo de un árbol de decisión. Esta estructura permite clasificar los datos según el atributo "Resultado Primer Parcial" en "Aprobado" = "A" y "No Aprobado" = "NA". La clase se deduce de los atributos "Primer T.P." que puede tomar los valores "Bien", "Muy Bien" y "Regular"; "Actividad Virtual" y "Actividad Presencial" pueden tomar los valores "Alta" o "Baja".

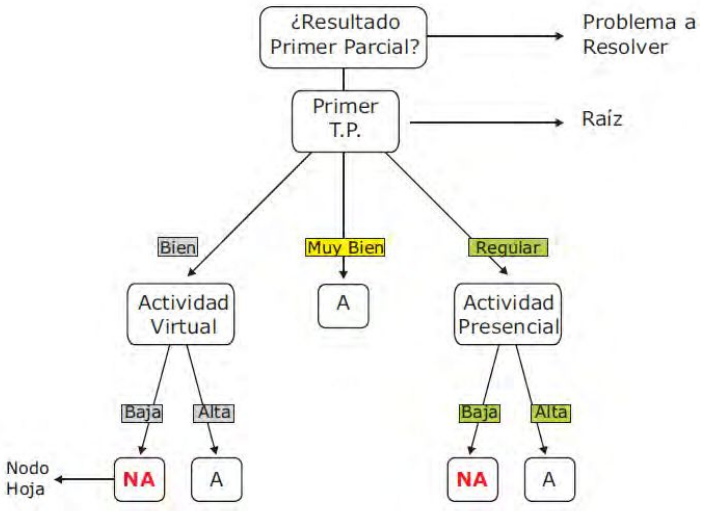


Figura 2.3 – Ejemplo de árbol de clasificación.

La mayoría de los algoritmos de árboles de decisión se basan en el enfoque de divide y vencerás, trabajan construyendo el árbol desde la raíz hasta las hojas, buscando en cada etapa cual es el "mejor" atributo para realizar la división y separar las clases. Los diferentes algoritmos de formación de árboles de decisión usan diferentes estrategias se selección del "mejor" atributo para crear la condición en un nodo dado.

También estos algoritmos se caracterizan por realizar una poda de la estructura con el objetivo de reducir su tamaño y complejidad, preservando una cota de precisión aceptable en la clasificación.

Entre los algoritmos o sistemas de aprendizajes de árboles de decisión más populares se encuentran:
ID3 [Quinlan 1986], y su extensión, el C4.5 [Quinlan 1993]. Son dos de los algoritmos más usados. Son métodos de "divide y vencerás", están basados en criterios de partición derivados de la ganancia (GainRatio). Tienen poda basados en reglas.
CART [Breiman et. al. 1984]. Son métodos de "divide y vencerás", se basan en el criterio de partición GINI y sirve tanto para clasificación como para regresión. La poda se basa en una estimación de la complejidad del error ("error-complexity"). Se pueden encontrar con el nombre C&RT.
IND, LMDT. Son sistemas híbridos, incorporan características de varios sistemas o añaden otras técnicas de aprendizaje en la construcción de árboles de decisión.
SLIQ y SPRINT. Modificaciones de árboles de decisión clásicos para conseguir la escalabilidad para grandes volúmenes de datos.

- **Redes bayesianas.** Las redes bayesianas se usan principalmente para resolver tareas de clasificación –aunque también suelen usarse con fines descriptivos –. Unas de las principales diferencias sobre las otras técnicas de clasificación es que permite calcular de forma explícita la probabilidad asociada a cada una de las hipótesis –o clases – posibles. Por este motivo también son denominados clasificadores probabilísticos.

> En la tarea "Predecir el desempeño –si aprobará o no – de un estudiante en un curso, antes de que el curso termine.". Puede ser más informativo, conocer la probabilidad que el estudiante tiene de aprobar o no.

Uno de los algoritmos para construcción de redes bayesianas es el Naïve Bayes [Duda & Hart 1973; Langley et. al. 1992], cuyo fundamento principal es la de suponer que todos los atributos son independientes una vez conocido el valor de la variable clase.
También se encuentran los algoritmos TAN y BAN [Friedman et. al 1997] que son extensiones del anterior; en el primero se supone que los atributos forman una red bayesiana con forma de árbol, con el objetivo de intentar mejorar la taza de aciertos durante la clasificación. En el segundo, primero se aprende una red bayesiana para los atributos y después se aumenta con la variable clase.

2.6 Evaluación de los modelos

Una vez construido el modelo de minería de datos, se necesita saber si el mismo es lo suficientemente válidos para el objetivo que se está buscando. Para este punto generalmente se usa un conjunto de datos diferente al usado en la parte de generación del modelo, que se denomina conjunto de test. Dependiendo de la tarea de minería de datos, existen diferentes medidas de evaluación de los modelos obtenidos.

En las tareas de clasificación, generalmente se evalúa la calidad del modelo con respecto a su *precisión*; la cual se calcula como el número de instancias del conjunto de test clasificadas correctamente dividido el número total de instancias del conjunto de prueba. El objetivo es obtener la mayor precisión posible sobre el conjunto de test.
En la tarea regresión, se utiliza el *error cuadrático medio* del valor predicho respecto al valor que indica la instancia del conjunto de test. Se promedia los errores.

En el caso de las tareas de reglas de asociación, se suele evaluar de forma separada cada una de las reglas, en base la *cobertura* y *confianza*, ejemplificados en el punto "**Reglas de Asociación**".

Para el agrupamiento, las medidas de evaluación suelen depender del método utilizado, aunque suelen ser fusión de la *cohesión* de cada grupo y de la *separación* entre grupos. La cohesión y separación entre grupos se puede formalizar, por ejemplo, utilizando la distancia media al centro del grupo de los miembros de un grupo y la distancia media entre grupos, respectivamente.

2.6.1 Técnicas de evaluación de modelos

Para evaluar un modelo, se debe utilizar un conjunto de datos diferente con el cual se generó el mismo para evitar un sobre ajuste y que el modelo sea lo más general posible.

- **Validación simple.** Es el método de validación más sencillo. Se reserva un porcentaje de datos para la generación de las pruebas, estos datos no son utilizados para la generación del modelo. Hay que tener en cuenta de reservar un conjunto de datos lo más aleatorio posible para que las pruebas sean correctas.
- **Validación cruzada.** Es el método más utilizado. Consiste en dividir el conjunto de datos en k grupos; uno se reserva para la prueba y se construye el modelo con la unión de los demás k – 1 grupos. Este proceso se repite k veces, dejando en cada iteración un conjunto diferente para la prueba. De esta manera se obtienen k porcentajes de errores. El modelo final queda definido con la totalidad de datos y su error es calculado con el promedio de los k errores diferentes obtenidos.

2.7 Selección de una Técnica de Minería de Datos

La correspondencia entre tareas y técnicas es variada. Algunas técnicas permiten resolver distintas tareas y otras permiten resolver solamente un tipo de tarea.
En la siguiente tabla se muestra la relación entre tareas y técnicas/algoritmos que se pueden usar para resolverlas:

Técnicas/Algoritmos	PREDICTIVO		DESCRIPTIVO	
	Clasificación	Regresión	Agrupamiento	Reglas de Asociación
Técnica de Redes Neuronales	✓	✓	✓	
Técnica de Árboles de decisión, Algoritmos ID3, C4.5, C5.0	✓			
Árboles de decisión CART	✓	✓		
Otras técnicas de árboles de decisión	✓	✓	✓	✓
Algoritmo A priori				✓
Algoritmo Naive Bayes	✓			
Técnica de Vecinos más próximos	✓	✓	✓	
Técnica de Máquinas de vectores soporte	✓	✓	✓	
Algoritmos genéticos y evolutivos	✓	✓	✓	✓
Algoritmo CN2 rules (cobertura)	✓			✓

Capítulo 3

MINERÍA DE DATOS EDUCATIVA - EDM

Desde hace unos años, se comenzó a dar lugar una nueva disciplina, dedicada a la extracción de conocimiento de tipos particulares de datos que provienen de los centros educativos, surgiendo así la minería de datos educativa. En este capítulo, se describirá el estado del arte de esta área.

El objetivo de este capítulo es investigar el estado del arte de la aplicación de minería de datos en el universo de datos de educación. Determinar los datos analizados y los modelos de minería de datos que mejor se adaptan en cada caso.

Si bien la minería de datos se aplica en los diferentes niveles del sistema educativo [Márquez-Vera et al. 2011], nos enfocaremos las investigaciones que se hayan realizado sobre los datos en el nivel universitario; debido a que la minería de datos a realizar en el presente trabajo, se basará en datos de alumnos de este nivel.

3.1 Introducción

En [IEDMS 2012], se define a la minería de datos educativa – EDM, por sus siglas del inglés Educational Data Mining –, como "*una disciplina emergente que se ocupa del desarrollo de métodos para la exploración de tipos particulares de datos que provienen de los centros educativos, y el uso de esos métodos con el objetivo de entender mejor a los estudiantes y las metodologías de estudios.*"

[Calders, Pechenizkiy 2011], dicen que *"es un área de investigación multidisciplinaria emergente, en la cual se han desarrollado métodos y técnicas para la exploración de datos originados de varios sistemas de información educativos. La minería de datos educativa es tanto una ciencia de aprendizaje como un área de aplicación rica para la minería de datos, debido a la creciente disponibilidad de datos provenientes del sector educativo. Contribuye al estudio de cómo aprenden los estudiantes y los contextos en que lo hacen. Permite la toma de decisiones para la mejora de las prácticas educativas actuales y el material de estudio.*

Como área multidisciplinaria, reúne a profesionales e investigadores de informática, educación, psicología y estadística, entre otras con el objetivo de descubrir información útil a partir de la gran cantidad de datos electrónicos obtenidos por los diversos sistemas educativos."

3.2 Ciclo y participantes de EDM

Actualmente se utilizan diversos sistemas de información que acompañan a los procesos educativos. En el ámbito de la universidad, los sistemas de información administrativa recopilan principalmente datos acerca de los estudiantes, su inscripción a planes de estudios y materias particulares, y las calificaciones de los exámenes; aunque también suelen estar disponible información sobre las clases, docentes, programas de estudio y cursos, entre otros datos. Si se habla de un curso o materia individual, se pueden considerar datos referentes al cumplimiento de tareas e inscripciones en los exámenes. El aumento del uso de los entornos de aprendizaje interactivos, sistemas de gestión de aprendizaje (LMS – Learning Management System), sistemas tutoriales inteligentes (ITS – Intelligent Tutoring System), y sistemas de hipermedia educativos (e-learning), ha permitido, obtener de grandes cantidades de datos en diferentes niveles de granularidad, como por ejemplo la interacción entre alumnos y docentes, el material de estudio consultado, participaciones en foros.
Los sistemas de bibliotecas también están siendo ampliamente usados en el sistema universitario, y almacenan información de estudiantes y docentes con respecto a los préstamos realizados y el material de bibliografía consultado de las diferentes materias y actividades que realizan en el transcurso de la carrera.
Estos datos pueden ser analizados desde diferentes niveles y perspectivas, mostrando de esta forma diversos aspectos y dando una visión más clara del sistema educativo en general [Trcka et al. 2011; Pechenizkiy et. al 2012; Calders & Pechenizkiy 2011].

Los datos almacenados en los diferentes sistemas, contienen información útil para los miembros del sistema educativo; que puede ser costosa de obtener manualmente debido a su volumen. Esta situación hace que sea necesario el uso de herramientas informáticas que ayuden a la obtención de la información que se desea.
Las técnicas tradicionales de minería de datos han sido ampliamente aplicadas con el objetivo de encontrar patrones interesantes, construir modelos descriptivos y predictivos de grandes volúmenes de datos acumulados. Los resultados de la minería de datos pueden ser utilizados también para conseguir una mejor comprensión de los procesos educativos, para generar recomendaciones y consejos a los estudiantes, para mejorar el manejo de los recursos, por ejemplo el rediseño de algún curso; como evidencia para la aplicación de nuevas evaluaciones y líneas de comunicación entre autoridades y estudiantes. [Trcka et al. 2011, Pechenizkiy et. al, 2012].

[Romero & Ventura 2007; Calders & Pechenizkiy 2011] proponen un ciclo iterativo para la aplicación de minería de datos en los sistemas educativos que se muestra en la Figura 3.1 – Ciclo de minería de datos en el sistema educativo. Los miembros de este ciclo son los responsables administrativos, docentes y alumnos; los dos primeros grupos son los

responsables de diseñar, planificar, construir y mantener los sistemas educativos, mientras que los estudiantes son los principales usuarios y los que interactúan con los sistemas, teniendo en cuenta que tanto docentes como responsables administrativos y académicos también son usuarios de estos sistemas, obteniendo diferentes beneficios de los mismos.

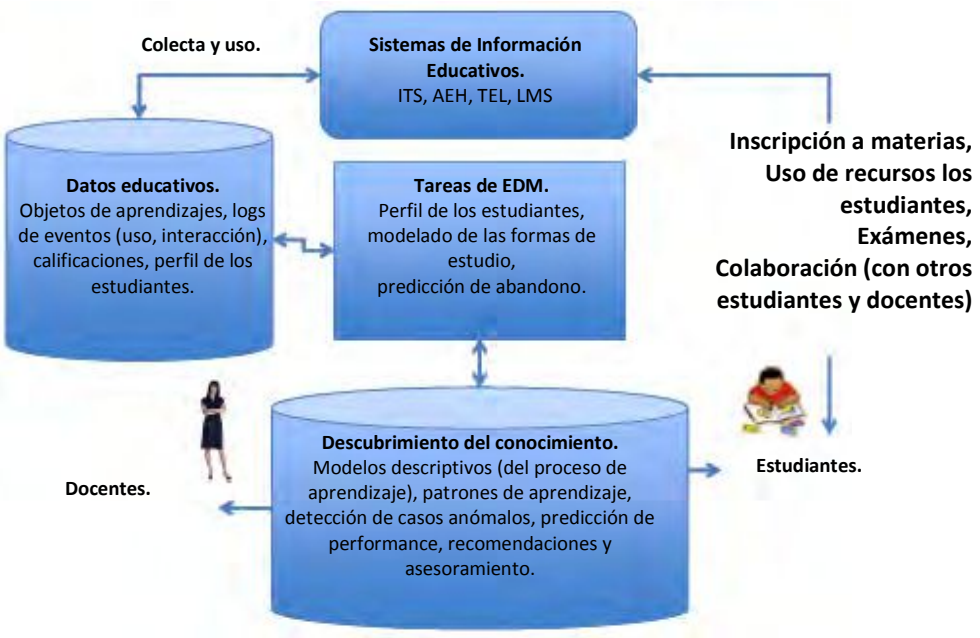


Figura 3.1. Ciclo de aplicación de minería de datos en el sistema educativo [Calders & Pechenizkiy 2011].

Por ejemplo los estudiantes pueden recibir consejos y recomendaciones acerca de cursos disponibles, actividades de aprendizaje, recursos o tareas teniendo en cuenta su conocimiento actual o los objetivos de aprendizaje. Los docentes pueden ver cuán efectivo es su material de estudios, cómo es el desarrollo de los estudiantes en las tareas que se proponen, identificar grupos de riesgo entre los estudiantes. Mientras que los responsables académicos pueden determinar falencias en el plan de estudios, la manera real de estudio de los alumnos y el desempeño de los docentes.

3.3 Clasificación de EDM

Según el conocimiento adquirido de la aplicación de técnicas de minería de datos al universo de educación, ésta puede estar orientada y beneficiar a los diferentes actores del sistema educativo –sector administrativo y responsables académicos, docentes y alumnos – [Romero & Ventura 2007]. Se agrupan los trabajos investigados, según esta clasificación.

3.3.1 EDM orientado al sector administrativo y responsables académicos

El objetivo es, entre otros, tener parámetros sobre cómo determinar y mejorar el nivel de estudiantes y docentes, la organización los recursos institucionales –humanos y materiales –, los programas educativos que se ofrecen y determinar la efectividad de nuevas propuestas.
Se analiza, en general, información relacionada con el desempeño de los alumnos; que incluyen información académica –inscripciones, promedios –, información socioeconómica y demográfica y forma de sustento. [Romero & Ventura 2007].

Muchos trabajos e investigaciones están dedicados al uso de minería de datos, para predecir la deserción de los estudiantes en el nivel universitario.
[Dekker, Pechenizkiy & Vleeshouwers 2009] realizan un caso de estudio, que intenta predecir una posible deserción, en el primer semestre de la carrera. Utilizan tres conjuntos de datos: uno contiene solamente datos universitarios de los alumnos, el segundo contiene solamente datos pre-universitarios y por último utilizan un conjunto de datos que contiene tanto datos universitarios como pre-universitarios. Realizan luego, una comparación entre clasificadores de árboles de decisiones y redes bayesianas, también realizan un modelo de aprendizaje basados en reglas.
[Timarán Pereira 2009] trata el tema de la deserción aplicando reglas de clasificación y asociación, con un modelo de árboles de decisión; determina de esta manera, diferentes perfiles de los estudiantes. Usa información académica y de precedencia.
[Kovacic 2010] explora variables socio-demográficas –edad, sexo, etnia, educación previa, situación laboral, y discapacidad – y el ambiente de estudio; que influyen en la persistencia o deserción de los estudiantes en un nivel universitario. Examina además en qué medida los datos de inscripción ayudarán en la pre-identificación de los estudiantes exitosos y no exitosos. Los modelos de minería de datos seleccionados en este trabajo fueron árboles de clasificación y árboles de regresión, haciendo una comparación entre los resultado obtenido de ambos.
[Obsivac et. al. 2012] además de tener en cuenta los datos académicos para predecir posibles abandonos en una etapa temprana de la carrera, también analizan datos del comportamiento social dentro de la institución. Estos datos, extraídos de correos electrónicos, foros y archivos compartidos, se relacionan con la comunicación de alumnos con sus compañeros y docentes. Los autores concluyen que el uso de estos datos sociales, incrementó la precisión en la predicción de abandonos. Usaron varios métodos de minería de datos, como árboles de decisión, máquinas de vectores soporte y redes bayesianas, haciendo luego una comparación.

[Thai-Nghe, Janecek & Haddawy 2007] realizan una comparación de la precisión obtenida al aplicar árboles de decisión y las redes bayesianas, para clasificar y predecir la performance académica de alumnos universitarios de grado y graduados de dos instituciones académicas diferentes. Los datos de los alumnos que se minaron fueron tanto académicos como sociales y culturales.
[Yadav & Pal 2012; Bhardwaj & Pal 2011] realizan estudios para predecir la performance de los estudiantes en los exámenes lo antes posible, y poder identificar los diferentes factores que afectan la conducta de aprendizaje y el rendimiento durante la carrera. En ambos trabajos se aplican clasificación. Los datos seleccionados se relacionan con la performance de los estudiantes en estudios anteriores, además de datos demográficos, nivel de estudios de los padres y hábitos sociales del estudiante. Yadav & Pal realizaron un modelo de árboles de decisión y los algoritmos de empleados para su realización fueron ID3, C4.5 y CART. Mientras que Bhardwaj & Pal seleccionaron redes bayesianas para minar los datos obtenidos.

En cuanto al uso de minería de datos para el análisis de los datos relacionados a los docentes de las instituciones, [Baracosa & Antunes 2011] proponen una metodología para anticipar el desempeño de los docentes, basado en el análisis de datos obtenidos de encuestas pedagógicas. El objetivo es encontrar comportamiento frecuente de los docentes en períodos lectivos pasados para poder anticipar la performance de los mismos en períodos futuros. Proponen de clasificación, usando árboles de decisión.
[Xu & Recker 2011] analizan los datos de los usuarios docentes de un sistema de bibliotecas. Utilizan agrupamiento, para identificar los diferentes grupos de docentes y sus características. Los datos usados se relacionan a los proyectos creados por los docentes, los recursos usados por proyectos y el perfil de navegación. Identifican 7 grupos de usuarios, que van desde los más dedicados a los menos dedicados, o de los usuarios aislados a aquellos que son clave para la comunidad.

3.3.2 EDM orientado a los docentes

La EDM orientada al docente, busca evaluar la estructura del contenido de los cursos y su eficacia en el proceso de aprendizaje; clasificar a los alumnos basados en sus necesidades en materia de orientación y supervisión; encontrar patrones de aprendizaje, errores más frecuentes cometidos por los alumnos en los temas dictados, actividades más eficaces; mejorar la adaptación y personalización de los cursos; en caso de los sistemas de e-learning, reestructurar los sitios personalizando los cursos, organizar los contenidos de manera eficiente para el progreso de los alumnos, mejorar las formas de evaluación.
Los datos analizados se refieren al desarrollo de los alumnos dentro de una materia particular. Si se los compara con los datos usados en la orientación administrativa, son más detallados y más referentes al día a día de la actividad del alumno dentro de la

materia; como por ejemplo evaluaciones parciales, entrega de actividades, comunicación con alumnos y docentes, participación en foros, consulta de bibliografía, entre otros. [Romero & Ventura, 2007].

[López, et al. 2012] muestran una correlación existente entre la participación de los alumnos en el foro de una materia y el resultado final obtienen en los exámenes de la misma. Proponen una clasificación mediante clusters, obteniendo resultados con similares precisiones que los algoritmos de clasificación tradicionales. Entre los datos seleccionados se encontraban, cantidad de mensajes, cantidad de conversaciones, cantidad de palabras, cantidad de oraciones, cantidad de mensajes leídos, cantidad de horas invertidas en el foro, resultado final obtenido por el alumno en la materia.

[Romero, et al. 2008] evalúan el rendimiento y la utilidad de diferentes algoritmos de clasificación, para la predicción de notas finales que los alumnos obtendrán en las materias que están cursando. El estudio se basa en la información de los logs de un sistema de e-learning que contienen datos sobre toda la actividad que los estudiantes realizan en el mismo. Entre los datos seleccionados se encuentran, cantidad de tareas y pruebas realizadas, cantidad de pruebas aprobadas y no aprobadas, cantidad de mensajes enviados y leídos, tiempo total utilizado en la realización de las pruebas y tiempo total de uso de la aplicación. Los clasificadores usados fueron árboles de decisión, reglas de inducción, redes neuronales y clasificadores de estadísticas.

[Dimic, et al. 2010] realizan una investigación para evaluar la calidad y el éxito de los cursos realizados en un sistema de e-learning, con el objetivo de predecir con antelación una posible falla del estudiante en los exámenes. Los datos analizados son referentes a las actividades de los estudiantes en el sistema: cantidad de lecturas realizadas, cantidad de tareas realizadas, cantidad de presentaciones ppts usadas, cantidad de auto-evaluaciones realizadas, resultado obtenido en el examen. Los métodos de minería de datos que usaron fueron clustering y clasificación mediante árboles de decisión, obteniendo mediante el mismo las reglas de clasificación.

[Falakmasir & Habibi 2010] investigan el impacto de las actividades en un sistema de e-learning en el aprendizaje de los alumnos, mostrando que la participación en las sesiones de aulas virtuales se relacionan con las calificaciones obtenidas en los exámenes. El estudio lo realizaron sobre los logs de uso del sistema e-learning; teniendo en cuenta las actividades de los estudiantes en el sistema, por ejemplo, cantidad de materiales de ayuda consultados, cantidad de participaciones en el aula virtual, cantidad de archivos vistos, cantidad de lectura y post en los foros, cantidad de conversaciones leídas y respondidas, resultado final del alumno en la materia. Realizaron una clasificación mediante árboles de decisión, usando el algoritmo C4.5, del cual luego se extrajeron reglas de clasificación.

[Antunes 2010] utiliza árboles de decisión para predecir, tan pronto como sea posible, si un alumno aprobará o no una materia, teniendo en cuenta la temporalidad de los datos, "*cada registro de los estudiantes pertenece a una secuencia ordenada de acciones y*

resultados". Usa clasificadores ASAP –as soon as possible–, que permite clasificar una nueva instancia sin la necesidad de conocer todos los valores de los atributos de la misma. Usa dos estrategias: clasifica solamente con los atributos para los cuales se conoce el valor –atributos observables –, y clasifica usando todos los atributos, completando el valor de los atributos no observables con valores estimados a partir de los valores de los atributos observables. El algoritmo usado es C4.5.

3.3.3 EDM orientado a los alumnos

El objetivo de la EDM orientado a los alumnos, es realizar diferentes sugerencias para acompañar el desarrollo del currículum universitario y ayudar a obtener buenos resultados en cada una de las materias. Teniendo en cuenta las experiencias de estudiantes anteriores con perfil académico y social similar, se pueden sugerir un desarrollo del itinerario académico; teniendo en cuenta el desarrollo del alumno dentro de una materia particular en un sistema e-learning, y sabiendo la bibliografía consultada por el alumno, se pueden sugerir nuevos links, recomendar actividades, recursos y tareas que favorezcan y mejoren su aprendizaje. [Romero & Ventura 2007].

[Vialardi, Bravo, et al. 2009] proponen el uso de técnicas de minería de datos para ayudar a los estudiantes en la elección del itinerario académico – elección de materias, horarios, aulas y profesores –. Teniendo en cuenta la performance académica de otros estudiantes con perfil similar, se sugiere al estudiante la elección de cuales y a cuantas materias inscribirse; el objetivo es que el estudiante logre terminar exitosamente cada uno de las materias que cursa. Los datos analizados comprenden información demográfica de los estudiantes, las materias en que se inscribieron, resultados obtenidos, cantidad de materias en cada ciclo lectivo, promedio y promedio acumulado. Realizan una clasificación mediante reglas, que sugieren a los estudiantes si la inscripción a un determinado curso puede ser exitosa o no.
En los sistemas de e-learning, los docentes proponen a los estudiantes Urls donde los alumnos pueden consultar sobre los temas del curso. La Web ofrece un conjunto mucho más amplio de fuentes de información extra, con las cuales el alumno puede enriquecer su conocimiento; esta cantidad de información puede resultar engorrosa y hasta invasiva si no se presenta de manera adecuada. [Godoy & Amandi 2010] Estudian el perfil de los estudiantes –compuesto por el material bibliográfico digital consultado –, para compararlos con el contenido de otras páginas Web y poder así, hacer recomendaciones personalizadas de nuevos links que puedan llegar a ser de interés. Usan clasificación mediante clúster (el vecino más cercano) para determinar los diferentes perfiles de los estudiantes y poder después hacer las recomendaciones.
[Tang & McCalla 2010; Nagata et al. 2009; Chen et al. 2008] aplicaron tareas y modelos de minería de datos, con el objetivo de proponer recomendaciones de bibliografía a los

usuarios de sistemas de bibliotecas. Usaron como base los registros de préstamos de la biblioteca, usaron reglas de asociación para buscar asociaciones con libros, enfocándose en el modo de préstamo del usuario, interés personal y características. Los datos usados se relacionan a los préstamos realizados al lector, así como también resultados de encuestas realizadas que funcionan como feedback de las recomendaciones ya realizadas, que ayudan a mejorar el aprendizaje y realizar mejores recomendaciones futuras. Utilizaron redes bayesianas para crear un sistema de recomendación de libros personalizado con el objetivo de generar diferentes recomendaciones asignándoles un ranking de mayor a menor para ayudar al lector a localizar la información del libro más adecuado a su necesidad.

3.4 Conclusión del capítulo

La EDM se comenzó a investigar hace varios años y se aplicada en instituciones académicas alrededor de todo el mundo. Especialmente ocupa a los investigadores proponer metodologías que puedan ayudar tanto a administrativos, docentes y alumnos a determinar aquellos casos de bajo nivel para poder tomar medidas apropiadas a tiempo y evitar así, la deserción a nivel universitario.

Vemos, en los trabajos de investigaciones consultados, que la tarea de minería de datos más frecuente, es la clasificación, siendo árboles de decisión el método más usado, aunque también se usan reglas de clasificación, clustering y redes bayesianas. Otra tarea aplicada con frecuencia, es el agrupamiento.

Es una práctica muy usada, la aplicación de más de una técnica de minería de datos sobre el mismo conjunto de datos, para poder luego comparar la eficiencia de cada uno.

PARTE II

HERRAMIENTAS Y DATOS

En esta parte, se hace una investigación y descripción de los diferentes y principales sistemas y herramientas de minería de datos que se encuentran disponibles actualmente en el mercado. También se describen los diferentes sistemas que se utilizan en la Facultad de Informática de la UNLP, para almacenar los datos de sus alumnos y egresados; que serán la fuente de datos para la realización de esta tesina.

CAPÍTULOS

Capítulo 4

SISTEMAS Y HERRAMIENTAS DE MINERÍA DE DATOS

En este capítulo, se investigan herramientas de minería de datos disponibles en el mercado; clasificándolas en tres grupos: librerías, suites y herramientas específicas. Las características descriptas de cada una hacen referencias, entre otros, a la facilidad de uso, la portabilidad, acceso a la información, los modelos y patrones disponibles.

4.1 Librerías de minería de datos

Comprenden un conjunto de métodos que implementan las funcionalidades básicas de la minería de datos: acceso a datos, inferencia de modelos, exportación y comprobación de datos. Facilitan el desarrollo de tareas de minería, como pueden ser el diseño de experimentos, el contraste de modelos, la creación de modelos combinados, y la integración de diversas técnicas de minería de datos. La principal desventaja que contiene es que son una interfaz para el desarrollo de aplicaciones de minería de datos, por lo que para su manejo, se precisa de conocimientos de programación.

4.1.1 Xelopes

Xelopes – eXtEnded Library for Prudsys Embedded Solutions – [prudsys XELOPES] es una librería con licencia pública GNU, implementada por *Prudys AG* en colaboración con *Russian MDA specialist ZSoft Ltd*. Implementa la mayoría de los algoritmos de aprendizaje. Permite la extensión con incorporación de nuevos métodos. Se caracteriza por:

- Acceso a datos: permite uniformidad a todos los modos de acceso a datos.
- Modelos: árboles de decisión, redes neuronales, métodos de agrupamiento, métodos de reglas de asociación.
- Exportación de datos: se pueden exportar los modelos y sus resultados a otros entornos de minería de datos, soportando el estándar PMML.

Xelopes librería está desarrollada bajo el estándar MDA –model drive architecture –, está disponible para C++ y Java, y también existe una interfaz para CORBA.

4.1.2 MLC++

MLC++ –Machine Learning library in C++ – [sgi – MLC++], es un conjunto de librerías para facilitar la comparación de resultados proporcionados por diversos algoritmos, sobre un mismo conjunto de datos. Desarrollada originalmente, de dominio público, por la Universidad de Stanford. Luego, se distribuyó por la compañía *Silicon Graphics* bajo dominio de investigación. Se integra con MineSet, de la compañía *Purple Insight* [Purple Insight - MineSet], lo que le proporciona extensiones para acceso a base de datos, tratamiento de la información y soporte visual.

A nivel general, MLC++ se caracteriza por:

- Acceso a datos: archivos con formato plano.
- Transformaciones de datos.
- Modelos de aprendizaje: Tablas de decisión, ID3, aprendizaje de árboles de decisión y métodos bayesianos. Además ofrece interfaces para invocar a C4.5, C5.0 y CART.

4.2 Suites

Una suite, integra en un mismo entorno, preprocesado de datos, diferentes modelos de análisis, diseño de experimentos y visualización de resultados. A diferencia de las librerías, el usuario no está condicionado a poseer conocimientos de programación, ya que la interfaz gráfica que posee, facilita la interacción.

4.2.1 IBM - SPSS

Distribuido por la empresa SPSS–IBM [IBM - SPSS]. Es un conjunto de herramientas visuales, orientado a la predicción. Se caracteriza por:

- Acceso a datos: ODBC, Excel, archivos planos ASCII y archivos SPSS.
- Preprocesado de datos.
- Técnicas de aprendizaje: árboles de decisión, redes neuronales, agrupamiento, reglas de asociación, regresión lineal y combinación de modelos.
- Técnicas para la evaluación de modelos.
- Visualización de resultados: ofrece un potente soporte gráfico que permite al usuario tener una visión global de todo el proceso, que comprende desde el análisis del problema hasta la imagen final del modelo aprendido.
- Exportación: generación automática de informes (HTML y texto), exportación de los modelos a distintos lenguajes (C, SPSS, HTML, estándar PMML, SQL para árboles de decisión y reglas).

4.2.2 RapidMiner

RapidMiner [RapidMiner] un sistema de código abierto de minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico, también se puede integrar en productos propios. Se distribuye bajo licencia AGPL y está hospedado en SourceForge desde el 2004. También permite utilizar los algoritmos incluidos en Weka. Entre sus características podemos destacar:

- Integración de datos de diferentes fuentes. ETL.
- Posee una interfaz gráfica intuitiva.
- Modelado y visualización de datos.
- Es extensible.

Proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento y transformación de datos y visualización.
Las siguientes figuras muestran la interacción con la herramienta, usando los datos obtenidos de la base de datos del sistema SIU-Guaraní para la realización de este trabajo de grado.

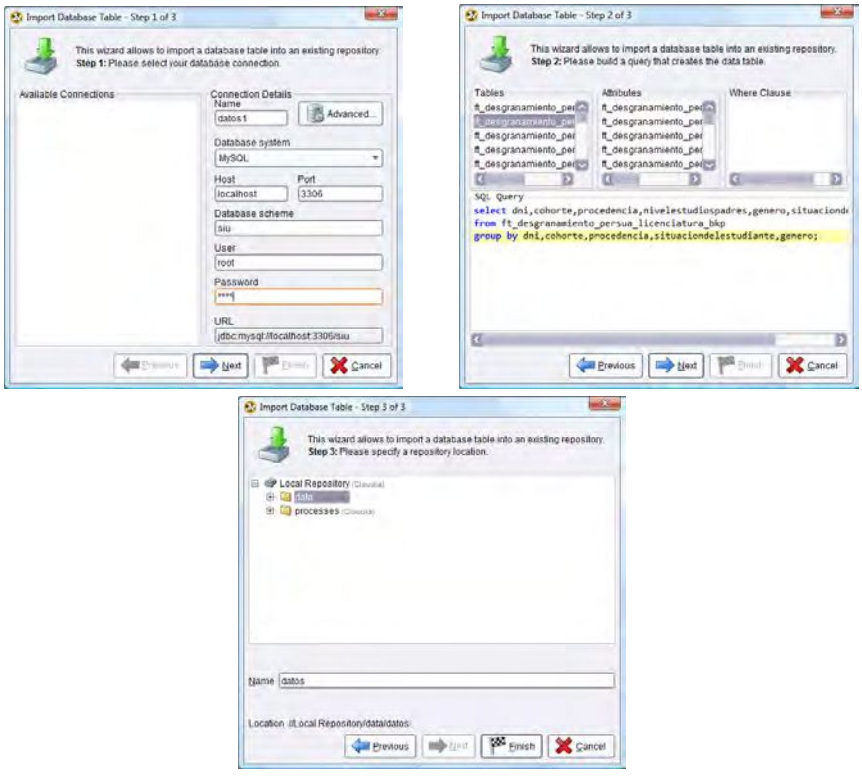


Figura 4.1 - Conexión a la base de datos con RapidMiner en 3 pasos: Seteo de los datos de conexión. Selección de la tabla de donde se extraerán los datos. Nombrado del repositorio localmente.

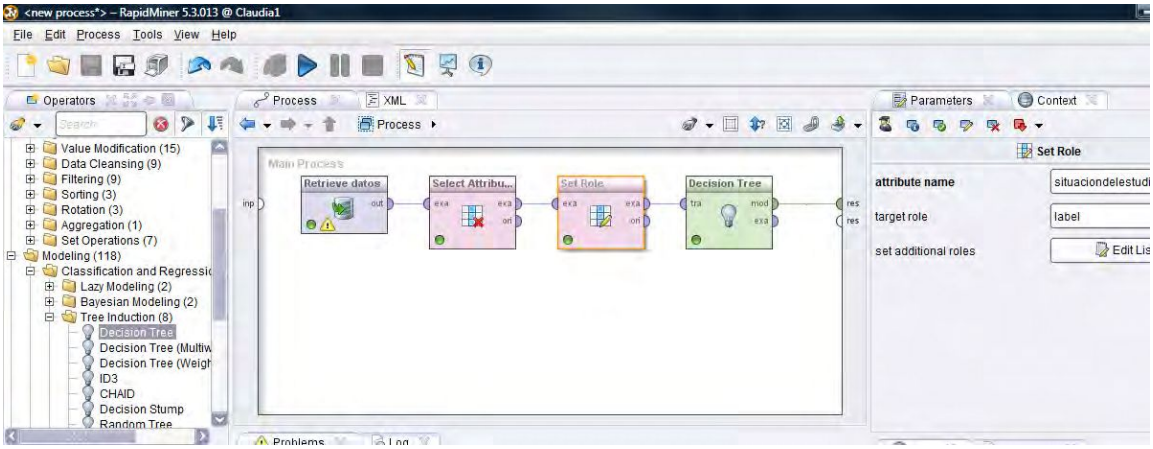


Figura 4.2 - Proceso de preprocesamiento de datos y generación de un árbol de decisión. Desde una misma pantalla de trabajo. De la misma manera se pueden realizar otras tareas de minería de datos como agrupamiento.

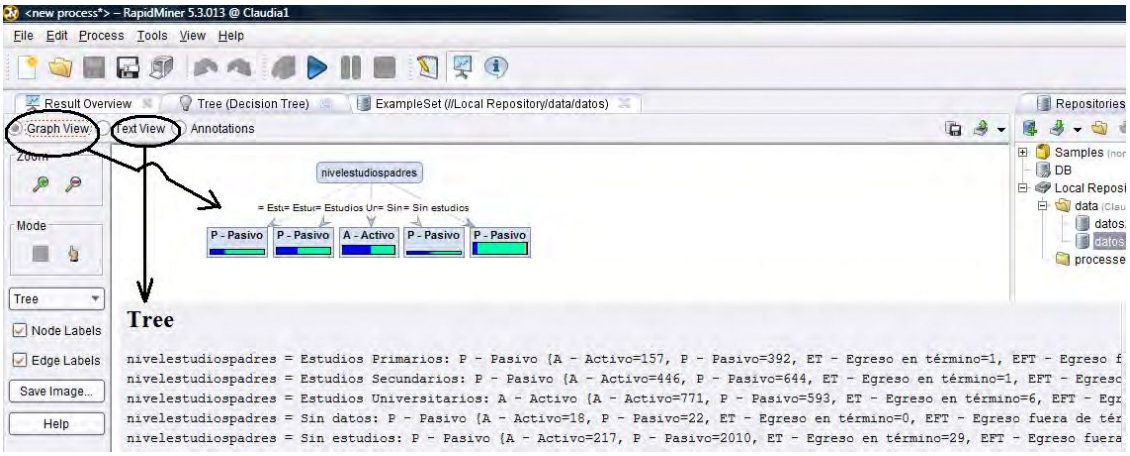


Figura 4.3 - Visualización de los resultados tanto en forma gráfica como en forma de texto.

4.2.3 WEKA

WEKA –Waikato Environment for Knowledge Analysis – [WEKA – The University of Waikato], es una herramienta visual de libre distribución (licencia GNU) desarrollada en JAVA, por un equipo de investigadores de la universidad de Waikato (Nueva Zelanda). En un inicio era sólo una librería, pero hoy en día es un paquete integrado (suite). Se puede destacar:

- Acceso a los datos: archivos en formato ARFF (archivo plano organizado en filas y columnas), aunque también acepta archivos de tipos Excel, csv, conexiones a bases de datos y Urls.
- Preprocesado de los datos: entre las tareas que se pueden realizar, se encuentran, entre otros, filtrado de datos, discretización, tratamiento de valores desconocidos, transformación de atributos numéricos.

- Modelos de aprendizaje: árboles de decisión (J4.8, versión propia del método C4.5), Reglas de asociación, Métodos de agrupamiento, Modelos combinados.
- Visualización: la interfaz gráfica se compone de diversos entornos
    - *Explorer*, permite el filtrado, selección del modelo, diseño de experimentos, visualización de resultados.
    - *CLI*, permite ejecutar todas las operaciones por línea de comandos.
    - *Experimenter*, facilita el diseño y realización de experimentos complejos.

KnowledgeFlow, de forma gráfica y a modo de flujo de operaciones, permite definir la totalidad del proceso.

Las siguientes figuras muestran la interacción con la herramienta, usando datos obtenidos de la base de datos del sistema SIU-Guaraní para la realización de este trabajo de grado.

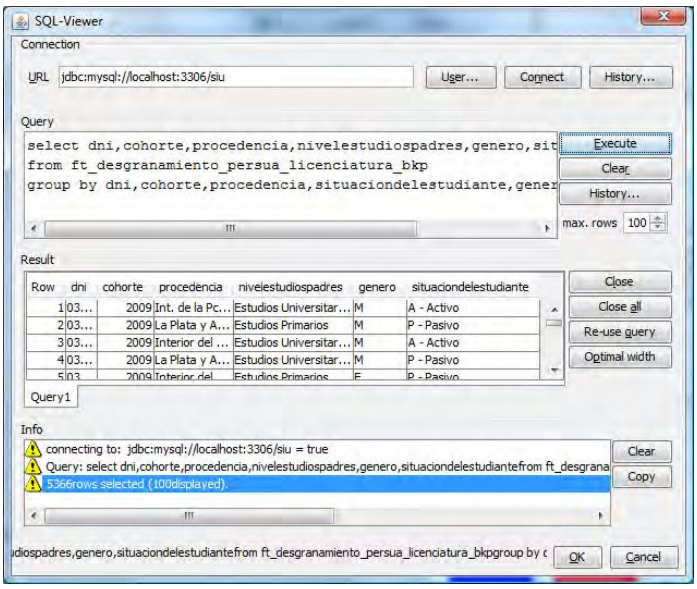


Figura 4.4 - Conexión a la base de datos con Weka en un solo paso.

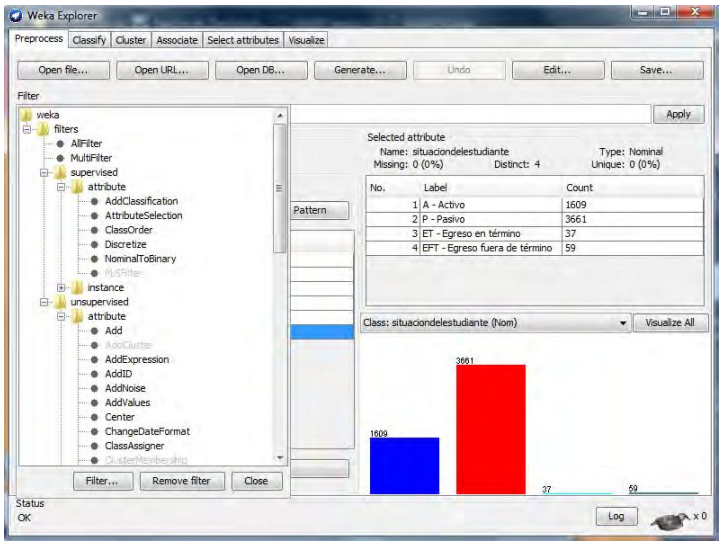


Figura 4.5 - Preprocesamiento de datos con weka.

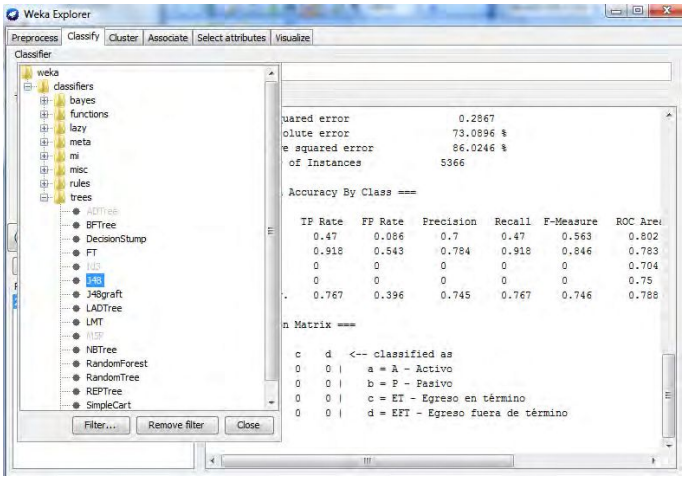


Figura 4.6 – Generación de un árbol de decisión y visualización de los resultados en forma de texto.

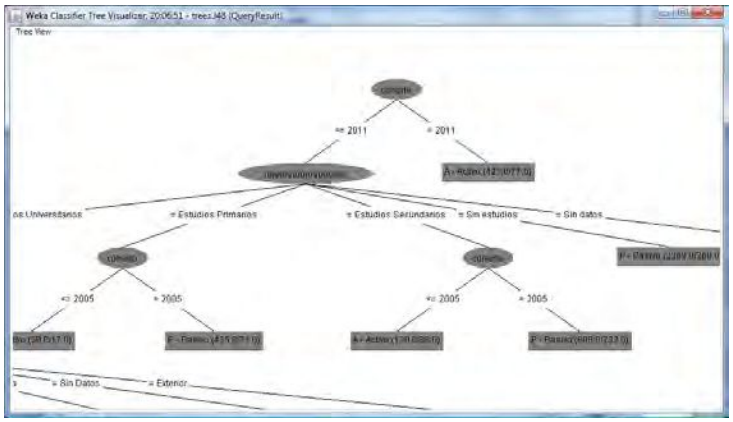


Figura 4.7 – Visualización de los resultados en formato gráfico. Se realiza seleccionando la opción de visualización luego del proceso de la Figura 4.6.

4.2.4 DBMiner

DBMiner es un sistema interactivo inicialmente desarrollado por *Data Base Systems Research Laboratory* de la *Universidad Simon Fraser* (Canadá) bajo licencia pública. La versión empresarial es desarrollada por *DBMiner Technology Inc.* [DBMiner Technology Inc.]. Es un sistema concebido para la extracción de conocimiento en grandes bases de datos relacionales, almacenes de datos y Web. Cuenta con tareas de clasificación, agrupamiento, asociación, entre otras.
Se destacan los módulos que están interconectados:

- OLAP –Online Analytic Processing –: funcionalidad de manejo multidimencional del almacén de datos.
- OLAM –Online Analytic Mining –: funcionalidad específica de minería de datos.
- DBMiner cuenta con dos modos de trabajo:

- Interfaz gráfica: permite al usuario solicitar cualquier operación OLAP y/o OLAM a través de una interfaz gráfica.
- Lenguaje de script: lenguaje de consulta similar al SQL denominado DMQL para especificar las diferentes tareas de minería.

4.2.5 SAS Enterprise Miner

Herramienta de minería de datos de SAS Institute [SAS - Enterprise Miner]. Posee una arquitectura distribuida, en donde toda la funcionalidad del sistema es accesible mediante una potente interfaz gráfica. Las tareas soportadas por el sistema son:

- Acceso a datos: formato de archivo propio de SAS, bases de datos.
- Preprocesado de datos: transformaciones, tratamiento de valores desconocidos, filtros, entre otras.
- Modelos: árboles de decisión, redes neuronales, construcción de modelos múltiples, entre otros.
- Evaluación: este módulo permite la comparación entre diferentes modelos de aprendizaje.

Visualización y presentación de resultados: gráficos en dos y tres dimensiones, visores de árboles de decisión, generador de informes HTML, presentación de la información en lenguaje natural.

4.3 Herramientas específicas

A diferencia de la generalidad propia de las suites, este tipo de entornos se caracteriza por centrarse en un determinado modelo (redes neuronales, árboles de decisión, modelos estadísticos, entre otros) o en una determinada tarea de minería de datos (clasificación, agrupamiento, entre otras). Pese a incorporar sólo un tipo de técnicas, son un entorno que permite realizar todo el proceso de minería de datos. Tampoco se requieren conocimiento de programación para poder ser utilizadas.

4.3.1 CART

Herramienta gráfica desarrollada y comercializada por Salford Systems [Salford Systems - CART]. Contiene utilidades para el análisis estadístico y la minería de datos orientada hacia la inferencia de árboles de decisión. Como entorno de minería de datos se destacan:

- Acceso a datos: tiene acceso a más de 70 formatos de archivos diferentes.
- Visualización: dispone de herramientas de visualización interactivas.
- Información estadística relativa al modelo: errores de clasificación, influencia en un atributo de la clasificación, entre otras.

4.3.2 NeuroShell

Son un conjunto de herramientas gráficas independientes,
de las que se destacan *NeuroShell 2*, *NeuroShell Predictor*, *NeuroShell Classifier* y *NeuroShell Trader*, desarrolladas y comercializadas por Ward System Group [Ward System Group, Inc.] para trabajar fundamentalmente con modelos de aprendizaje basados en redes neuronales. La primera de éstas, *NeuroShell 2*, es una herramienta de uso muy intuitivo pero cuya aplicación se restringe al ámbito académico. El resto de las aplicaciones son más robustas y se utilizan para problemas más reales.

- *NeuroShell Predictor,* se utiliza para la predicción de variables numéricas tales como índices de venta, precios de mercado, costos, entre otros. Se basa en el desarrollo de unas técnicas propias de redes neuronales y estimadores estadísticos guiados por algoritmos genéticos. El entorno dispone de capacidades de representación gráfica de la información estadística asociada al modelo.
- *NeuroShell Classifier,* optimiza las técnicas del Predictor con el fin de utilizar dichos modelos para tareas de clasificación.
- *NeuroShell Treder*, añade técnicas de lógica difusa e indicadores de agrupamiento enfocados ambos hacia el reconocimiento de patrones.

4.3.3 See5 / C5.0

Es una herramienta desarrollada y comercializada por la empresa *RuleQuest Research Pty Ltd* dirigida por Ross Quinlan [RuleQuest Research See5 / C5.0]. Se centra en la construcción de modelos de clasificación basados en árboles de decisión y conjuntos de reglas. Además permite la combinación de modelos.
La herramienta ha sido diseñada para operar sobre grandes volúmenes de datos. Los modelos aprendidos pueden ser exportados a código en C, por lo que pueden ser incorporados como parte de otros sistemas de aprendizaje.

4.4 Cuadro comparativo

En el siguiente cuadro, se realiza una comparación entre las diferentes herramientas descriptas anteriormente, con las características principales que se tendrán en cuenta luego, en el momento de la selección de una de ellas, para la realización de este trabajo de grado.

Características	Librerías		Suites					Herramientas Específicas		
	Xelopes	MLC++	Clementine	RapidMiner	Weka	DBMiner	SAS	Cart	NeuroShell	See5
Licencia Libre.	Sí	No	No	Sí	Sí	No	No	No	No	No
Requiere otros conocimientos.	Sí	Sí	No	No	No	No	No	No	No	No
Diferentes fuentes de datos.	Sí	Acotado	Sí	Sí	Sí	Sí	Sí	Sí	No Aplica	Sí
Multiplataforma.	No Aplica	No Aplica	No Aplica	Sí	Sí	No Aplica	No	No	No	No
Preprocesamiento y filtrado de datos.	No	Sí	Sí	Sí	Sí	No	Sí	No	No	No
Múltiples tareas de minería de datos.	Acotado	Acotado	Sí	Sí	Sí	Sí	Sí	No	No	No
Visualización de resultados.	No	Acotado	Sí	Sí	Sí	Sí	Sí	Sí	Sí	No
Integración en una aplicación propia.	Sí	No	No	Sí (Java)	Sí (Java)	No	No	No	No	Sí
Extensible.	Sí	No	No	Sí	Sí	No	No	No	No	No
Puede combinar modelos.	No	No	No Aplica	Sí	Sí	No	No	No	No	Sí
Interfaz gráfica amigable.	No	No	Sí	Sí	Sí	Sí	Sí	Sí	No	No

Capítulo 5

POSIBLES FUENTES DE DATOS

Los datos que se analizarán en este trabajo serán obtenidos de las bases de datos de los sistemas basados en software libre que se utilizan en la Facultad de Informática y registran la información de sus estudiantes y egresados. En este capítulo se describen las funcionalidades principales que proveen todos los sistemas involucrados y en el capitulo *6. Presentación del problema*, se indica cuáles de ellos se seleccionaron y qué datos son de utilidad para el desarrollo de esta tesina.

5.1 SIU

El SIU –Sistema de Información Universitaria – [Consorcio SIU], desarrolla soluciones informáticas y brinda servicios para el Sistema Universitario Nacional. Entre sus objetivos se encuentran:

- Promover un sistema de información integral a las Universidades.
- Garantizar la disponibilidad, integridad, seguridad y calidad de la información.
- Colaborar en el análisis de la información producida.
- Contribuir con la transparencia de la gestión.

Entre los sistemas informáticos que desarrolla este consorcio, se encuentran:

- SIU-Guaraní: Registra y administra las actividades académicas de la universidad, desde que los alumnos ingresan como aspirantes hasta que obtienen el diploma.
- SIU-Kolla: Permite un seguimiento de graduados, a fin de obtener información sobre su inserción laboral, su relación con la universidad.
- SIU-Tehuelche: Es un sistema de gestión de becas universitarias.
- SIU-Araucano: Es un sistema de información estadística de alumnos.

5.1.1 SIU Guaraní

Una de las soluciones del SIU, que especialmente nos interesa para obtener información de los estudiantes para este trabajo de grado, es el SIU-Guaraní [SIU-G]. Fue concebido para administrar la gestión de alumnos en forma segura, con la finalidad de obtener información consistente para los niveles operativos y directivos. Entre los servicios que el SIU-Guaraní provee, se pueden citar:

- Inscripciones a cursadas y exámenes.

- Resultados de cursadas, exámenes.
- Solicitud de certificados.
- Consulta de actas de examen, planes de estudios y promociones.

El SIU-Guaraní provee además, un módulo que exporta varios Data Mart de una Data Warehouse. Los mismos están orientados al OLAP –On-Line Analytical Processing – y abarcan diferentes temáticas como ser:

- Cubo 02. Rendimiento Académico. Desarrollado para evaluar el trabajo de los docentes y el rendimiento de los alumnos. Permite analizar resultados de cursados, exámenes y equivalencias de las materias y cátedras".
- Cubo 03. Procedencia. Permite analizar la evolución de la matrícula de cada carrera, permitiendo discriminar por país, provincia, localidad y colegio secundario.
- Cubo 04. Desgranamiento. Creado para ayudar en la determinación de las causas de la deserción que se produce en las distintas carreras dentro de una unidad académica. Relaciona el rendimiento académico de los estudiantes con factores sociales y la procedencia de los mismos [Cubo 04 - Desgranamiento].
- Cubo 05. Alumnos. Análisis de matrícula, para los diferentes años académicos, comparaciones históricas. Rendimiento.

5.2 Moodle

Moodle [Moodle] es un Sistema de Gestión de Cursos de Código Abierto (*Open Source Course Management System, CMS*), también conocido como Sistema de Gestión del Aprendizaje (LMS). Es una aplicación Web gratuita, que los educadores pueden utilizar para crear sitios de aprendizaje en línea ó como complemento el aprendizaje presencial. Los módulos principales de Moodle son:

- Tareas: Permite la entrega de tareas por parte de los estudiantes, el seguimiento, evaluación y observaciones de las mismas por parte de los docentes.
- Foros: Permite que docentes y estudiantes se comuniquen mediante conversaciones iniciadas en diferentes foros.
- Cuestionarios: Los profesores pueden armar cuestionarios con diferentes preguntas, que el estudiante responderá. Se pueden calificar automáticamente.
- Recursos: Admite la presentación de contenido digital.
- Wiki: Permite a los alumnos trabajar en grupo sobre un mismo documento.

La aplicación Moodle en su versión 2.2, es usada por muchas de las cátedras de la Facultad de Informática – UNLP, como entorno virtual de aprendizaje de sus alumnos.

5.3 WebUNLP

WebUNLP [WebUNLP] es un entorno virtual de enseñanza y aprendizaje, creado por el Instituto de Investigación en Informática LIDI de la Universidad Nacional de La Plata. Permite a docentes y alumnos compartir materiales de estudio, comunicarse y generar una experiencia educativa en forma virtual; permitiendo flexibilizar el proceso de enseñanza y aprendizaje.
Dentro de un curso, podrán trabajar con diferentes unidades pedagógicas ó áreas:

- Comunicación (mensajería, foros y cartelera de novedades).
- Información General y Contenidos.
- Recursos Educativos.
- Trabajo Colaborativo.
- Evaluación.
- Gestión y Seguimiento.

La aplicación WebUNLP, es usada por muchas de las cátedras de la Facultad de Informática - UNLP como entorno virtual de aprendizaje de sus alumnos.

5.4 Merán

Meran [Meran] es un Sistema Integrado de Gestión de Bibliotecas (SIGB) que permite administrar los procesos bibliotecarios y gestionar servicios a los usuarios en forma integrada. Desarrollado por el grupo de desarrollo interdisciplinario del Centro Superior para el Procesamiento de la Información (CeSPI), dependiente de la Universidad Nacional de La Plata (UNLP). Profesionales informáticos, bibliotecarias y diseñadores se encargan del desarrollo y ofrecen capacitación y asistencia técnica a la comunidad de usuarios.
Meran es utilizado actualmente en bibliotecas de la Universidad Nacional de La Plata. Ha sido liberado como producto Open Source, bajo licencia GPL v3. Entre las funcionalidades que provee se encuentran las siguientes:

- Consulta del catálogo.
- Gestión virtual de reservas y renovaciones de bibliografía.
- Interacción de los docentes con la biblioteca.
- Alertas y notificación a los usuarios por correo electrónico.
- Creación de estantes virtuales, que permite estructurar la consulta de bibliografía disponible por cada cátedra, por carrera u otro criterio de asociación.
- Gestión de tareas administrativas; obtención de estadísticas y reportes.

PARTE III

SOLUCIÓN PROPUESTA

En esta parte del trabajo, se presenta la solución propuesta, y el trabajo realizado para alcanzar el objetivo.

CAPÍTULOS

Capítulo 6
PRESENTACIÓN DE LA SOLUCIÓN PROPUESTA

Se introduce en este capítulo, el trabajo realizado para la solución propuesta al análisis de los datos obtenidos de los sistemas de la Facultad de Informática. En los capítulos siguientes, se detallarán los pasos de cada una de las etapas de solución.

6.1 Recursos Disponibles

Para poder contar con los datos necesarios para la realización de este trabajo; en Agosto de 2012, se presentó una nota a las autoridades de la Facultad de Informática de la UNLP, donde se solicitó autorización para el acceso a los datos de los sistemas utilizados por esta unidad académica, para el manejo de información de sus alumnos y egresados; previo compromiso a que los mismos permanezcan en total anonimato, sin comprometer la identidad de ninguno de los alumnos ni egresados.
La Secretaria Académica de la Facultad de Informática, UNLP, Lic. Claudia Queiruga, en respuesta a esta solicitud, considera pertinente acceder al pedido – Expediente 3300-4008/11-001–.
De esta manera, se logra obtener el acceso a la información necesaria. Se anexa en el Apéndice B, las copias obtenidas de la resolución.

6.2 Sistemas y Datos Seleccionados

Teniendo en cuenta los sistemas usados en la Facultad, para el almacenamiento de datos de sus estudiantes y graduados, se decide utilizar, datos provenientes de los sistemas SIU-Guaraní, Moodle y Merán. Esta elección se debe a que se prefiere, en la medida de lo posible, seleccionar datos de sistemas de código abierto, que faciliten mediante bibliografía publicada o la facilidad de acceso, el estudio de la estructura de su base de datos. El sistema SIU-Guaraní es el único de los utilizados en este trabajo de grado, cuyo código no está distribuido bajo una licencia de software libre; sin embargo, se distribuye en forma libre en todas las instituciones educativas de nivel superior de Argentina, que tengan la intención de usarlo.
La integración de los datos entre estos tres sistemas es posible debido a que todos ellos almacenan al menos un dato en común con el cual se puede identificar unívocamente al estudiante, este dato es el DNI del estudiante y el legajo del mismo. Si bien esta información se mantiene oculta en los análisis realizados, son de gran ayuda en la

correlación de los datos. De la misma manera que se identifica al estudiante, es posible identificar en los tres sistemas a las materias mediante un código o el nombre de la misma.

6.2.1 Datos seleccionados del sistema SIU-Guaraní

Luego de analizar detalladamente la documentación que describe cada uno de los Data Mart del sistema SIU-Guaraní, mencionado en el *Capítulo 5. Posibles Fuentes de datos*; sección *5.1. SIU*; se determinó que el *Cubo 04 – Desgranamiento*, puede ser de gran utilidad para este trabajo de grado, ya que aborda la temática de la deserción desde el punto de vista académico, social y demográfico y nos aporta información sobre la situación actual del estudiante. También se extrajeron datos de las tablas que contiene información de los egresados para poder determinar, de cada uno de ellos, el tiempo total empleado en la conclusión de la carrera. Por último, fue necesaria la selección de los datos de las tablas que contienen información de las materias aprobadas por los estudiantes.

6.2.2 Datos seleccionados del sistema Moodle

Como se mencionó anteriormente en el *Capítulo 5. Posibles Fuentes de Datos*, en la sección *5.2. Moodle*, este sistema cuenta con varios módulos que permiten la interacción entre estudiantes y docentes. El módulo más usado entre las cátedras, es el módulo de Foros. La figura 6.1 – Distribución de uso de los diferentes módulos del sistema Moodle; muestra la proporción de uso de los diferentes módulos del sistema Moodle. Es por este motivo que se seleccionaran, de este sistema, los datos de las tablas que almacenan toda la información de los foros y el uso de los mismos, como por ejemplo conversaciones, post y tiempo empleado en el mismo. Una tabla muy importante para este análisis es la tabla de logs, contiene información del uso del sistema de cada una de las sesiones iniciadas.

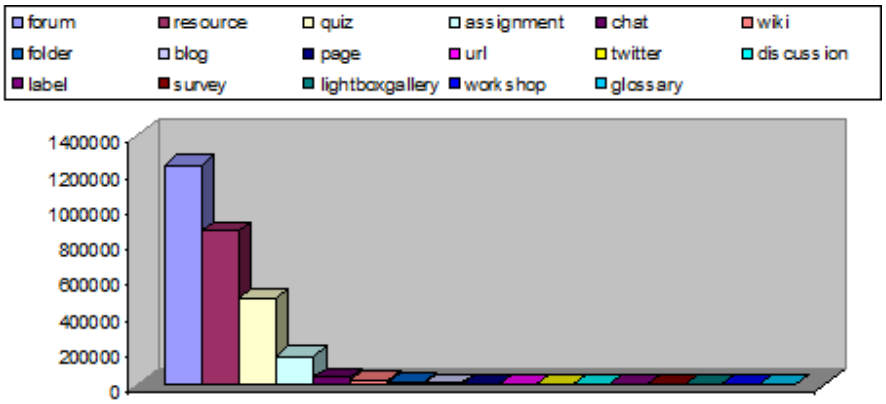


Figura 6.1 – Distribución del uso de los diferentes módulos del sistema Moodle.

Para tener una relación entre la imagen y las referencias, la imagen debe verse se izquierda a derecha y las referencias leerse de izquierda a derecha y de arriba hacia abajo.

6.2.3 Datos seleccionados del sistema Merán

Del sistema Merán, se seleccionaron los datos relacionados a la consulta de bibliografía relacionada a las materias que cursan –préstamos, renovaciones, reservas –. Esta información se pudo obtener del módulo de estantes virtuales con el que cuenta el sistema Merán. Los estantes virtuales, permiten relacionar el material de consulta de la biblioteca, con las materias de la Facultad. Las condiciones específicas de extracción de datos que se pidieron fueron las siguientes:
"*Estudiantes de las carreras de la Facultad de Informática - UNLP, con cohorte lo más antigua posible. Sin información detallada por bibliografía, es suficiente saber el estante virtual (materia/carrera) en el que está incluido el material. Del estante virtual, como mínimo se necesita saber el nombre de la materia. Tampoco es necesario tener el historial de sanciones del estudiante, sólo usaríamos los datos de los préstamos y reservas. En resumen, los datos que usaríamos para el análisis serían los siguientes:*
Nombre, DNI ó legajo del estudiante. Fecha de la reserva ó préstamo del material. Materia del estante virtual al que pertenece el material."

6.3 Análisis propuestos

El objetivo de este trabajo de grado, es la aplicación de tareas y técnicas de minería de datos, sobre los datos de los alumnos y egresados de la Facultad de Informática - UNLP; para poder determinar si existen patrones de comportamiento en los mismos.
Se propone analizar estos datos desde dos puntos de vistas diferentes.

- Analizar los datos de los estudiantes y egresados de la Facultad de Informática de la UNLP, utilizando un único data set; es decir, un conjunto reducido de datos obtenidos de un único sistema utilizado en la Facultad de Informática. Para este análisis se decidió trabajar con datos extraídos del cubo 04 del sistema SIU-Guaraní; analizando los datos desde el punto de vista de factores sociales y demográficos, para intentar establecer si existe una relación entre éstos y la situación académica de los estudiantes.
- Realizar una correlación de datos de los estudiantes, obtenidos de diferentes sistemas utilizados por la Facultad, con el objetivo de obtener un análisis más integrado. Para este análisis se decidió correlacionar datos pertenecientes a los sistemas SIU Guaraní, Moodle y Merán. Se analizará el comportamiento que tienen los estudiantes en las diferentes materias que cursa, que lo lleven a aprobar la misma. Para esto, se tendrá en cuenta la interacción social que los alumnos tienen en las materias con sus compañeros y docentes y el uso que los estudiantes hacen de bibliografía relacionada a la materia que cursan, disponible en la biblioteca de la Facultad.

6.4 Herramienta de minería de datos seleccionada

Para la elección de la herramienta a utilizar se analizaron las distintas alternativas, descriptas en la sección de Suites, del *Capítulo 4. Herramientas de Minería de Datos*, ya que éstas ofrecen una funcionalidad más amplia con respecto a las descriptas bajo los títulos de librerías y herramientas específicas.
De las herramientas Suites descriptas, se tenía conocimiento previo de la herramienta WEKA. Teniendo en cuenta las funcionalidades que ésta provee, y analizando las funcionalidades necesarias para la realización de este trabajo, se decide que WEKA es una buena opción para la realización de los análisis que aquí se proponen.
Otras de las ventajas que podemos describir de WEKA, son las siguientes:

- Contiene una extensa colección de técnicas para preprocesamiento de datos y modelado. Esta facilidad permite la comparación de resultados, al crear más de un modelo sobre un mismo conjunto de datos.
- Es portable, porque está completamente implementado en Java y puede correr en casi cualquier plataforma. También facilita la integración de sus librerías en un sistema propio que requieran del uso de algoritmos de minería de datos.
- Para su uso no requiere conocimientos previos de programación, lo que lo convierte en un sistema al cual pueden acceder usuarios no relacionados con la programación.
- Permite a los usuarios expertos en el área de informática extender su librería para agregar nuevos métodos y algoritmos de minería de datos ó corregir alguno de los existentes.
- Posee una interfaz gráfica amigable que permite el uso de todas las funcionalidades del sistema de manera intuitiva.
- Está disponible libremente bajo la licencia pública general de GNU.
- Un nuevo paquete, a partir de la versión 3.7.5 de WEKA, agrega un conector a la base de datos Cassandra, una base de datos NoSQL.

6.5 Extracción de conocimiento

Para la generación de cada uno los análisis, se ejecutaron los pasos sugeridos en el proceso de KDD; descriptos en el *Capítulo 1. BI, KDD y DM*, sección *1.3.3. Pasos del proceso de KDD*. Se pueden resumir en:

- Generación de la vista minable. Compuesto por la selección, preprocesamiento y transformación de los datos extraídos de las diferentes fuentes de datos.
- Generación y evaluación de los modelos.

6.5.1 Generación de las vistas minables

Se creará una vista minable diferente para cada uno de los análisis propuestos. Para trabajar con los datos obtenidos de manera más óptima, se creó un almacén de datos local, con una estructura de tablas similares a los sistemas fuentes, en donde se almacenó toda la información obtenida y a partir del cual se realizó el preprocesamiento luego. Este almacén de datos se creó sobre una base de datos MySQL; utilizando la herramienta MySQL Workbench [MySQL Workbench].

MySQL Workbench es una herramienta visual de base de datos para arquitectos, desarrolladores y administradores de bases de datos. Ofrece, entre otras prestaciones, modelado de datos, desarrollo de SQL y herramientas completas de configuración de servidor y administración de usuarios.

Los diagramas de Entidad-Relación, que se pueden observar en el capítulo siguiente, se realizaron con la herramienta DIA [DIA].

DIA es una aplicación informática de propósito general para la creación de diagramas, desarrollada como parte del proyecto GNOME. Incluye diagramas entidad-relación, diagramas UML, diagramas de flujo, diagramas de redes, diagramas de circuitos eléctricos, etc. Nuevas formas pueden ser fácilmente agregadas. Puede producir salida en los formatos EPS, SVG, PNG, JPG entre otros.

El preprocesamiento y transformación de los datos se realizaron mediante consultas SQL, no se consideró necesario para este trabajo, la utilización de herramienta de ETL específicas; ya que los datos extraídos de las bases orígenes fueron bastantes concretas –especialmente los extraídos del sistema SIU y Merán –, mientras que los datos utilizados del sistema Moodle, provinieron de un grupo reducido de tablas.

6.5.2 Generación de los modelos

Como se mencionó en el *Capítulo 2. Minería de datos y estilos de aprendizaje*, en la sección *2.3. Pasos para el desarrollo de la fase de minería de datos*; es necesario tomar una serie de decisiones en el momento de realizar la fase de minería de datos del proceso de KDD, ellas son las siguientes:

- Seleccionar el tipo o tarea de minería de datos.

- Seleccionar del método o técnica que resuelva la tarea.
- Seleccionar del algoritmo que resuelva la tarea seleccionada.

6.5.3 Tareas y técnicas seleccionadas

En el *Capítulo 3. Minería de datos educativa - EDM*, sección *3.5. Conclusión del capítulo*; se menciona que las tareas de minería de datos más comúnmente aplicadas en minería de datos educativa, son las siguientes:

- Clasificación como tarea predictiva. El método seleccionado fue árboles de decisión.
- Agrupamiento (ó clustering) en caso de querer resolver una tarea descriptiva.

Se decide entonces, aplicar estos dos tipos de minería de datos para la realización de los análisis propuestos para este trabajo.

6.5.4 Algoritmos aplicados

Para cada uno de los análisis propuestos, aplicará, el algoritmo J48; la implementación de WEKA, del algoritmo C4.5. El parámetro más importante de este algoritmo es el ***factor de confianza de la poda***; que controla el tamaño y complejidad del árbol generado. El valor por defecto de este factor es de 0.25, y conforme va bajando, se permiten más operaciones de poda, llegando a árboles cada vez más pequeños. Se ejecutará, para cada análisis, el algoritmo J48, con los valores 0.25, 0.001 y 1.0 en el parámetro de factor de confianza de la poda; comparando los resultados obtenidos.
Para evaluar los modelos de clasificación generados, se usará la opción Cross-Validation, que calcula el porcentaje de aciertos haciendo una validación cruzada, dividendo el conjunto de datos de entrenamiento en k segmentos, por defecto el valor de K es 10.

Para la generación de clusters, se usará el algoritmo SimpleKMeans que es la implementación en Weka del algoritmo K-Medias, una técnica básica que agrupa las instancias de un conjunto de datos por la similitud de sus propiedades. El parámetro más importante para configurar, es el que indica la cantidad de grupos –ó clusters – a generar. Se ejecutará este algoritmo para formar 3 y 5 clusters.

La aplicación de estas tareas y técnicas en las vistas minables creadas, así como también la visualización de los resultados obtenidos, se detallan en el *Capítulo 8 – Modelado y evaluación*.

6.6 Interfaz de interacción

Para que el usuario final pueda interactuar con los análisis realizados sin necesidad de tener conocimiento del uso de la herramienta WEKA; se realizó un aplicativo que permite la ejecución de estos algoritmos de manera amigable.
La aplicación se desarrolló en JAVA. Se utilizó Eclipse, en su versión Eclipse Classic 4.2.1 – JUNO [Eclipse], como entorno de desarrollo. Para poder aplicar las diferentes funcionalidades provistas por WEKA, se utilizaron las librerías de WEKA escritas en JAVA [Weka]. En el capítulo *9. Aplicativo*, se presenta una guía de uso del sistema.

Java es un lenguaje de programación de propósito general, concurrente, basado en clases, orientado a objetos e independiente a la plataforma, para lo cual es necesario tener instalado una máquina virtual java en el dispositivo en que se vaya a ejecutar el programa.

Eclipse consiste en un Entorno de Desarrollo Integrado (IDE), abierto y extensible. Sirve como IDE Java y cuenta con numerosas herramientas de desarrollo de software. También da soporte a otros lenguajes de programación, como son C/C++, Cobol, Fortran, PHP o Python. A la plataforma base de Eclipse se le pueden añadir extensiones (plugins) para extender la funcionalidad.

Capítulo 7

SELECCIÓN, PREPROCESAMIENTO Y TRANSFORMACIÓN

Este capítulo detalla la realización de los primeros tres pasos del proceso de KDD, Selección, Preprocesamiento y Transformación de datos; para los análisis propuestos en el *Capítulo 6 – Presentación del problema, sección 6.3 – Análisis propuestos*.

7.1 Análisis socio-demográfico

En esta sección, se describe la creación de la vista minable para el análisis de los datos desde el punto de vista de factores sociales y demográficos; para el cual se utiliza un único data set obtenido del cubo 04-desgranamiento del sistema SIU-Guaraní. Se indican los datos exportados del sistema fuente, su almacenamiento en una base de datos local, y la estructura final de la vista minable.

7.1.1 Exportación e importación de los datos

Para este análisis, se extrajeron los datos del modelo estrella del *Cubo 04 – Desgranamiento* del sistema SIU-Guaraní. La Figura 7.1 - Cubo 04 – desgranamiento, muestra el diagrama de Entidad-Relación de este modelo estrella.
También se extrajo información de otras tablas del mismo sistema, que contienen información de egresados. La Figura 7.2 – Egresados, contiene su estructura.
Los datos extraídos del sistema fuente, se guardaron en archivos con extensión .csv, y luego se importaron a las tablas creadas en el ambiente local.
En la Figura 7.3 – Base de datos SIU en ambiente local, se puede ver la estructura de la base de datos SIU, creada en el ambiente local.
Los scripts de exportación/importación de los datos, se encuentran en el *Apéndice A; 1. Creación y llenado de las tablas en el ambiente local*.

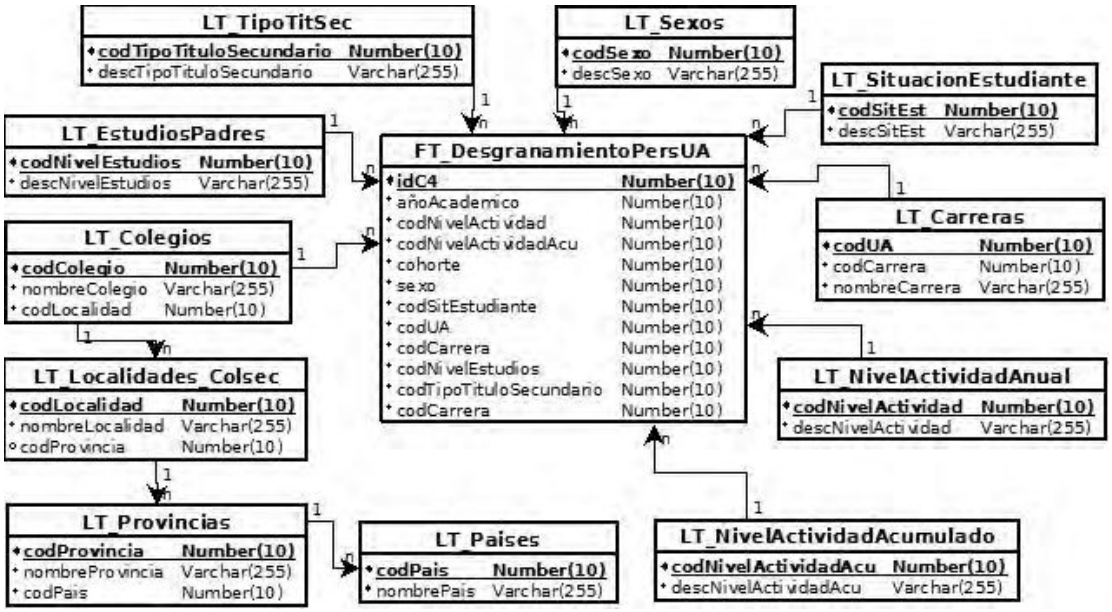


Figura 7.1 - Cubo 04 – desgranamiento.

Figura 7.2 – Egresados.

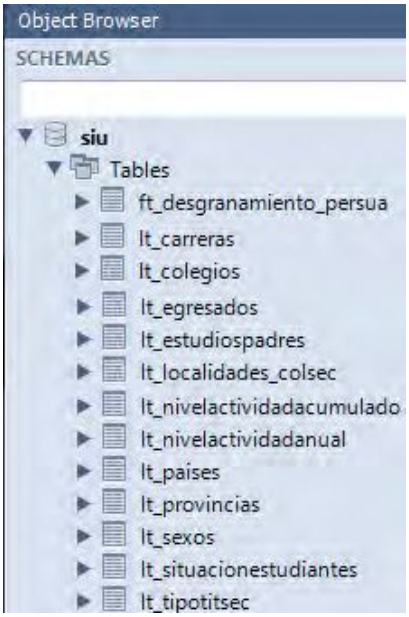


Figura 7.3 – Base de datos SIU en ambiente local.

7.1.2 Descripción de los datos recolectados

En la siguiente tabla, se describen los datos del Cubo 04 – Desgranamiento.

Atributo	Descripción
Año académico	Año académico, en que se evalúa la actividad de las personas. Esta dimensión es obligatoria en las consultas, ya que es parte de la definición de la medida.
Nivel de actividad anual	Se define en base a la cantidad de actividad realizada durante el año académico correspondiente, en la unidad académica. Está referido a la cantidad de exámenes rendidos, promociones y equivalencias (no importa el resultado de los mismos). Es decir todo lo que refleje intención de aprobar una materia.
Nivel de actividad acumulado	Se define en base a la cantidad de actividad realizada desde el ingreso a la unidad académica. Está referido a la cantidad de exámenes rendidos, promociones y equivalencias (no importa el resultado de los mismos). Es decir todo lo que refleje intención de aprobar una materia.
Cohorte	Año académico en el cual la persona se inscribe a la carrera en la que ingresa a la unidad académica. En caso de ser alumno de más de una carrera se considerará el año académico correspondiente a primera carrera en la que ingresa.
Sexo	Género de la persona.
Unidad Académica	La unidad académica a la que pertenece la carrera.
Carrera	Primera carrera en la que el estudiante figura como ingresante o alumno.
Situación del Estudiante	Se establecen cuatro categorías de estudiantes: los que nunca tuvieron actividad, los que tuvieron muy escasa actividad, los que habiendo tenido actividad nunca egresaron y los egresados.
Nivel de Estudios Padres	Indica el mayor nivel de estudios alcanzado por sus padres.
Colegio	Colegio secundario del que egresó el estudiante.
Localidad	Localidad en la que se encuentra el colegio secundario del que egresó el estudiante.
Provincia	Provincia a la que corresponde dicha localidad.
País	País al que corresponde la provincia.

En la siguiente tabla, se describen los datos de la tabla LT_Egresados.

Atributo	Descripción
Título	Título con el cual se recibe el estudiante.
Carrera	Carrera a la que pertenece el título obtenido por el estudiante.
Legajo	Legajo del estudiante.

Fecha de egreso	Fecha en que el estudiante obtiene el título.
Período de inscripción	Año de inscripción en la carrera, coincide con la cohorte de la tabla FT_Desgranamiento_PersUA.
Duración Carrera	Cantidad en meses que el estudiante tarda en obtener el título.

7.1.3 Exploración y transformación de los datos

Se realizó un análisis de los datos obtenidos, con el objetivo de determinar las propiedades y los datos que formarán parte de la vista minable. Para este análisis, utilizó la funcionalidad de WEKA, que permite la inspección de los datos.
Fue importante determinar que ninguna de las tablas importadas contenía valores nulos, por lo cual no hubo que tomar ninguna decisión sobre estos valores.

7.1.3.1 Corte horizontal de FT_Desgranamiento_PersUA

Se realizan dos tipos de cortes horizontales sobre la tabla FT_Desgranamiento_PersUA, uno para obtener aquellos inscriptos en las carreras de licenciatura; y otro para obtener aquellos con cohorte mayor o igual al año 2000. La información filtrada se almacena en la tabla FT_desgranamiento_PersUA_Licenciatura. Esta tabla tiene un total de 37.497 registros, representando a 5.366 estudiantes y egresados de las carreras de Licenciatura de la Facultad de Informática de la UNLP.
El script correspondiente se puede ver en el *Apéndice A, 2. Creación y llenado de la tabla FT_Desgranamiento_PersUA_Licenciatura*.

7.1.3.2 Atributos de FT_Desgranamiento_PersUA_Licenciatura

La siguiente tabla, contiene los posibles valores de los atributos de la tabla FT_Desgranamiento_PersUA_Licenciatura.

Atributo	Valores posibles (en rango o por extensión)
Año académico	2000-2012.
Nivel de actividad anual	0; 1; 3; 6.
Nivel de actividad acumulado	0; 1; 6; 16; 26; 36.
Cohorte	2000 - 2012.
Sexo	1; 2.
Unidad Académica	33.
Carrera	LS; LI.
Situación del Estudiante	N; A; P; S.
Nivel de Estudios Padres	0-7; 12-13.
Tipo título secundario	S.
Colegio	Valor mínimo: 0; valor máximo 2600055. (650 valores diferentes)

7.1.3.3 Atributos Unidad Académica, Tipo Título Secundario y Carrera

Observamos, que los atributos *Unidad Académica* y *Tipo Título Secundario* contienen un valor constante en todos los registros, estos valores son 33 para el atributo *Unidad Académica*, que representa a la Facultad de Informática y 'S' para el atributo *Tipo Título Secundario*, cuyo significado es "Sin Referencias". Estos valores no aportan información relevante para el análisis de los datos, por este motivo, no serán tenidos en cuenta en el momento de la creación de la vista minable.
Tampoco se tendrá en cuenta el atributo *carrera*, ya que no es relevante para el análisis, saber a qué licenciatura está inscripto el estudiante.

7.1.3.4 Atributos Colegio, creación del atributo Procedencia

El atributo *colegio* no se usará de manera directa, porque no nos interesa saber en qué colegio terminaron sus estudios secundarios los alumnos; pero es un atributo importante, porque de él podemos saber la localidad de la cual son originarios los estudiantes.
Se asume que la localidad de la cual proviene un estudiante, es la misma a la cual pertenece el colegio en donde terminó sus estudios secundarios.
Se creó, un nuevo atributo, *procedencia*, que indica la zona geográfica del país de la cual es originario el alumno, ó si proviene del exterior del país. Este atributo toma los siguientes valores posibles.

Valor del atributo.	Descripción.
La Plata y alrededores.	Agrupa las localidades de La Plata, Berisso, Ensenada, Brandsen, Magdalena, Berazategui, Florencio Varela, San Vicente. Corresponden al 59.39% de los estudiantes.
Gran La Plata.	Agrupa las localidades no limítrofes a la ciudad de La Plata, que se encuentran a una distancia NO mayor de 150 Km. Corresponden al 10.04% de los estudiantes.
Interior de la provincia de Buenos Aires.	Agrupa las localidades que se encuentran en la provincia de Buenos Aires, a una distancia mayor a 150 Km. Corresponden al 11.44 % de los estudiantes.
Interior del país.	Agrupa todas las localidades de la República Argentina que se encuentran fuera de la provincia de Buenos Aires. Corresponden al 8.29% de los estudiantes.
Exterior	Agrupa los registros que tienen un código de colegio que pertenece a un país con nombre "Otros". Corresponden al 2.18% de los estudiantes.
Sin Datos.	Agrupa todos los registros con alguna de las siguientes condiciones: Código de colegio no válido. Código de colegio con localidad "indeterminado", de la Pcia. de Bs. As. Son el 12.00% de los estudiantes.

Para más detalles del proceso de creación del atributo *procedencia*, consultar el *Apéndice A, 3. Atributo procedencia*.

7.1.3.5 Atributo Situación del Estudiante

La siguiente tabla contiene la descripción de los valores del atributo *situación del estudiante*.

Valor del atributo.	Descripción.
Nunca tuvo actividad (A)	Para los casos de los alumnos que no tienen ninguna materia en la historia académica.
Tuvo escasa actividad (P)	El alumno tiene hasta 5 materias en su historia académica.
No egreso (N)	El alumno tiene 6 ó más materias en su historia académica y no egresó aún.
Egresado (S)	Alumno egresado.

Con las categorías de alumnos existentes, se pueden tener alumnos tanto activos como pasivos dentro de las categorías P y N; además la categoría S, no determina el tiempo medio en que el alumno concluyó sus estudios. Se propone entonces, la siguiente división de categorías de este atributo.

Nuevo valor del atributo.	Descripción.
Activos (A)	Alumnos con actividad en todos los años de carrera. Indica que podrían terminar la carrera en un tiempo promedio.
Pasivos (P)	Alumnos sin actividad en su historia académica ó tienen más años sin actividad que con actividad.
Egresado en término (ET)	Alumnos que concluyeron la carrera en un tiempo aproximado al establecido en el plan de carrera (se toman 7 años promedio).
Egresado fuera de término (EFT)	Egresó en un tiempo mucho mayor al establecido por el plan de carrera (más de 7 años).

El script de transformación se encuentra en el *Apéndice A, 4. Atributo Situación del Estudiante*.

7.1.3.6 Atributo Nivel Estudio Padres

Este atributo, indica el último nivel de estudios alcanzado por los padres del alumno. Los valores posibles son los siguientes:

Código	Descripción
1	No hizo estudios.
2	Escuela primaria incompleta.
3	Escuela primaria completa.
4	Colegio secundario incompleto.
5	Colegio secundario completo.
6	Estudios Universitarios o superiores incompletos.
7	Estudios Universitarios o superiores completos.
12	Estudios de postgrado.

Se propone el siguiente agrupamiento, indicando el último nivel completo alcanzado.

Categorías actuales	Nuevas categorías
No hizo estudios o Escuela primaria incompleta.	Sin estudios. Representa el 42.84% de los estudiantes.
Escuela primaria completa ó Colegio secundario incompleto.	Estudios Primarios. Representa el 10.26% de los estudiantes.
Colegio secundario completo ó estudios Universitarios o superiores Incompletos	Estudios Secundarios. Representa el 20.38% de los estudiantes.
Estudios Universitarios o superiores completos ó Estudios de postgrado.	Estudios Universitarios. Representa el 25.75% de los estudiantes.
Sin datos.	Registros con valor original 0 ó 13. Estos códigos no son valores válidos de LT_EstudiosPadres. Representa el 0.74% de los estudiantes.

Los scripts de transformación del atributo *nivel estudio padres*, se puede consultar en el *Apéndice A, 5. Atributo Nivel Estudio Padres*.

7.1.3.7 Atributos Sexo, creación del atributo Género

El atributo sexo indica el género del estudiante. Las descripciones de los valores posibles para este atributo son 0 y 1, indicando el género masculino y femenino respectivamente. Se propone cambiar este atributo por el atributo género, cuyos valores posibles serán 'Masculino' y 'Femenino'; obteniendo una mejorar la comprensión de los valores en el momento de la lectura de los resultados.

Los scripts de transformación del atributo *sexos*, se pueden consultar el *Apéndice A, 6. Atributo Género*.

7.1.3.8 Valores de los atributos de la tabla LT_Egresados

De la tabla LT_Egresados, se exportaron los datos referentes a las carreras de licenciatura. Los posibles valores de sus atributos son los siguientes.

Atributo	Valores posibles (en rango o por extensión)
Título	LS; LI.
Carrera	LS; LI.
Período de inscripción	2000 – 2008. Tener en cuenta que si hubo un cambio de carrera, el período de inscripción indica el año en que se realizó el cambio.
Duración de la Carrera	12-145. Los valores pequeños es debido a lo explicado en el atributo período de inscripción.
Fecha de Egreso	El valor más pequeño corresponde al año 2004, el valor mayor corresponde al año 2012.

7.1.4 Vista minable

Después de las tareas de selección, preprocesamiento y transformación de datos recolectados, la vista minable para el análisis de los datos de los estudiantes y egresados, desde el punto de vista socio demográfico, queda formada por los siguientes atributos:

Atributo	Descripción
Género.	Género del estudiante o egresado.
Nivel de estudios padres.	Indica el mayor nivel de estudios alcanzado por sus padres.
Procedencia.	Zona geográfica del país de la cual es oriundo el estudiante. Está relacionado con la cercanía que tiene a la unidad académica.
Situación del estudiante.	Se establecen cinco categorías de estudiantes: activos, activos con escasa actividad, pasivos, egresado en término y los egresados fuera de término.

7.2 Análisis de participación en las materias

En esta sección, se describe la creación de la vista minable, para el análisis de los datos relacionados a las acciones de los estudiantes, que lo lleven a aprobar la materia en la que se encuentran inscriptos. Las acciones que se tendrán en cuenta para este análisis son la participación social que los alumnos tienen con sus compañeros y docentes en las materias que cursan y el uso de material bibliográfico relacionado a la materia que hacen de la biblioteca de la Facultad.

Las materias analizadas, fueron las que utilizan el sistema Moodle como plataforma de aprendizaje. No todas las materias de la Facultad utilizan Moodle, es por eso que el

análisis se realiza sobre a un grupo de materias de la Facultad. Se decidió analizar estas materias ya que la tesina se enfoca principalmente en el uso de herramientas de código abierto.

Nota: Tener en cuenta que, el término materias, se hará referencia a este subconjunto de materias de la Facultad de Informática de la UNLP.

7.2.1 Exportación e importación de los datos

En este análisis, se lleva a cabo uno de los principales objetivos de este trabajo de grado, que es la correlación de datos de diferentes sistemas. Esta correlación se va a ver en el hecho que se utilizaron principalmente, los datos de la base de datos del sistema Moodle; pero también fueron necesarios datos de la base de datos del sistema SIU-Guaraní y el sistema de bibliotecas Merán.
Como se verá más adelante, esta correlación es posible realizarla porque los tres sistemas utilizan identificadores únicos para los estudiantes, como son el DNI ó el legajo; si bien estos datos no son visibles en los análisis, son muy importantes para la realización de la vista minable cruzando datos de los tres sistemas. También se tuvieron que correlacionar datos pertenecientes a las materias de los tres sistemas; esto fue posible mediante un código de materia que se almacenan en el sistema SIU-Guaraní y en el sistema Moodle.

7.2.1.1 Exportación e importación de los datos de Moodle

La base de datos de Moodle está formada por más de 300 tablas. Nos interesan aquellas con información relacionada a la interacción de los estudiantes; lo que redujo a 14, el número de tablas utilizadas.
Con reingeniería y documentación [Base de datos Moodle], se formó el diagrama de entidad relación que se muestra en la Figura 7.4 – Estructura de datos de Moodle.

A diferencia de la exportación e importación de datos del sistema SIU-Guaraní, descripta en la sección *7.1.1 – Exportación e importación de los datos*; la base de datos del sistema Moodle se exportó e importó en su totalidad.
La Figura 7.5 – Base de datos *catedras* en ambiente local, muestra la estructura de tablas de la base de datos en el ambiente local. Solamente se muestra el conjunto de tablas usadas para este análisis.
Los scripts de exportación/importación de los datos de la base de datos *catedras*, se pueden consultar en el *Apéndice A, 7. Importación de la base de datos catedras al ambiente local*.

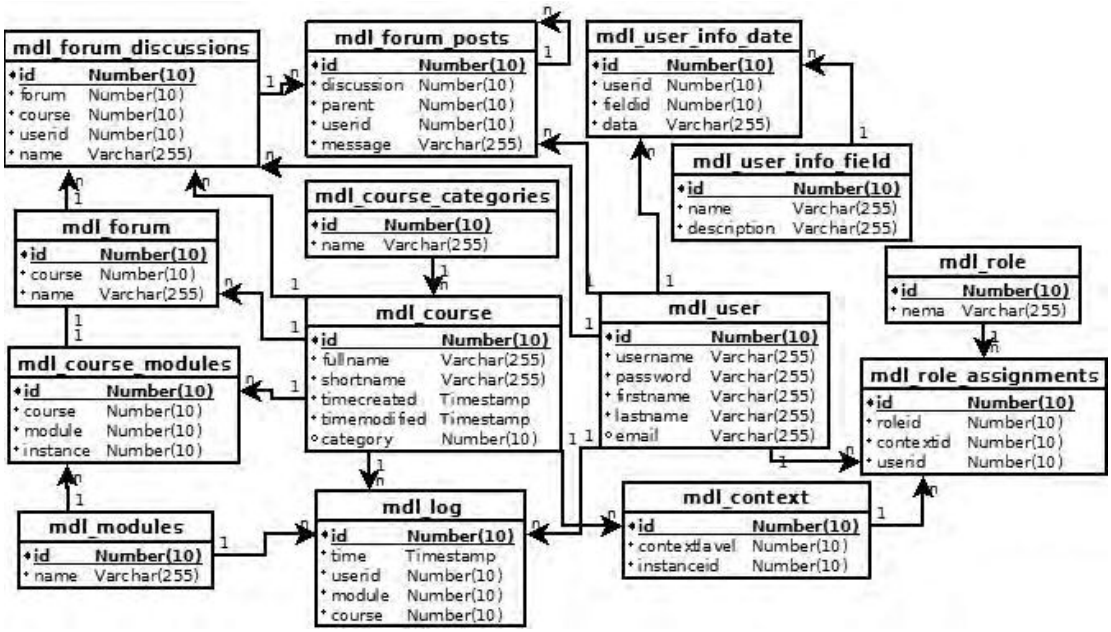


Figura 7.4 – Estructura de datos Moodle.

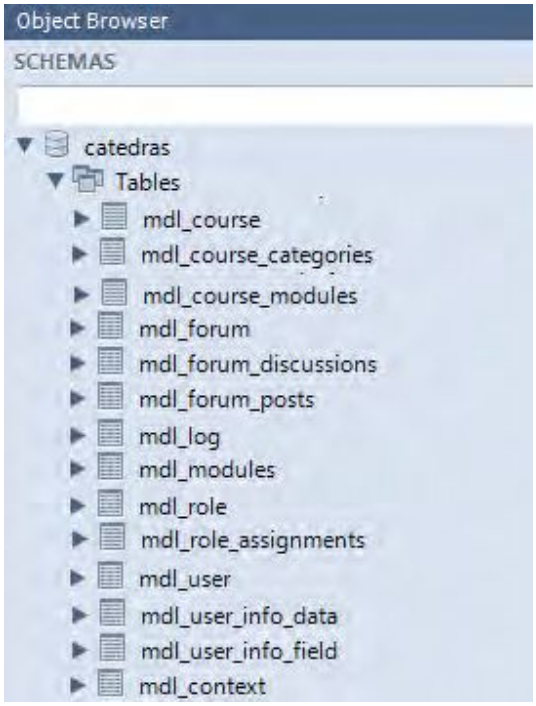


Figura 7.5 – Base de datos *catedras* en ambiente local.

El sistema Moodle ofrece diversos recursos de comunicación entre docentes y alumnos, entre los que se encuentran foros, Wikis, Chat, blogs. Se decidió utilizar los datos de la participación de los estudiantes en foros ya que es el recurso de comunicación más utilizado entre las diferentes materias.

Esta información se obtuvo con el script disponible en el *Apéndice A, 8. Recursos más usados por las materias en Moodle*.

7.2.1.2 Exportación e importación de los datos de SIU-Guaraní

Para este análisis, fue necesario obtener también, datos del sistema SIU-Guaraní. Con el cruce de estos datos, se pudo determinar los cursos de Moodle correspondientes a materias de la Facultad, y el resultado de los estudiantes en las materias. En la figura 7.6 – Datos del SIU-Guaraní para el análisis de participación social en las materias, se muestran las estructuras de las tablas del SIU-Guaraní exportadas para este análisis.

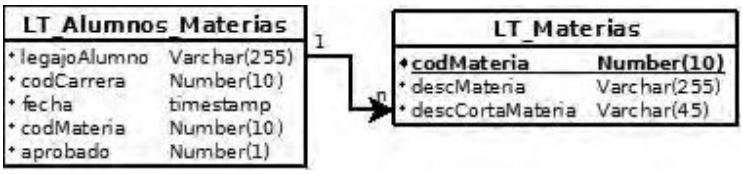


Figura 7.6 – Datos del SIU-Guaraní para el análisis de participación social en las materias.

Los scripts de exportación/importación de estas tablas se pueden consultar en el *Apéndice A, 9. Tablas del SIU-Guaraní para el análisis de participación social en las materias*.

7.2.1.3 Exportación e importación de los datos de Merán

Los datos disponibles de Merán, fueron otorgados por responsables de este sistema, quienes realizaron las consultas correspondientes con las condiciones pedidas. Extrajeron los datos en un archivo .csv, los cuales fueron luego importados a una tabla en el ambiente local. En la Figura 7.7 – Estructura de datos extraídos de Merán, se puede observar la estructura de los datos extraídos.

Figura 7.7 – Estructura de datos extraídos de Merán.

Los scripts de exportación/importación de estas tablas se pueden consultar en el *Apéndice A, 10. Importación de los datos del sistema Merán*.

7.2.2 Descripción de las tablas y datos seleccionados

En la siguiente tabla, se describen las tablas seleccionadas del sistema Moodle.

Tabla	Descripción
mdl_course	Son los cursos dados de alta en el sistema. Para cada materia, se da de alta un curso por año académico.
mdl_modules	Representa las actividades del sistema, Chat, foro, resource, URL, Workshops. Nos interesa para este trabajo, el módulo *forum*.
mdl_course_modules	Son las actividades creadas por curse. Nos interesan especialmente los registros *module = 7* –módulo *forum* –. La propiedad *instance* contiene una clave foránea a la tabla *mdl_forum*.
mdl_course_categories	Los cursos en Moodle están agrupados en una categoría.
mdl_context	El concepto de contexto se utiliza para el chequeo de permisos, los roles en Moodle se asignan en un contexto. Nos interesa el valor *contextlevel = 50*, –contexto de un curso –. La propiedad *instanteid* contiene una clave foránea a la tabla mdl_course.
mdl_user	Contiene todos los usuarios dados de alta en el sistema.
mdl_role	Los posibles roles que se les pueden asignar a los usuarios en algún contexto. Nos interesa el rol 5 –rol de *estudiante* –.
mdl_role_assignments	Contiene la relación entre el usuario, el rol y el contexto. Nos interesa las asignaciones del rol estudiante (*roleid = 5*) en contexto de un curso (*contextlevel = 50*).
mdl_forum	Contiene todos los foros creados en el sistema.
mdl_forum_discussion	Contiene todas las conversaciones de los foros.
mdl_forum_posts	Contiene los diferentes post de las conversaciones de los foros.
mdl_log	Contiene el log de toda la actividad en el sistema. Nos interesan las actividades de los estudiantes en los cursos, principalmente en el modulo *forum*.
mdl_user_info_field	Contiene propiedades de datos personales de los usuarios adicionales a la tabla *mdl_users*. Por ejemplo DNI, legajo, twitterusername. Nos interesa el legajo del alumno.
mdl_user_info_data	Contiene para cada usuario el dato concreto de la propiedad *mdl_user_info_field*.

En la siguiente tabla, se describen las tablas del sistema SIU-Guaraní utilizadas para el análisis de la participación de los estudiantes en las materias.

Tabla	Descripción
LT_materias	Contiene todas las materias dictadas en la Facultad de Informática.
LT_alumnos_materias	Contiene la información de las materias que aprobaron cada uno de los estudiantes. Si la relación no se encuentra en esta tabla, se asume el alumno no aprobó la misma.

En la siguiente tabla, se describen los datos otorgados del sistema Merán almacenados localmente en la tabla tesis_merán.

Atributo	Descripción
apellido	Apellido del usuario que realizó la operación en el sistema.
nombre	Nombre del usuario que realizó la operación en el sistema.

DNI	El DNI del usuario que realizó la operación en el sistema.
tipo_operacion	El tipo de operación que realizó el usuario en el sistema. Puede tomar los siguientes valores: préstamo, reserva, devolución, renovación, espera.
fecha	La fecha en que el usuario realizó la operación en el sistema.
materias	Lista separada por comas, de las diferentes materias a las cuales pertenece la bibliografía sobre la que realizó la operación el usuario en el sistema.

7.2.3 Exploración de los datos

Una vez importados todos los datos necesarios, se realizó un análisis inicial de los mismos, con el objetivo de determinar las propiedades y los registros que formarán parte de la vista minable.

7.2.3.1 Selección de los cursos

No todos los cursos de Moodle, se corresponden a materias de la Facultad de Informática; por ejemplo, podemos encontrar "Escuela de CACIC2008", "Curso de Verano - AyED - 2010" entre otros. Solamente nos interesan aquellos cursos que correspondan a materias de los planes de estudios de las carreras dictadas en la Facultad de Informática. Para poder obtener estos datos, se realizó una correlación entre las tablas *mdl_course* y *Lt_materias*, creando la tabla *mdl_tesis_course_fac_informatica*.

La correlación con ambas tablas se hicieron principalmente mediante los atributos codMateria de *LT_materias*, y shortname de *mdl_course*; aunque hubo casos en que estos códigos no coincidían y se tuvo que correlacionar mediante el nombre de la materia.
Se descartaron las materias correspondientes a las sedes de Tres Arroyos ó Las Flores; quedando un total de 354 cursos válidos.

Los script de creación y llenado de la tabla *mdl_tesis_course_fac_informatica* se pueden consultar el *Apéndice A, 11. Creación y llenado de la tabla mdl_tesis_course_fac_informatica*.

7.2.3.2 Alumnos inscriptos en cada materia

Una vez seleccionados los cursos de Moodle, que se corresponden con materias de la Facultad de Informática, se obtuvieron los alumnos inscriptos a los mismos.
Para almacenar estos datos se creó una tabla auxiliar, *mdl_tesis_courses_students*, que contiene la información de los alumnos y las materias que cursaron. La estructura de esta tabla se puede ver en la Figura 7.7 – Tabla *mdl_tesis_courses_students*.

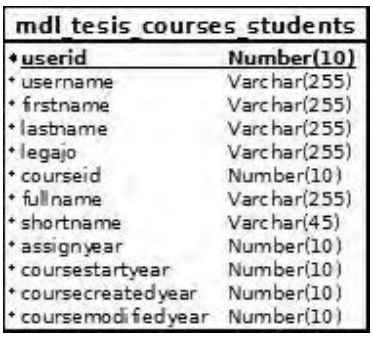


Figura 7.7 – Tabla *mdl_tesis_courses_students*.

Determinar los alumnos que cursaron una materia es posible mediante el cruce de información de las tablas *mdl_context* – el contexto de curso es el 50–; *mdl_role_assignments* –el rol de estudiante es el 5 –; *mdl_user* y *mdl_tesis_course_fac_informatica*.

Los scripts de creación y llenado de la tabla *mdl_tesis_courses_students*, se pueden consultar en el *Apéndice A, 12. Creación y llenado de la taba mdl_tesis_courses_students*.

7.2.3.3 Resultados obtenidos en las materias

Para determinar si el alumno aprobó o no la materia que cursó, se realizó un cruce de información de las tablas *mdl_tesis_courses_students* y *lt_alumnos_materias*. Se creó otra tabla llamada *mdl_tesis_course_students_note*, que contiene datos del alumno, la materia que el alumno curso, el año en que la curso y la información de si aprobó o no la materia. La Figura 7.8 – Tabla auxiliar *mdl_tesis_course_students_note*, muestra la estructura de esta tabla auxiliar.

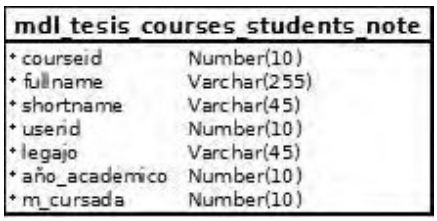


Figura 7.8 – Tabla auxiliar *mdl_tesis_course_students_note*.

Los scripts de creación y llenado de la tabla *mdl_tesis_course_students_note*, se pueden consultar el *Apéndice A, 13. Creación y llenado de la taba mdl_tesis_course_students_note*.

7.2.3.4 Participación en los foros de las materias

Para el análisis de la participación social de los estudiantes en las materias; de cada alumno y materia, uno se obtuvo la siguiente información:

- Participación total en los foros, esto es, cantidad de veces que el alumno realizó alguna acción en Moodle relacionado con los foros de la materia.
- Cantidad de respuestas que el alumno realizó a post de compañeros del curso.
- Cantidad de respuestas que el alumno realizó a post de docentes del curso.
- Cantidad de conversaciones iniciadas por el alumno en los foros de la materia.
- Tiempo total –en minutos –, que el alumno le dedicó a los foros de la materia.
- Tiempo total –en minutos –, que el alumno le dedicó mediante Moodle, a la materia.

Para el cálculo de estos datos, se utiliza la información de la tabla *mdl_log*, junto con los datos de las tablas auxiliares que se crearon anteriormente.
Para que la realización de la vista minable sea más eficiente, se decide guardar los datos de las diferentes participaciones en los foros listadas anteriormente, en tablas separadas; las estructuras de estas tablas, se muestran en la Figura 7.9 – Tablas auxiliares de participación en foros.

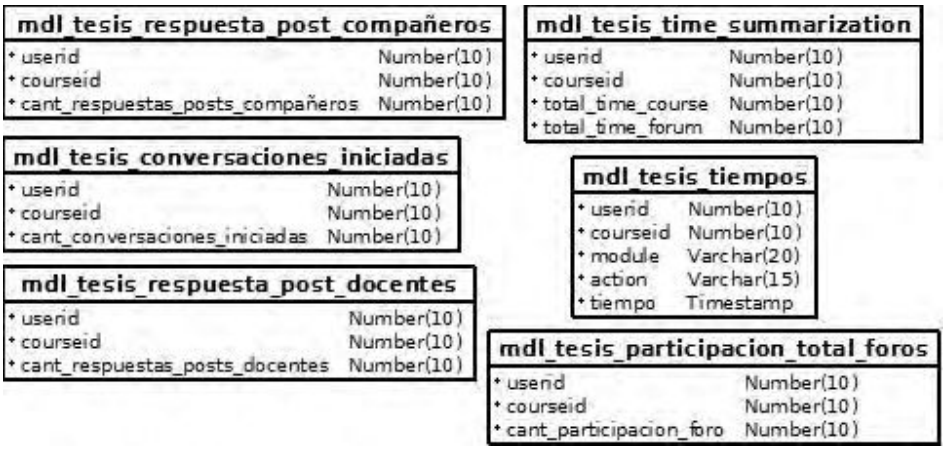


Figura 7.9 – Tablas auxiliares de participación en foros.

Se detalla a continuación la información que cada una de estas tablas almacena.

Tabla	Descripción
mdl_tesis_participacion_total_foros	Contiene la cantidad de participaciones en los foros de las materias del alumno.
mdl_tesis_respuesta_post_compañeros	Contiene la cantidad de respuestas del alumno a compañeros de clase.
mdl_tesis_respuesta_post_docentes	Contiene la cantidad de respuestas del alumno a docentes de la materia.
mdl_tesis_conversaciones_iniciadas	Contiene la cantidad de conversaciones iniciadas por el alumno en foros de la materia.
mdl_tesis_tiempos	Contiene el tiempo en minuto de las acciones de los

	alumnos sobre las materias que cursan, esta información se obtiene de la tabla mdl_log. Estos datos se usarán luego, para el cálculo de la cantidad de tiempo empleado por el alumno en la materia y en los foros de las mismas.
mdl_tesis_time_summarization	Contiene la suma el tiempo total empleado por los estudiantes en las materias que cursaron y en los foros de las mismas.

Se decidió pasar los valores numéricos de los tiempos a valores discretos, que indiquen la frecuencia –alto, medio ó bajo –, con la que el estudiante hace uso de las tareas de foro de las materias. Cada atributo tomará los siguientes valores: Alto, Medio ó Bajo.

Los scripts de creación y llenado de estas tablas se pueden consultar en el *Apéndice A, 14. Participación de los alumnos en los foros*.

7.2.3.5 Uso de biblioteca

A partir de los datos del sistema Merán, se puede obtener información acerca de si el alumno hizo o no, uso de material de la biblioteca, relacionado a las materias que cursó. Para esto se procesó la información obtenida del sistema Merán para poder relacionarla de manera más fácil con la información obtenida del sistema Moodle.
Se representó, a partir de los datos obtenidos, la información del uso de la biblioteca relacionado con los estantes virtuales, separando la información en dos tablas que se pueden observas en la Figura 7.10 – Tablas auxiliares del uso de bibliotecas.

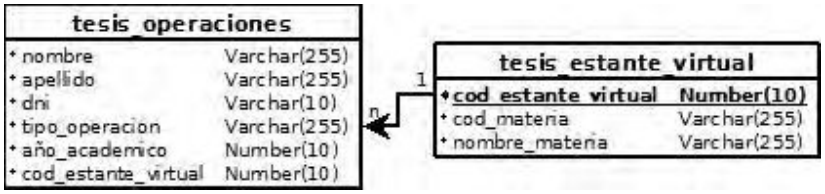


Figura 7.10 – Tablas auxiliares del uso de bibliotecas.

Se detalla a continuación la información de cada una de estas tablas.

Tabla	Descripción
Tesis_operaciones	Operaciones del usuario sobre bibliografía perteneciente a la biblioteca. Relacionado con el estante virtual al que pertenece dicha bibliografía y el año académico en que se produjo la operación.
Tesis_estante_virtual	Es el estante virtual, que relaciona el material de la biblioteca con las materias que se dictan en la Facultad. También contiene el código de materia, obtenido de las materias del sistema Moodle y SIU-Guaraní.

Luego de la separación de los datos obtenidos del sistema Merán, se asocia la misma con los datos correspondientes del sistema Moodle; los datos de esta asociación se guarda en la tabla tesis_uso_bibliotecas, para luego hacer más fácil agregar la información del uso de bibliotecas en la creación de la vista minable.

Los scripts de creación y llenado de esta tabla se pueden consultar en el *Apéndice A, 15. Uso de biblioteca*.

7.2.4 Vista minable

La vista minable para la realización de este análisis, queda formada por los siguientes atributos.

Atributo	Descripción
Participación en Foros	La cantidad de veces que el usuario participó en alguno de los foros de la materia. Puede tomar los valores Bajo, Medio, Alto.
Respuestas a Compañeros	La cantidad de veces que el usuario respondió algún mensaje posteado por un compañero de curso.
Respuestas a Docentes	La cantidad de veces que el usuario respondió algún mensaje posteado por algún docente de la materia.
Conversaciones Iniciadas	La cantidad de veces que el usuario inició una conversación en alguno de los foros de la materia.
Tiempo Total Empleado en Foros	El tiempo en minutos empleado por el usuario en los foros de la materia.
Tiempo Total Empleado en la Materia	El tiempo total empleado por el usuario en la materia mediante el sistema Moodle.
Uso de bibliotecas	Indica si el estudiante hizo uso o no de bibliografía (extraída de la biblioteca) referente a la materia el año que la curso.
Aprobado	Indica si el usuario aprobó o no la materia.

Para mayor detalle sobre la creación de la vista minable, consultar el Apéndice *A, 16. Vista Minable*.

Capítulo 8
MODELADO Y EVALUACIÓN

En este capítulo, se describen las etapas de minería de datos del proceso de KDD, para los análisis propuestos en el *Capítulo 6 – Presentación del problema*.

8.1 Análisis socio-demográfico

En esta sección, se presenta el análisis de los datos desde el punto de vista de factores sociales y demográficos, para intentar establecer si existe una relación entre éstos y la situación académica de los estudiantes.

8.1.1 Distribución de los datos

En los histogramas de la Figura 8.1 – Distribución de los datos, vemos la distribución de los atributos con respecto al atributo Situación del estudiante. El atributo situación del estudiante será la clase a predecir en la tarea de clasificación.

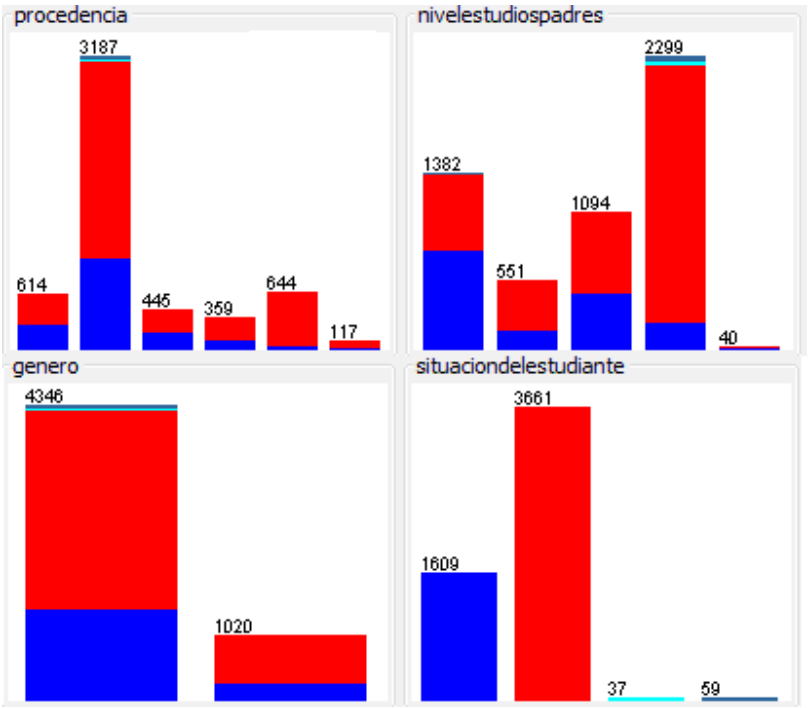


Figura 8.1 – Distribución de los datos.

En la siguiente tabla se indican los valores de los atributos, de izquierda a derecha.

Atributos	Posibles valores (de izquierda a derecha según la ubicación en los histogramas)
Procedencia	Interior de la provincia de Buenos Aires; La Plata y alrededores; Interior del país; Gran La Plata; Sin Datos; Exterior.
Nivel estudio de los padres	Universitarios; Primarios; Secundarios; Sin estudios; Sin datos.
género	Masculino; Femenino.
Situación del estudiante	Activo; Pasivo; Egresado en término; Egresado fuera de término.

En la distribución de los estudiantes según la gama de colores, en cada una de las barras de los atributos procedencia, nivel estudio de los padres y género es, desde la base de la barra hacia arriba: Activo (Azul), Pasivo (Rojo), Egresado en término (Verde) y Egresado fuera de término (Gris oscuro).

Las clases no están balanceadas, siendo P-Pasivo la clase mayoritaria. Ningún atributo separa bien las clases.
Los histogramas muestran, una distribución más equilibrada de estudiantes activos y pasivos, en los procedentes del interior de la provincia de Buenos Aires y del interior de país. En las demás procedencias, se observa una clara diferencia con más estudiantes pasivos que activos.
También la distribución entre estudiantes activos y pasivos es más pareja para los niveles de estudio de los padres más altos –universitarios y secundarios –, habiendo en los demás valores una clara diferencia de pasivos frente a activos.

8.1.2 Clasificación – Árbol de decisión

Como se describió en al *Capítulo 6 – Presentación del problema*; se ejecutará el algoritmo J48, con diferentes valores en el parámetro de factor de confianza de la poda, para comparar resultados.
Primero ejecutamos el algoritmo con el valor por defecto del parámetro. El resultado se puede observar en la Figura 8.2 – Algoritmo J48, con los parámetros por defecto.

```
=== Run information ===
J48 pruned tree
------------------
nivelestudiospadres = Estudios Universitarios
|   procedencia = Int. de la Pcia. de Buenos Aires: A - Activo (202.0/80.0)
|   procedencia = La Plata y Alrededores: A - Activo (881.0/385.0)
|   procedencia = Interior del país: A - Activo (164.0/67.0)
|   procedencia = Gran La Plata: P - Pasivo (90.0/41.0)
|   procedencia = Sin Datos
|   |   genero = M: P - Pasivo (13.0/5.0)
|   |   genero = F: A - Activo (2.0)
|   procedencia = Exterior: P - Pasivo (30.0/9.0)
nivelestudiospadres = Estudios Primarios: P - Pasivo (551.0/159.0)
nivelestudiospadres = Estudios Secundarios
|   procedencia = Int. de la Pcia. de Buenos Aires: A - Activo (149.0/68.0)
|   procedencia = La Plata y Alrededores: P - Pasivo (692.0/276.0)
|   procedencia = Interior del país: P - Pasivo (124.0/48.0)
|   procedencia = Gran La Plata: P - Pasivo (82.0/30.0)
|   procedencia = Sin Datos: A - Activo (11.0/4.0)
|   procedencia = Exterior: P - Pasivo (36.0/8.0)
nivelestudiospadres = Sin estudios: P - Pasivo (2299.0/289.0)
nivelestudiospadres = Sin datos: P - Pasivo (40.0/18.0)

Number of Leaves  :     16
Size of the tree :      20

=== Summary ===
Correctly Classified Instances        3838               71.5244 %
Incorrectly Classified Instances      1528               28.4756 %
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Class
  0.468     0.158    A - Activo
  0.843     0.548    P - Pasivo
  0         0        ET - Egreso en término
  0         0        EFT - Egreso fuera de término
=== Confusion Matrix ===
   a    b    c    d   <-- classified as
 753  856    0    0 |    a = A - Activo
 576 3085    0    0 |    b = P - Pasivo
   5   32    0    0 |    c = ET - Egreso en término
  12   47    0    0 |    d = EFT - Egreso fuera de término
```

Árbol Generado

Tamaño del árbol.

Precisión del clasificador

Detalles por clase.

Matriz de confusión.

Figura 8.2 – Algoritmo J48, con los parámetros por defecto.

Se generó un árbol con un tamaño total de 20 nodos. El clasificador obtenido, tiene un 71.52% de precisión. La razón de aciertos de la clase P – Pasivo, es de 0.843, y la razón de aciertos de la clase A – Activo es de 0.462.

En la matriz de confusión, se observa que el modelo no clasifica a estudiantes egresados. Esto se debe a que la cantidad de registros que tenemos en el universo de datos usados para entrenar el modelo que cumplen con ser egresados, es muy pequeña en comparación con la cantidad total de datos del universo –96 egresados en un universo de

más de 5300 estudiantes – entonces, el clasificador no puede asignar, con precisión, registros a esta clase.

El modelo generado, indica que el atributo que mejor clasifica a los estudiantes, es el nivel de estudio de los padres.
Clasifica con la clase P - Pasivo a los estudiantes cuyos padres no tienen estudios realizados, tienen estudios primarios o se desconoce el nivel de estudio. Sin importar la procedencia de los estudiantes.
Los estudiantes cuyos padres tienen nivel de estudio secundario, son clasificados también como pasivos salvo aquellos cuya procedencia es el interior de la provincia de Buenos Aires o se desconoce la procedencia.
Los estudiantes cuyos padres tienen un nivel de estudio universitario, son clasificados mayoritariamente con la clase A - Activo, salvo aquellos con procedencia de Gran La Plata, ó los de género masculino a los cuales se le desconoce la procedencia; quienes son clasificados con la clase P - Pasivo.

Se ejecutará el algoritmo modificando el valor del parámetro factor de confianza de la poda, para ver si se obtiene un modelo más simple, sin sacrificar la precisión; ó más complejo que mejore significativamente su precisión.

Con el factor de confianza = 0.001, obtenemos los siguientes resultados en cuanto al tamaño del árbol y la precisión del mismo:

```
Number of Leaves  :     5
Size of the tree :      6
=== Summary ===
Correctly Classified Instances     3839  71.543  %
Incorrectly Classified Instances   1527  28.457  %
```

Figura 8.3 – Algoritmo J48, con factor de confianza = 0.001.

Vemos que el tamaño del árbol se reduce considerablemente, con un tamaño total de 6 nodos, y la precisión del mismo mejora aunque no considerablemente.

Con el factor de confianza = 1.0 en la poda, obtenemos los siguientes resultados:

```
Number of Leaves  :     22
Size of the tree :      28
=== Summary ===
Correctly Classified Instances    3850  71.748  %
Incorrectly Classified Instances  1516  28.252  %
```

Figura 8.4 – Algoritmo J48, con factor de confianza = 1.

En este caso obtenemos un modelo un poco más complejo que el primero, con un aumento en la precisión no muy significativa.

En los modelos generados, observamos una relación entre el nivel de estudios de los padres y la situación del estudiante, siendo más activos aquellos estudiantes cuyos padres poseen estudios secundarios o universitarios.

Los modelos generados con las modificaciones del parámetro confianza de la poda, se pueden ejecutar y ver desde el aplicativo que se entrega junto con este informe.

8.1.3 Agrupamiento

Como se describió en al *Capítulo 6 – Presentación del problema*; se ejecutará el algoritmo SimpleKMeans para la tarea de agrupamiento. Se realizarán diferentes ejecuciones formando diferentes cantidades de grupos para comparar los resultados entre ellos.

Al ejecutar el algoritmo para formar 3 clusters, obtenemos el siguiente resultado:

```
kMeans
======
Cluster centroids:
                                    Cluster#
Attribute                                0                    1                           2
                                    (3883)               (1254)                      (229)
=============================================================================================================
procedencia            La Plata y Alrededores  La Plata y Alrededores Int. de la Pcia. de Buenos Aires
nivelestudiospadres              Sin estudios Estudios Universitarios              Estudios Primarios
genero                                      M                       M                           M
situaciondelestudiante             P - Pasivo             A - Activo                  A - Activo
=== Model and evaluation on training set ===
  Clustered Instances
   0      3883 ( 72%)
   1      1254 ( 23%)
   2       229 (  4%)
```

Figura 8.5 – Algoritmo SimpleKMeans con 3 clusters.

Observamos, que el primer cluster, agrupa al 72% de la población y contiene en su mayoría a estudiantes masculinos, cercanos a al Facultad, con padres sin estudios, y pasivos. Vemos en el segundo de los grupos a estudiantes también cercanos a la Facultad cuyos padres tienen estudios universitarios y son activos. El tercero de los clusters, contiene en su mayoría a estudiantes activos, procedentes del interior de la provincia de Buenos Aires cuyos padres realizaron estudios primarios.

Al ejecutar el algoritmo para formar 5 clusters, obtenemos el siguiente resultado

```
kMeans
======
Cluster centroids:
                                    Cluster#
Attribute                                0                    1                          2
                                    (3584)                (887)                      (229)
=============================================================================================================
procedencia                 La Plata y Alrededores  La Plata y Alrededores Int. de la Pcia. de Buenos Aires
nivelestudiospadres                  Sin estudios Estudios Universitarios              Estudios Primarios
genero                                          M                       M                               M
situaciondelestudiante                  P - Pasivo              A - Activo                      A - Activo

=== Model and evaluation on training set ===
                                                                        3                               4
  Clustered Instances                                               (369)                           (297)
                                                    ========================================================
   0      3584 ( 67%)                              La Plata y Alrededores Int. de la Pcia. de Buenos Aires
   1       887 ( 17%)                                Estudios Secundarios          Estudios Universitarios
   2       229 (  4%)                                                   M                               M
   3       369 (  7%)                                          A - Activo                      P - Pasivo
   4       297 (  6%)
```

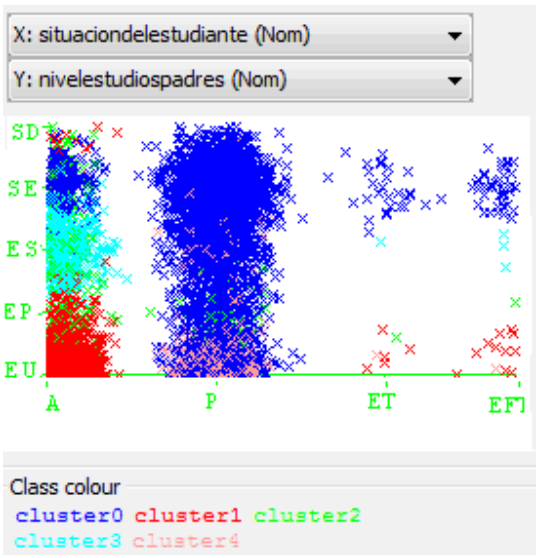
Figura 8.6 – Algoritmo SimpleKMeans con 5 clusters.

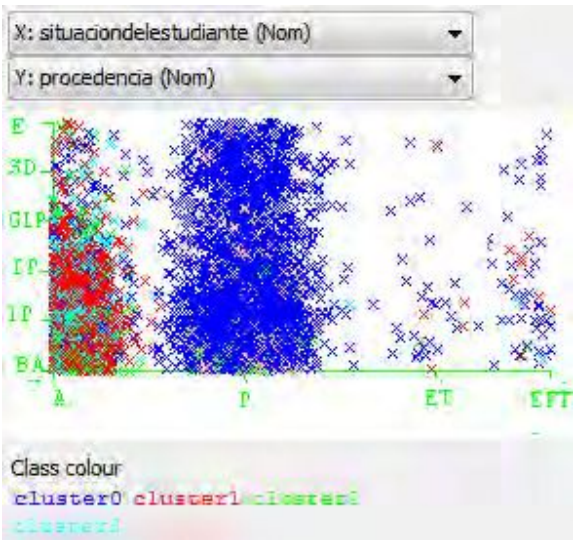
Vemos que los tres primeros clusters generados son iguales a los generados en la primera ejecución del algoritmo. El cluster 3 se caracteriza por estudiantes activos, cercanos a la Facultad, cuyos padres poseen estudios secundarios. El último de los clusters, contiene estudiantes pasivos procedentes del interior de la provincia de Buenos Aires cuyos padres poseen estudios universitarios.

Todos los clusters contienen, en su mayoría, a estudiantes procedentes de la provincia de Buenos Aires, más o menos cercanos a la Facultad, podemos concluir, con esta distribución que es muy alto el porcentaje de estudiantes que cursan en la Facultad de Informática procedentes de estas localidades.

8.1.3.1 Visualización de los grupos

La herramienta WEKA, nos permite visualizar gráficamente la distribución de las instancias en los diferentes clusters. En la siguiente figura, podemos observar la distribución de los estudiantes en los diferentes clusters según el atributo situación del estudiante, relacionado con los atributos nivel de estudio de los padres y procedencia.





A - Situación del estudiante y nivel de estudio padres B - Situación del estudiante y procedencia.

Figura 8.7 – Distribución de los datos según los atributos

Vemos, en la Figura 8.7 (A) que los estudiantes pasivos y los que tienen padres sin estudios realizados, se agrupan en su mayoría juntos, en este caso en el cluster 0 –segundo agrupamiento contando desde izquierda a derecha –. También se agrupan juntos los estudiantes egresados y activos cuyos padres tienen estudios universitarios o superiores –cluster 1 –. Por último, podemos ver, cómo en un mismo cluster se destacan los estudiantes activos cuyos padres tienen estudios secundarios –cluster 3 –.

En la Figura 8.7 (B) no se observa una separación entre los diferentes clusters tan clara como se observa en la Figura 8.7 (A); podemos señalar que los estudiantes que no son activos, en su mayoría se encuentran en un mismo cluster sin importar la procedencia de los mismos, en este caso el cluster 0; pero los estudiantes activos están más divididos entre los demás clusters.

8.2 Análisis de la participación en las materias

En esta sección se presenta el análisis de los datos referentes al comportamiento del estudiante en las materias que cursa, que lo lleve a aprobar la misma. Los datos obtenidos para este análisis fueron correlacionados de diferentes sistemas utilizados por la Facultad de Informática para almacenar los datos de sus estudiantes.

8.2.1 Distribución de los datos

En los histogramas de la Figura 8.8 – Distribución de los datos, podemos ver la distribución de los atributos con respecto al atributo clase. El atributo m_aprobado, será la clase a predecir en la tarea de clasificación.

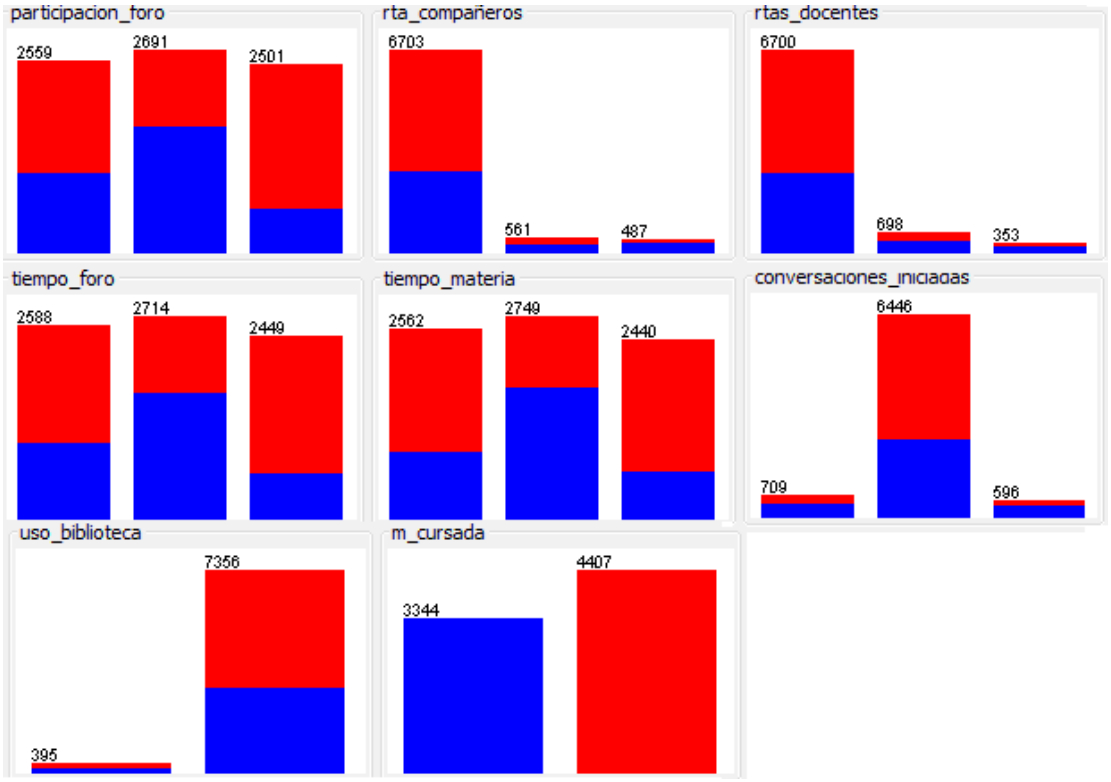


Figura 8.8 – Distribución del atributo aprobado sobre los demás atributos.

En la siguiente tabla se indican, de izquierda a derecha según la imagen, los posibles valores que pueden tomar los atributos.

Atributo	Valores posibles (de izquierda a derecha según la ubicación en los histogramas)
participacion_foro	Medio, Alto, Bajo
rta_compañeros	Bajo, Medio, Alto
rtas_docentes	Bajo, Medio, Alto
tiempo_foro	Medio, Alto, Bajo
tiempo_materia	Medio, Alto, Bajo
conversaciones_iniciadas	Medio, Bajo, Alto
uso_biblioteca	Sí, No
m_cursada	Aprobado (1), Desaprobado (0)

En la distribución de los estudiantes según la gama de colores, en cada una de las barras de los atributos participación_foro, rta_compañeros, rtas_docentes, tiempo_foro, tiempo_materia, conversaciones_iniciadas, uso_biblioteca es, desde la base de la barra hacia arriba: Aprobado (Azul), Desaprobado (Rojo).

Las clases no están balanceadas, siendo Desaprobado la clase mayoritaria. Ningún atributo separa bien las clases.
En los histogramas vemos, cuanta más participación tiene el estudiante en los foros –participación_foro –, cuanto más tiempo emplea en la navegación y uso de los foros y mayor tiempo emplea en las actividades virtuales de las materias en general, –tiempo_foro y tiempo_materia respectivamente –, mayor es la cantidad de estudiantes aprobados.
Se observa también, que una participación activa en los foros, como por ejemplo, la cantidad de conversaciones iniciadas o posts realizados, no influye en el resultado de la materia.
En cuanto al uso de bibliotecas, la cantidad de estudiantes que hicieron uso de la biblioteca es muy inferior en comparación a aquellos que no hicieron uso de la misma.

8.2.2 Clasificación – Árbol de decisión

De la misma manera que en el análisis de la situación socio-demográfica de los estudiantes, se ejecutará el algoritmo J48 con diferentes valores en el parámetro de factor de confianza de la poda, para comparar resultados.
Se genera una primera clasificación con el valor del parámetro por defecto. La salida se puede observar en la Figura 8.10 – Algoritmo J48, con los parámetros por defecto.

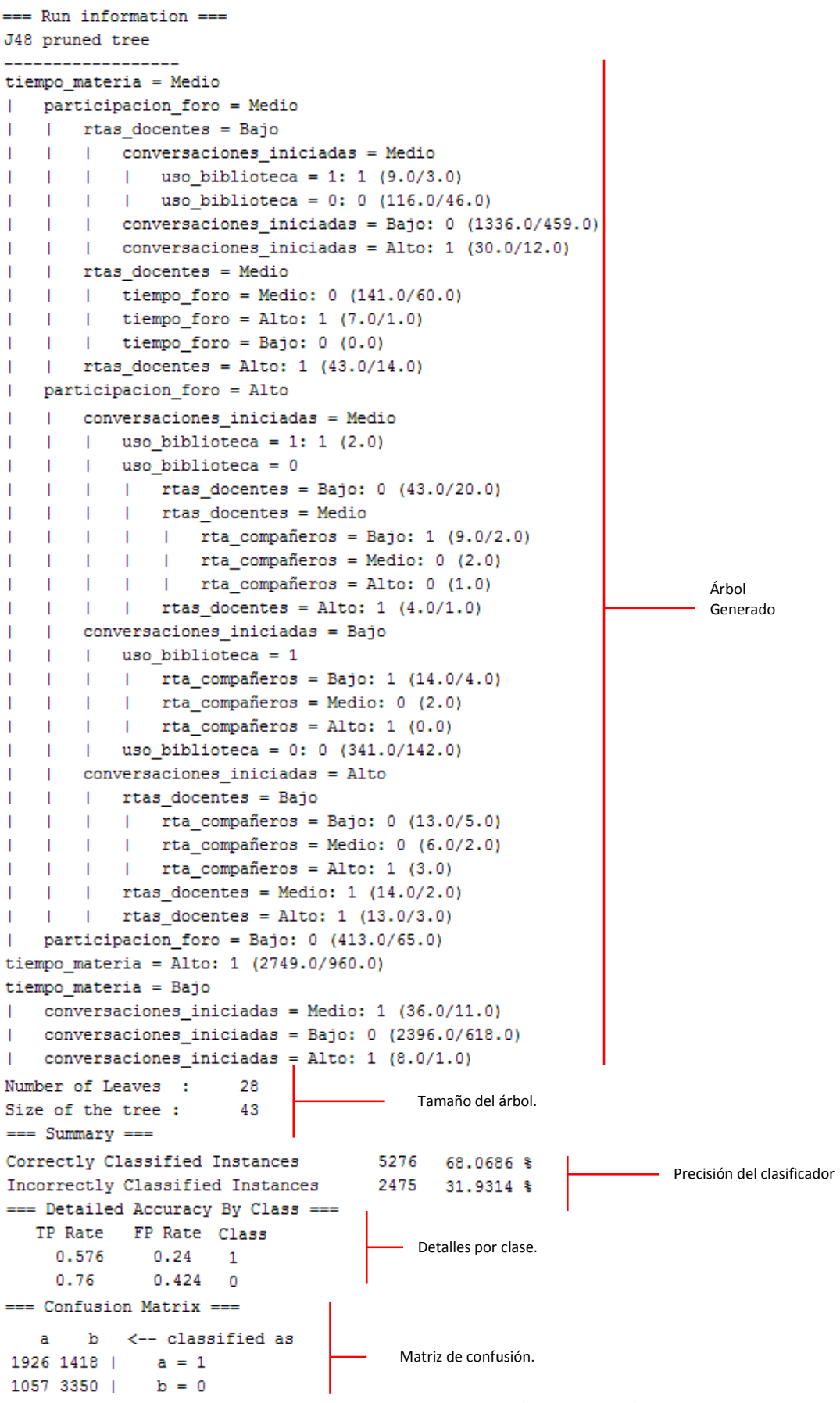


Figura 8.10 – Algoritmo J48, con los parámetros por defecto.

Se generó un árbol con un tamaño total de 43 nodos. El clasificador obtenido, tiene una precisión del 68.06%. La razón de aciertos de la clase Desaprobado es de 0.76, y la razón de aciertos de la clase Aprobado es de 0.576. En la matriz de confusión se puede observar cómo es la distribución entre los datos bien y mal clasificados.

El modelo generado, indica que el atributo que mejor clasifica a los estudiantes es el tiempo total de uso del sistema Moodle relacionado con la materia.
Clasifica con la clase Aprobado a aquellos estudiantes que tienen un tiempo de uso del sistema Moodle alto.
La rama del árbol que hace referencia al uso Medio del sistema Moodle, es la más compleja que se formó en este clasificador; ya que se tienen en cuenta luego, el uso activo y pasivo de los foros, y el uso de bibliotecas del estudiante.
Los estudiantes con tiempo de uso Bajo del sistema Moodle, son clasificados dependiendo de la participación activa que tienen en los foros; se clasifican con la clase Aprobado a los estudiantes que tienen una participación activa alta o media y se clasifican con la clase desaprobados los que tienen una participación activa baja.
Se observa también que en la mayoría de las ramificaciones del árbol generado, en que se evalúa el uso de bibliotecas, si el estudiante hace uso de bibliografía de la materia proveniente de la biblioteca, es clasificado con la clase Aprobado.

Se modificará el parámetro de factor de confianza de la poda, para ver si se puede obtener un modelo más simple, sin sacrificar la precisión ya obtenida; ó más compleja que mejore significativamente su precisión.

Con el factor de confianza = 0.001, obtenemos los siguientes resultados en cuanto al tamaño del árbol y la precisión del mismo:

```
Number of Leaves  :     5
Size of the tree :      7
=== Summary ===
Correctly Classified Instances        5256  67.8106 %
Incorrectly Classified Instances      2495  32.1894 %
```

Figura 8.11 – Algoritmo J48, con factor de confianza = 0.001.

Vemos que el tamaño del árbol se reduce considerablemente, con un tamaño total de 7 nodos, pero también se reduce la precisión del mismo.

Con el factor de confianza = 1.0, obtenemos los siguientes resultados:

```
Number of Leaves  :     92
Size of the tree :      141
=== Summary ===
Correctly Classified Instances        5276   68.0686 %
Incorrectly Classified Instances      2475   31.9314 %
```

Figura 8.12 – Algoritmo J48, con factor de confianza = 1.

En este caso obtenemos un modelo mucho más complejo que el primer árbol generado, sin ganar en la precisión del mismo. El árbol generado tiene un tamaño que supera las 3 veces del tamaño del primer árbol generado, lo que dificulta considerablemente su lectura.

Los modelos generados con las modificaciones del parámetro confianza de la poda, se pueden ejecutar y ver desde el aplicativo que se entrega junto con este informe.

8.2.3 Agrupamiento

De la misma manera que en el análisis de la situación socio-demográfica de los estudiantes, se ejecutará el algoritmo SimpleKMeans para la tarea de agrupamiento. Se realizarán diferentes ejecuciones formando diferentes cantidades de grupos para comparar los resultados entre ellos.
Ejecutamos el algoritmo para obtener tres grupos de estudiantes con características similares.

```
kMeans
======
Cluster centroids:
                          Cluster#
Attribute                       0          1          2
                            (3160)     (3951)      (640)
=========================================================
participacion_foro             Bajo      Medio       Bajo
rta_compañeros                 Bajo       Bajo       Bajo
rtas_docentes                  Bajo       Bajo       Bajo
conversaciones_iniciadas       Bajo       Bajo       Bajo
tiempo_foro                    Bajo      Medio       Bajo
tiempo_materia                 Bajo       Alto       Bajo
uso_biblioteca                    0          0          0
m_cursada                         0          1          1
=== Model and evaluation on training set ===
 Clustered Instances
  0      3160 ( 41%)
  1      3951 ( 51%)
  2       640 (  8%)
```

Figura 8.13 – Algoritmo SimpleKMeans con 3 clusters.

Vemos, en la Figura 8.13 – Algoritmo SimpleKMeans con 3 clusters, que el primero de los clusters agrupa a estudiantes con participación baja en los foros, tanto activa como pasiva, no hacen uso de la biblioteca y están desaprobados. El segundo de los clusters contiene en su mayoría a estudiantes con una mejor participación en los foros, no hacen uso de la biblioteca y están aprobados. El tercer de los clusters, tiene los mismos valores en la mayoría de los atributos que el cluster 0, con la diferencia de que agrupa estudiantes aprobados. Hay que tener en cuenta que este último cluster representa solamente al 8% de la población.

La ejecución del algoritmo para generar cinco clusters arroja los resultados que se muestran en la Figura 8.16 – Algoritmo SimpleKMeans con 5 clusters, y se describen a continuación.

Los primeros tres clusters son muy similares a los clusters de la ejecución anterior; mientras que el cluster 3 se caracteriza por estudiantes aprobados con alta participación pasiva y baja participación activa en los foros y el cluster 4 tiene baja participación en los foros y una participación media en la materia en general.

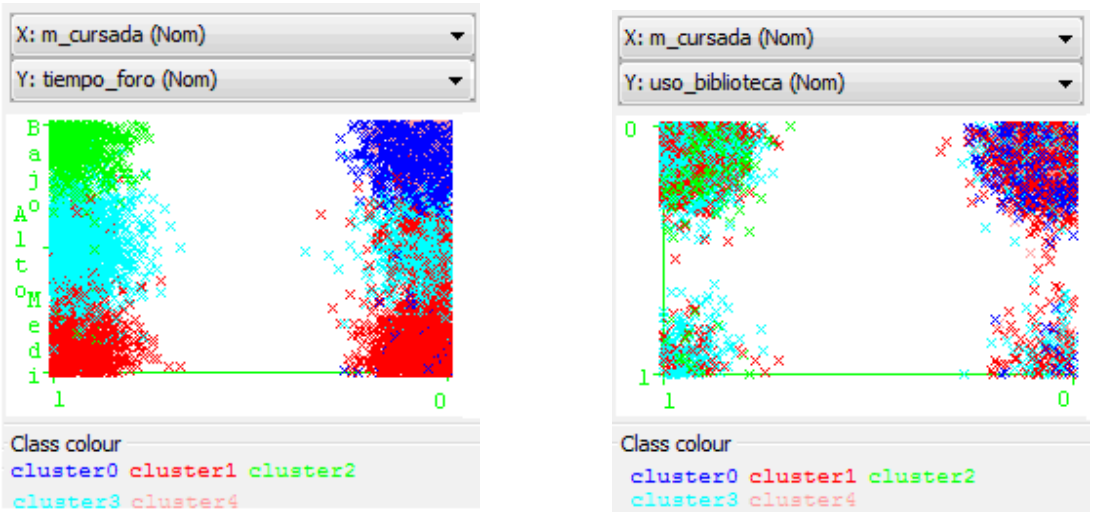
Podemos observar que ninguno de los cluster se caracteriza por contener estudiantes que hacen uso de bibliotecas. Esto puede deberse a la baja cantidad de estudiantes que se sabe que hicieron uso de material de la biblioteca de la Facultad.

```
kMeans
======
Cluster centroids:
                          Cluster#
Attribute                       0       1       2       3       4
                            (1633)  (2823)  (644)  (2385)  (266)
=============================================================
participacion_foro           Bajo   Medio    Bajo    Alto    Bajo
rta_compañeros               Bajo    Bajo    Bajo    Bajo    Bajo
rtas_docentes                Bajo    Bajo    Bajo    Bajo    Bajo
conversaciones_iniciadas     Bajo    Bajo    Bajo    Bajo    Bajo
tiempo_foro                  Bajo   Medio    Bajo    Alto    Bajo
tiempo_materia               Bajo   Medio    Bajo    Alto   Medio
uso_biblioteca                  0       0       0       0       0
m_cursada                       0       0       1       1       0
=== Model and evaluation on training set ===
 Clustered Instances
  0      1633 ( 21%)
  1      2823 ( 36%)
  2       644 (  8%)
  3      2385 ( 31%)
  4       266 (  3%)
```

Figura 8.14 – Algoritmo SimpleKMeans con 5 clusters.

8.2.3.1 Visualización de los grupos

En las siguientes figuras vemos gráficamente la distribución de los estudiantes en los diferentes clusters, relacionados con los atributos m_aprobado, uso de biblioteca y tiempo empleado en el foro.



A – m_cursada y tiempo_foro

B – m_cursada y uso de biblioteca.

Figura 8.15 – Distribución de los datos según los atributos.

Podemos observar, en la Figura 8.15 (A) que los estudiantes con nivel de actividad media en el foro se encuentran, en su mayoría en un mismo cluster –cluster 1, parte inferior de la imagen –. Los estudiantes que tienen participación alta en el foro también se agrupan en un mismo cluster –cluster 3, parte media de la imagen –. Los estudiantes con participación baja en el foro, se dividen en dos grupos, en uno se encuentran los aprobados –cluster 2, parte superior izquierda de la imagen – y en otro se encuentran los estudiantes desaprobados, –cluster 0, parte superior derecha de la imagen –.

En la Figura 8.15 (B) no se observa una separación de estudiantes entre los diferentes clusters tan clara como se puede observar en la Figura 8.15 (A); teniendo una división un poco más compleja de analizar. La mayoría de los estudiantes aprobados se encuentran en el cluster 3, pero el uso de biblioteca parece no afectar la nota final obtenida en la materia. La mayoría de los estudiantes desaprobados se encuentran en el cluster 0 y el uso de biblioteca tampoco parece afectar la nota obtenida en la materia cursada.

Capítulo 9
APLICATIVO

En el presente capítulo, se describe el aplicativo realizado para este trabajo de grado, el cual se entrega junto con este informe.

9.1 Funcionalidad provista

El objetivo del aplicativo, es facilitar el uso de tareas y algoritmos de minería de datos a usuarios finales, evitando que tengan herramientas similares a WEKA lo que puede resultar dificultoso para usuarios que no están relacionados con temas de minería de datos de manera directa.
El aplicativo se centra en la ejecución y presentación de los resultados de las tareas y algoritmos de minería de datos descriptos en el Capítulo *8 – Modelado y evaluación*. Especialmente facilita los análisis desarrollados en los incisos 8.2 – *Análisis socio demográfico* y 8.3 – *Análisis de la participación social en las materias*.
El usuario será capaz de:

- Seleccionar el tipo de minería de datos a realizar; las posibles selecciones son clasificación y agrupamiento.
- Seleccionar los principales valores de los parámetros para estas dos taras, esto es la confianza de la poda para la clasificación, y la cantidad de clusters para la tarea de agrupamiento.
- Seleccionar el tipo de análisis a realizar, esto es el análisis socio demográfico o el análisis de la participación social en las materias ó seleccionar un archivo de datos para la generación del modelo.
- Observar y analizar los modelos generados.
- Aplicar los modelos generados a datos desconocidos.

En caso de no seleccionar un archivo de datos para la generación de los modelos, se utilizarán archivos con datos de las vistas minables utilizadas en el desarrollo de los incisos 8.2 y 8.3 del capítulo 8 del presente trabajo de grado.
Como se utilizan las librerías de WEKA para la realización y ejecución de los modelos; es necesario que los archivos de datos, tanto para la generación de modelos o la ejecución de los mismos, cumplan con el formato .arff de WEKA.

9.2 Ejecución de la Aplicación

Para poder ejecutar la aplicación, es requisito indispensable tener instalado una máquina virtual de java versión 1.6 ó superior.
Junto con este informe, se entrega un archivo comprimido llamado tesisKruzylkoOstojic(exe).zip; para poder ejecutar la aplicación, se debe descomprimir dicho archivo en un directorio deseado.
La estructura de carpetas y archivos que contiene el .zip es el siguiente:

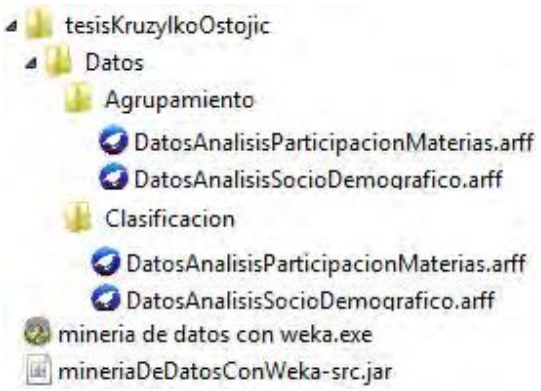


Figura 9.1 - Estructura de archivos entregados.

mineriaDeDatosConWeka.exe, es el ejecutable de la aplicación.
El directorio /tesisKruzylkoOstojic/datos/Agrupamiento, contiene los archivos que se pueden usar para probar las agrupaciones de nuevas instancias, ver sección *9.7 - Tarea de Agrupamiento*.
El directorio /tesisKruzylkoOstojic/datos/Clasificacion, contiene los archivos con datos desconocidos, que se pueden usar como ejemplo para probar las clasificaciones de nuevas instancias, ver sección *9.6 - Tarea de Clasificación*.
El archivo mineriaDeDatosConWeka-src.jar; contiene el código fuente completo del aplicativo. Se debe abrir utilizando una herramienta de descompresión de archivos, como por ejemplo winzip ó winrar.

En caso de que el archivo .exe no sea compatible con el sistema operativo utilizado por el usuario, se entrega también un archivo tesisKruzylkoOstojic(jar).zip; el cual tiene la misma estructura que el archivo tesisKruzylkoOstojic(exe).zip con la diferencia que contiene el archivo mineriaDeDatosConWeka.jar en vez de mineriaDeDatosConWeka.exe.
El archivo mineriaDeDatosConWeka.jar es un jar ejecutable que se puede ejecutar desde línea de comandos y de esta manera, para poder probar la aplicación. Sólo basta con situarse en el directorio que contiene el archivo mineriaDeDatosConWeka.jar y ejecutar el comando
java –jar mineriaDeDatosConWeka.jar

9.3 Pantalla inicial del aplicativo

Al iniciar la aplicación, se mostrará la pantalla que se observa en la Figura 9.1 – Pantalla inicial del aplicativo.

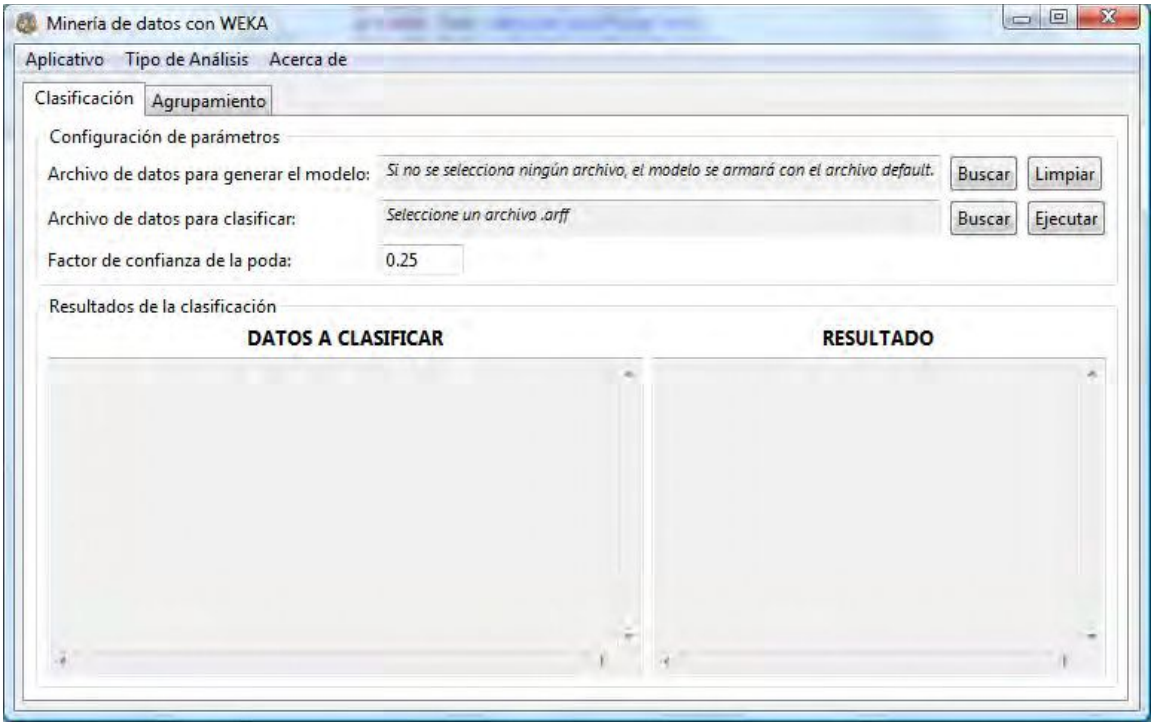


Figura 9.2 – Pantalla inicial del aplicativo.

Se puede observar, las dos solapas del panel principal de la pantalla, las mismas deben seleccionarse adecuadamente según se requiera una tarea de clasificación o una tarea de agrupamiento. Cada una de las solapas permite la selección del archivo de datos para armar el modelo –se usará un archivo por defecto en cado de no seleccionar ninguno –; la selección de un archivo de datos desconocidos; la configuración de parámetros particulares y por último se puede observar en la parte inferior de la pantalla, dos cuadros de texto en donde se mostrarán los datos desconocidos seleccionados –cuadro de la izquierda – y los resultados de la ejecución –cuadro de la derecha –.

9.4 Tipos de análisis

La opción más importante de la barra de menú es “Tipo de Análisis”, donde se podrá seleccionar si se desea realizar el análisis socio demográfico ó el análisis de la participación social de los estudiantes en las materias. La selección de un tipo de análisis, permitirá setear adecuadamente, el archivo de datos por defecto que se usará para la realización del modelo deseado.

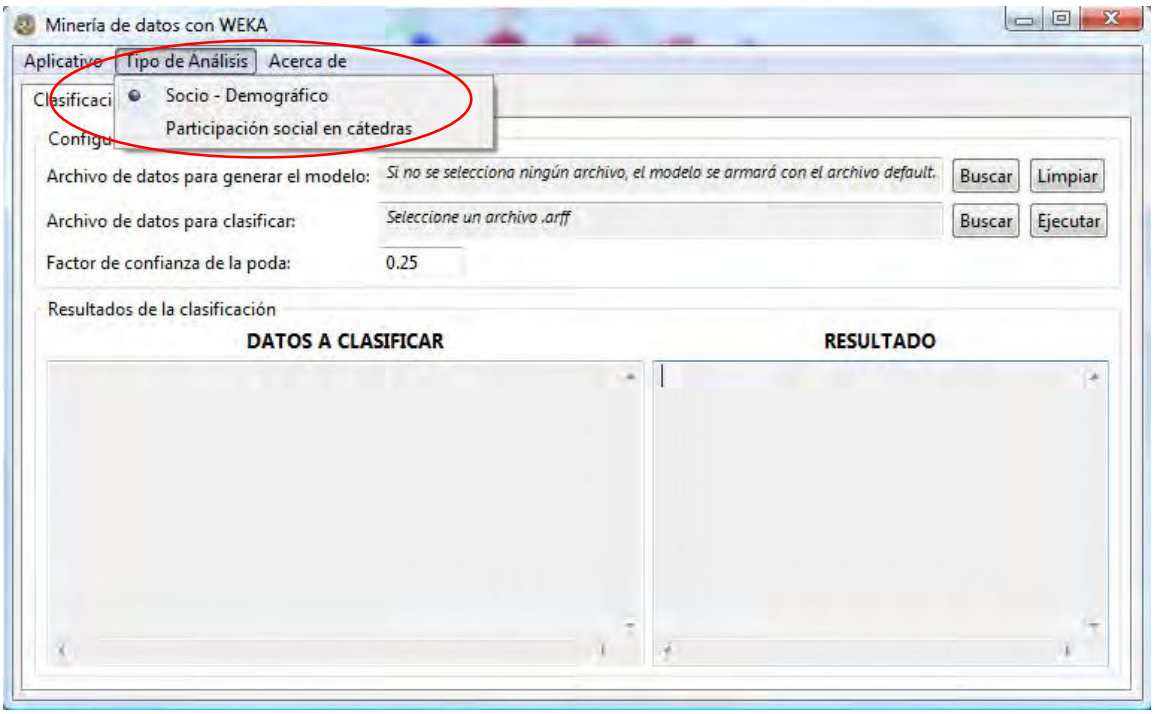


Figura 9.3 - Selección del tipo de análisis.

9.5 Tarea de Clasificación

Para ejecutar una tarea de clasificación, en la solapa de clasificación es necesario tener seteado los siguientes valores:

- El archivo de datos para generar el modelo, tener en cuenta que si no se selecciona ningún archivo particular, el modelo de árbol de decisión se creará con un archivo de datos por defecto, dependiendo de la selección de análisis hecha desde el menú “Tipo de Análisis”.
- Archivo de datos desconocidos para clasificarlos adecuadamente utilizando el modelo generado. En el caso de la tarea de clasificación, la selección de este archivo es obligatoria. Se pueden usar los archivos del la carpeta /tesisKruzylkoOstojic/datos/clasificacion, que acompaña la aplicación, para realizar la prueba según el tipo de análisis que se esté ejecutando.
- El parámetro de confianza de la poda del árbol. Un valor que tiene que estar entre 0 y 1.

Una vez seleccionado el archivo de datos desconocidos, éstos se visualizarán en el campo de texto “DATOS A CLASIFICAR”.
Cuando se ejecute la aplicación, la clasificación de los datos desconocidos y el modelo generado se visualizará en el campo de texto “RESULTADO”.

Supongamos que queremos clasificar un grupo de estudiantes en la situación académica de cada uno, según la situación socio-demográfica que posean.
Vamos a generar el modelo de árbol de clasificación con el archivo de datos por default, entonces nos tenemos que asegurar de tener seleccionado el tipo de análisis socio-demográfico en la barra de tareas y dejar el valor de “Archivo de datos para generar el modelo” con la leyenda original.
Cargamos el archivo de datos que contiene los registros desconocidos, haciendo clic en el botón “Buscar” que se encuentra a la derecha de la opción “Archivo de datos para clasificar”, se abrirá el diálogo que se muestra en la Figura 9.3 - Diálogo de selección de archivo. Del cual se podrá seleccionar el archivo de datos deseado.

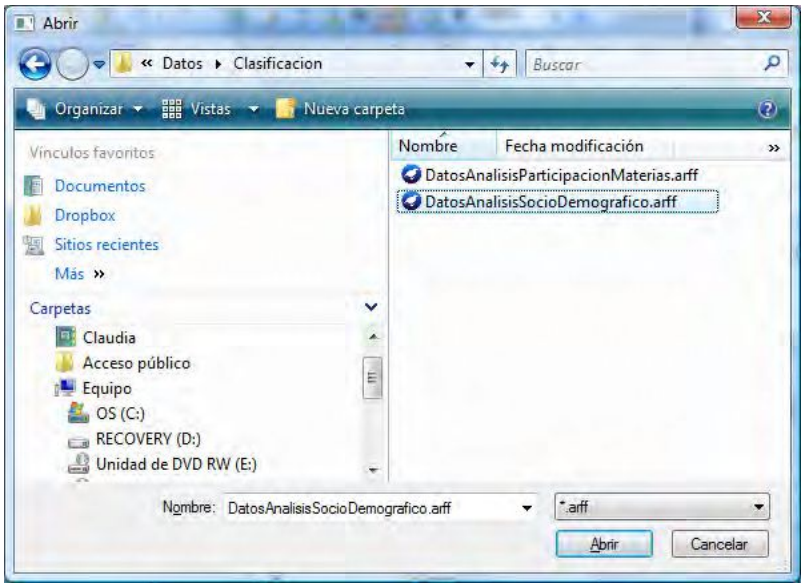


Figura 9.4 - Diálogo de selección de archivo.

Una vez seleccionado el archivo, los valores que el mismo contiene, se mostrarán en el cuadro de texto de la izquierda de la pantalla, como muestra la Figura 9.4 - Datos seleccionados para clasificar.

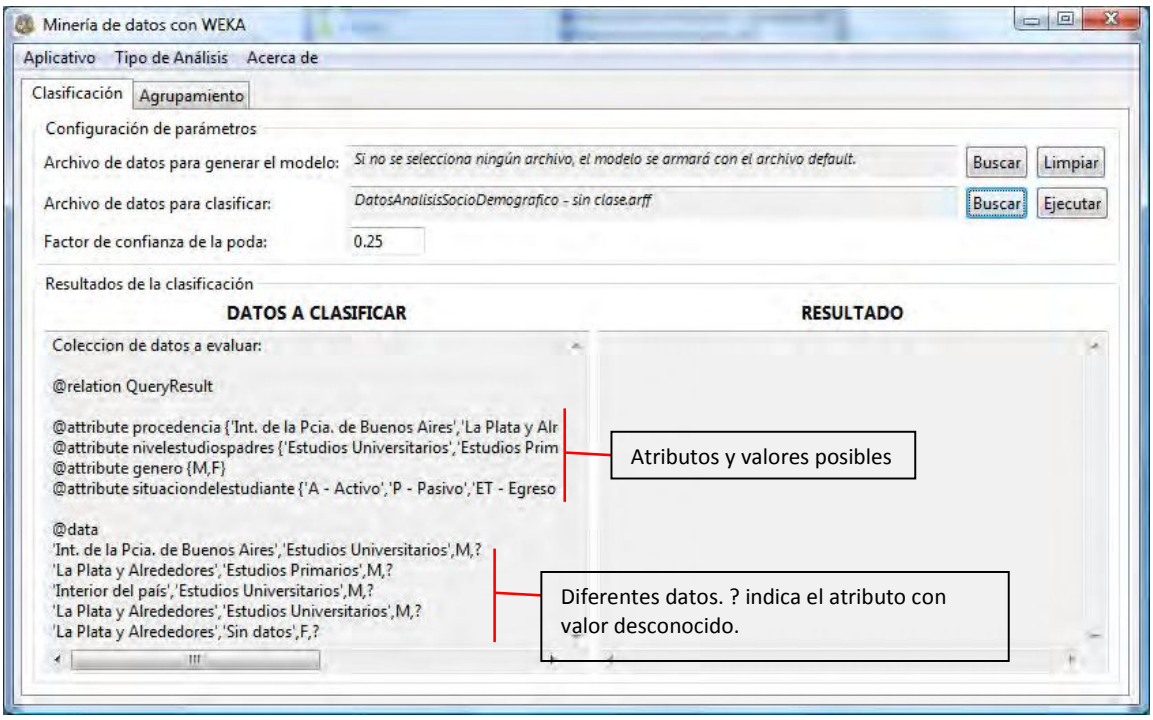


Figura 9.5 - Datos seleccionados para clasificar.

En los datos seleccionados primero se observa la lista de atributos, junto con los posibles valores de cada uno de ellos y luego se muestra en diferentes renglones los diferentes datos, indicando para cada uno de ellos el valor del atributo conocido y con el símbolo '?', se indica el valor del atributo desconocido; este valor es el que tendrá que deducir el modelo que se generará.

Dejamos el factor de confianza de la poda con su valor por defecto y seleccionamos la opción "Ejecutar". Esto hará que se genere un árbol de decisión; con dicho modelo se clasificarán las instancias del archivo de datos desconocidos y se mostrará el resultado obtenido en el cuadro de texto "RESULTADO".

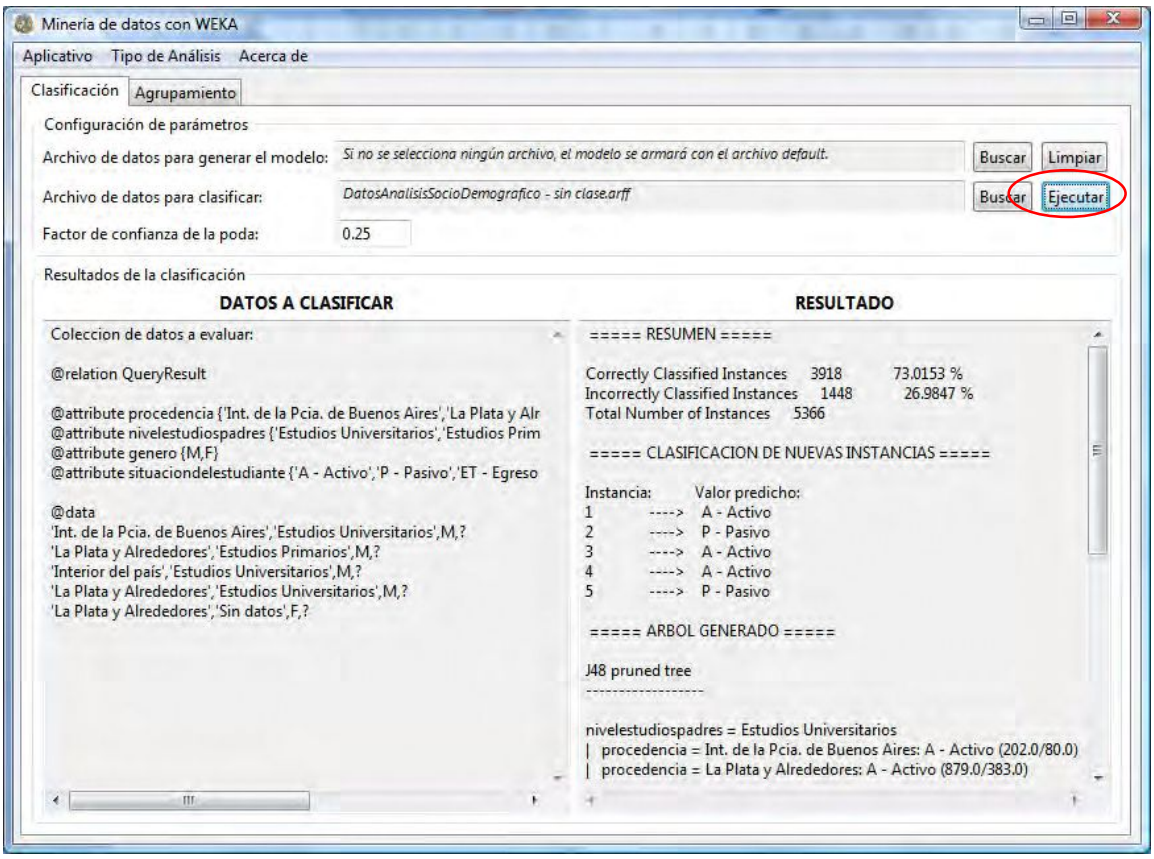


Figura 9.6 - Resultados de la clasificación.

Primero se observa un resumen del modelo generado que indica la precisión del mismo, en este caso, se generó un modelo que tiene más de un 73% de precisión.
En segundo lugar se observa cómo clasificaron cada una de las instancias del conjunto de datos desconocidos. El resultado se muestra por orden según como aparecen en el archivo; en este ejemplo, la primera instancia del archivo de datos clasifica como A-Activo, la segunda instancia clasifica como P – Pasivo y así sucesivamente.
Luego se puede observar el árbol generado y si se desplaza el scroll hasta el final, se podrá ver la matriz de confusión.

La opción de “Limpiar” que se puede ver arriba de la opción de “Ejecutar”, se encarga de volver los valores de los parámetros a como estaban por default.

9.6 Tarea de agrupamiento

Para ejecutar una tarea de agrupamiento, en la solapa de agrupamiento es necesario tener configurados los siguientes parámetros:

- El archivo de datos para generar el modelo, tener en cuenta que si no se selecciona ningún archivo particular, el modelo se creará con un archivo de datos por defecto, dependiendo de la selección de análisis hecha desde el menú “Tipo de Análisis”.

- Archivo de datos que se quiera agrupar, utilizando el modelo generado. En el caso de la tarea de agrupamiento, la selección de este archivo no es obligatoria; esto se debe a que la mayoría de las veces una tarea de agrupamiento es utilizada como tarea descriptiva de los datos ya existentes. Se pueden usar los archivos del la carpeta /tesisKruzylkoOstojic/datos/agrupamiento, que acompaña la aplicación, para realizar la prueba según el tipo de análisis que se esté ejecutando.
- El parámetro de la cantidad de clusters o grupos a generar.

En caso de seleccionar un archivo de datos a agrupar, éstos se visualizarán en el campo de texto “DATOS A AGRUPAR”.
Cuando se ejecute la tarea, el grupo al que se asocian los datos nuevos y el modelo generado, se visualizarán en el campo de texto “RESULTADO”.

Supongamos que queremos ver como se agrupan los estudiantes según la participación social que tengan en las materias mediante el sistema Moodle.
Vamos a generar el modelo de grupos, con el archivo de datos por default para este análisis, nos tenemos que asegurar, entonces, de tener seleccionado el tipo de análisis “Participación social en cátedras”, de la barra de tareas y no seleccionar ninguna opción en “Archivo de datos para generar el modelo”.
Dejamos el valor de “Cantidad de grupos a generar” en 3 para que se generen 3 clusters, y por último, seleccionamos la opción “Ejecutar”. Esto hará que se generen 3 grupos que caracterizan a los estudiantes, según la participación social que tienen en las materias mediante el sistema Moodle. Los resultados obtenidos, se mostrarán en el cuadro de texto “RESULTADO”.

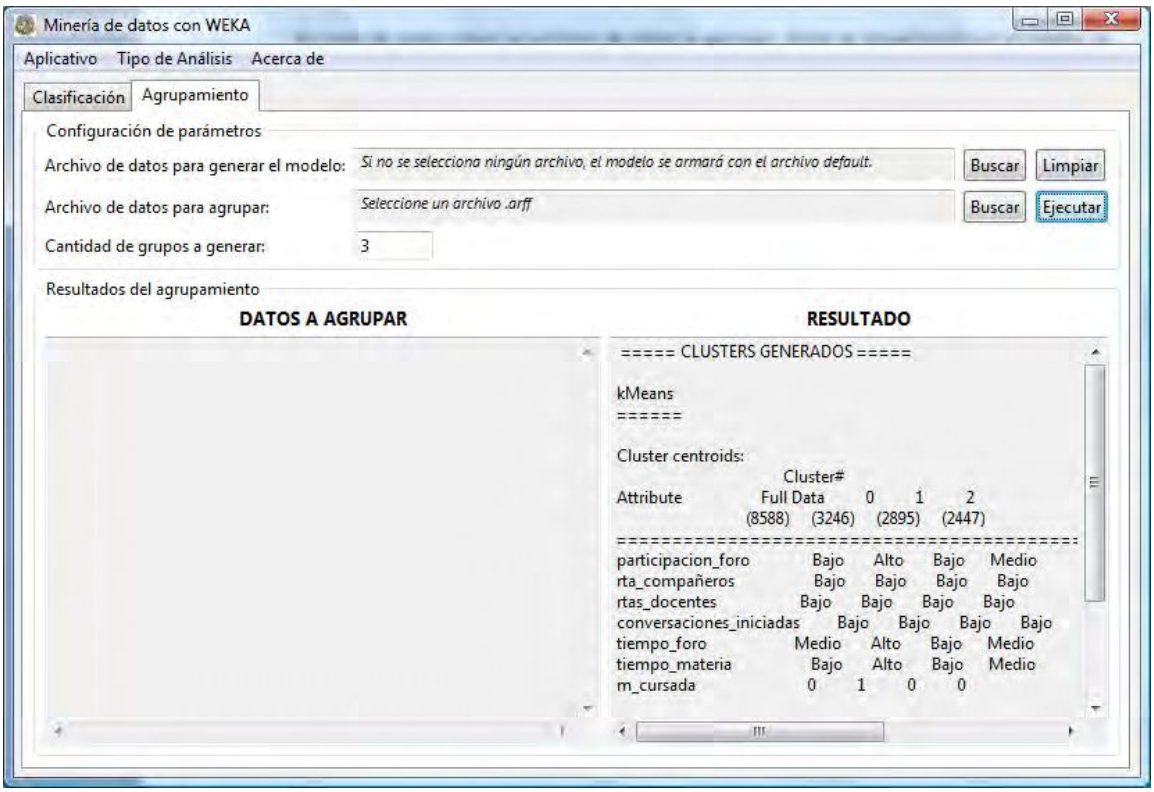


Figura 9.7 - Resultados del agrupamiento.

Se observa la composición de cada uno de los clusters generados. Se muestra el valor mayoritario de cada uno de los atributos por cluster.

Si hubiésemos seleccionado un conjunto de datos para agrupar; como primera información del resultado, hubiese aparecido una lista indicando, para cada instancia, el grupo al cual se la asociaba.

Si se desplaza el scroll hasta el final del cuadro de texto de resultados, se podrá ver el porcentaje de elementos con el cual está conformado cada uno de los clusters.

La opción de “Limpiar” que se puede ver arriba de la opción de “Ejecutar”, se encarga de volver los valores de los parámetros a como estaban por default.

9.7 Otras opciones del aplicativo

En la barra de menú se observan dos opciones más que no se han mencionado hasta el momento. El menú “Aplicativo”; solamente contiene la opción “Salir” que permite salir de la aplicación y el menú “Acerca de” que abre un diálogo que contiene un breve explicación sobre la funcionalidad de la aplicación.

Capítulo 10
EXTENSIÓN DEL TRABAJO

Las principales fuentes de datos para la realización de este trabajo, fueron las bases de datos de los sistemas utilizados por la Facultad de Informática – UNLP para el almacenamiento de los datos de sus estudiantes y graduados. Pero, también se pueden obtener datos referentes a los alumnos y egresados, de sistemas externos a la Facultad, como pueden ser las redes sociales tan, ampliamente usadas en la actualidad.
Se propone una extensión de este trabajo de grado, relacionando datos obtenidos de los sistemas de da Facultad de Informática, con datos obtenidos de las redes sociales Facebook y Twitter.

10.1 Las redes sociales en la Web

Las redes sociales en la Web, son principalmente un lugar de interacción virtual, en el que millones de personas alrededor del mundo se conectan con diversos intereses en común. Las herramientas informáticas para potencias la eficacia de las redes sociales online, operan en tres ámbitos: comunicación –ayudan a poner conocimiento en común –, comunidad –ayudan a encontrar e integrar comunidades – y cooperación –ayudan a realizar cosas juntos –. Las redes sociales sirven para dar a conocer todo tipo de información entre las personas. Su propósito es facilitar la comunicación y otros temas sociales en el sitio Web. Gracias a las redes sociales las formas en que las personas se comunican ha cambiado considerablemente. Dos de las redes sociales más populares actualmente son Facebook y Twitter.

10.2 Uso de las redes sociales en las materias

Aprovechando la difusión de información que permiten las redes sociales y el gran uso de las mismas; algunas cátedras de la Facultad de Informática, comenzaron a hacer uso de las mismas para comunicarse con sus estudiantes.
Entre las cátedras que comenzaron a hacer uso de las redes sociales para comunicación don sus estudiantes, se pueden nombrar, Introducción a los sistemas operativos, Sistemas operativos y Tecnologías aplicadas para Business Intelligence. En particular, los docentes de la materia Tecnologías aplicadas para Business Intelligence utilizan tanto la red social Facebook como Twitter como medio de comunicación secundario con sus estudiantes.

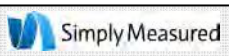
10.3 Análisis de las materias incluyendo las redes sociales

Aprovechando esta iniciativa, se decide hacer una extensión de este trabajo de grado, analizando si la participación de los estudiantes con las materias, incluyendo, además del uso de la plataforma Moodle, el uso de las redes sociales, influye o no en el resultado que pueden obtener en las mismas.
Se reduce el problema, haciendo este análisis sobre los registros de estudiantes de una sola de las materias que hacen uso de las redes sociales. La materia seleccionada es "Tecnologías Aplicadas para Business Intelligence" ya que es la que hace uso de las dos redes sociales más usadas actualmente.

10.4 Obtención de datos

Seleccionamos, de la vista minable creada para el análisis de la participación de los estudiantes en las materias, aquellos registros pertenecientes a la materia Tecnologías Aplicadas para Business Intelligence; le agregamos información que indica si el estudiante es seguidor o no de la cuenta de Twitter de la cátedra y si es amigo de la página de Facebook que la materia posee.

La información de los seguidores de Twitter de la cuenta de la cátedra, se consiguió mediante el sistema Simply Measured, un producto certificado por Twitter, el cual permite obtener un reporte de los seguidores de una cuenta de manera muy sencilla.

Simply Measured es una herramienta para la obtención de  *información de redes sociales y crear con ellos informes en Excel, dashboard en la Web y en PowerPoint. Comenzó como una herramienta analítica de Twitter.*

La información de los amigos de la página de Facebook se consiguió mediante el sistema Netvizz junto con el sistema Ghephi. Si bien estas herramientas permiten un análisis más complejo de la relación entre los usuarios de redes sociales, solamente las utilizamos en este trabajo para obtener los "amigos" mediante Facebook, de la cátedra Tecnologías Aplicadas para Business Intelligence.

Netvizz es una aplicación que accede a la base de datos de Facebook, y obtiene datos del usuario y sus contactos. Permite la creación de un archivo .gdf que contiene las relaciones de la cuenta personal del usuario y de los grupos a los que pertenece. Los archivos .gdf pueden ser analizados y visualizados usando software gráfico de visualización como la plataforma Gephi.

Gephi es una plataforma interactiva para la exploración, visualización y consulta de grafos. El análisis de redes sociales, se realiza aplican teoría de grafos, indicando las entidades como nodos y las relaciones como aristas.

10.5 Análisis

Una vez armada la vista minable con esta información, se realizó el análisis de estos datos de manera similar a los dos análisis descriptos en el *Capítulo 8 – Modelado y Evaluación*.
La vista minable está compuesta por los mismos atributos que la vista minable creada para el análisis descripto en el Capítulo 8, sección 8.2 – Análisis de participación en las materias; excepto el atributo de uso de bibliotecas –esto es debido a que no se registraron en el sistema Merán uso de biblioteca relacionado a la materia Tecnologías aplicadas para Business Intelligence –; agregándole los atributos Twitter y Facebook que indican si los estudiantes usan estas redes sociales para comunicación con los integrantes de la materia.

10.5.1 Distribución de los datos

En los histogramas de la Figura 10.1 – Distribución de los datos, podemos ver la distribución de los atributos con respecto a la clase. El atributo m_cursada situación del estudiante será la clase a predecir en la tarea de clasificación.

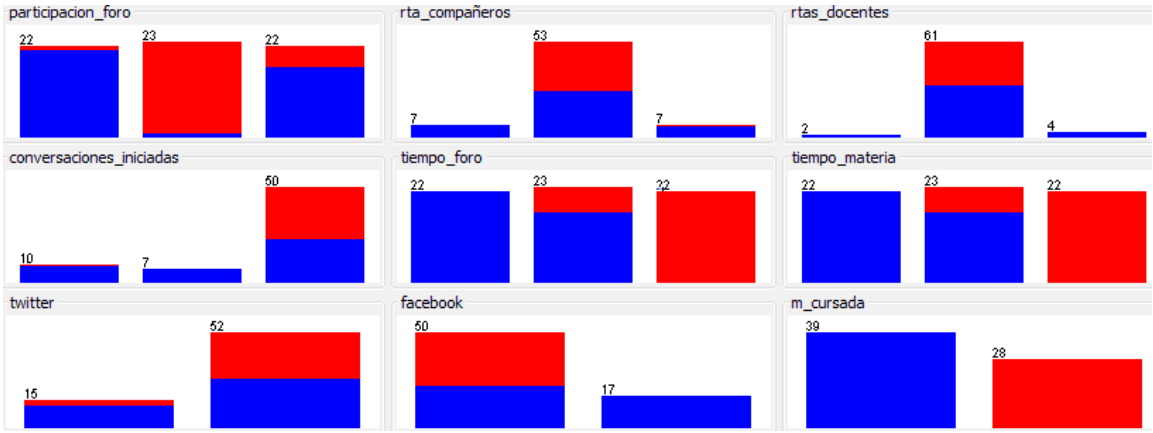


Figura 10.1 – Distribución de los datos.

A continuación se listan los valores posibles, de izquierda a derecha, de cada uno de los atributos.

Atributo	Valores posibles (de izquierda a derecha según la Figura 10.1)
participacion_foro	Alto – Bajo – Medio
rta_compañeros	Alto – Bajo – Medio
rtas_docentes	Alto – Bajo – Medio

conversaciones_iniciadas	Medio – Alto – Bajo
tiempo_foro	Alto – Medio – Bajo
tiempo_materia	Alto – Medio – Bajo
Twitter	Sí – No
Facebook	No – Sí
m_cursada	Sí – No

En la distribución de los estudiantes según la gama de colores, en cada una de las barras de los atributos participacion_foro, rta_compañeros, rtas_docentes, tiempo_foro, tiempo_materia, conversaciones_iniciadas, Twitter, Facebook es, desde la base de la barra hacia arriba: Aprobado (Azul), Desaprobado (Rojo).

Haciendo un análisis de esta distribución de los datos, podemos ver que el uso de las redes sociales no es una elección de comunicación masiva por parte de los estudiantes; ya que más del 50% de los mismos elige no seguir a la cuenta de Twitter de la materia al igual que ser amigos por medio de Facebook. También podemos ver que la mayoría de los estudiantes que sí eligen este medio de comunicación, tienen un resultado exitoso en la materia.
Los atributos en este análisis hacen una mejor separación de las clases en comparación con los análisis descriptos en el *Capítulo 8 – Modelado y evaluación*. Por ejemplo, tiempo_foro y tiempo_materia separan casi perfectamente las clases; ya que si los valores de estos atributos son Alto o Medio, es muy probable que la clase sea Aprobado, mientras que si el valor es Bajo, la clase es Desaprobado.

10.5.2 Árbol de decisión

De la misma manera que ambos análisis realizados en el *Capítulo 8. Modelado y Evaluación*; se ejecutará el algoritmo J48 con diferentes valores en el parámetro de factor de confianza de la poda, para comparar resultados.
Se genera una primera clasificación con los valores de los parámetros por defecto que sugiere Weka para el algoritmo J48. La salida se puede observar en la Figura 10.3 – Algoritmo J48, con los parámetros por defecto.

```
=== Run information ===
J48 pruned tree
------------------
tiempo_foro = Alto: 1 (22.0)
tiempo_foro = Medio
|   tiempo_materia = Alto: 1 (6.0)
|   tiempo_materia = Medio: 1 (14.0/3.0)
|   tiempo_materia = Bajo: 0 (3.0)
tiempo_foro = Bajo: 0 (22.0)

Number of Leaves  :     5
Size of the tree :      7
=== Summary ===
Correctly Classified Instances     58  86.5672 %
Incorrectly Classified Instances    9  13.4328 %
=== Detailed Accuracy By Class ===
  TP Rate   FP Rate   Precision  Class
    1         0.321      0.813    1
    0.679     0          1        0
=== Confusion Matrix ===
  a  b   <-- classified as
 39  0 |  a = 1
  9 19 |  b = 0
```

Figura 10.3 – Algoritmo J48, con los parámetros por defecto.

Se generó un árbol con un tamaño total de 7 nodos. El clasificador obtenido, tiene una precisión del 86.56%. La razón de aciertos de la clase Desaprobado es de 0.67, y la razón de aciertos de la clase Aprobado es de 1. En la matriz de confusión se puede observar cómo es la distribución entre los datos bien y mal clasificados.

El modelo generado, indica que el atributo que mejor clasifica a los estudiantes es el tiempo empleado en el foro; clasificando con la clase Aprobado a aquellos que tienen un tiempo de de participación Alto y con la clase Desaprobado a los que tienen un tiempo de participación Bajo. El segundo atributo más importante es el tiempo empleado en el sistema Moodle, relacionado con la materia, clasificando con la clase Aprobado a los que tienen tiempo de uso Medio y Alto y clasificando con la clase Desaprobado a los que tienen tiempo de uso Bajo.

Se modificará el parámetro de factor de confianza de la poda, para ver si se puede obtener un modelo más simple, sin sacrificar la precisión ya obtenida; ó más compleja que mejore significativamente su precisión.
Con el factor de confianza = 0.001, obtenemos los siguientes resultados en cuanto al tamaño del árbol y la precisión del mismo:

```
Number of Leaves  :     3
Size of the tree :      4
=== Summary ===
Correctly Classified Instances     58  86.5672 %
Incorrectly Classified Instances    9  13.4328 %
```

Figura 10.4 – Algoritmo J48, con factor de confianza = 0.001.

Vemos que el tamaño del árbol se reduce a 4 nodos totales y la precisión del mismo no cambia con respecto a los resultados de la ejecución anterior.

Con el factor de confianza = 1.0, obtenemos los siguientes resultados:

```
Number of Leaves  :     5
Size of the tree :      7
=== Summary ===
Correctly Classified Instances     64  95.5224 %
Incorrectly Classified Instances    3   4.4776 %
```

Figura 10.5 – Algoritmo J48, con factor de confianza = 1.

Vemos que el tamaño del árbol no cambia con respecto a la primera ejecución del algoritmo J48, pero su precisión aumenta a un 95.52% de instancias bien clasificadas.

Podemos notar que en ninguna de las ejecuciones del algoritmo J48 se utilizaron los atributos Twitter y Facebook para la clasificación de los estudiantes. Esto se puede deber a la poca cantidad de estudiantes que por el momento optaron por estas herramientas como medio de comunicación con los responsables de la cátedra y sus compañeros.

10.5.3 Agrupamiento

Como se realizó en el *Capítulo 8. Modelado y Evaluación*; se ejecutará el algoritmo SimpleKMeans para la tarea de agrupamiento. Se realizarán diferentes ejecuciones formando diferentes cantidades de grupos para comparar los resultados entre ellos.

Al ejecutar el algoritmo para obtener tres grupos de estudiantes con características similares. Se obtuvo el siguiente resultado:

```
kMeans
======
Cluster centroids:
                         Cluster#
Attribute                      0          1          2
                            (21)       (25)       (21)
=============================================================
participacion_foro          Alto       Bajo      Medio
rta_compañeros              Bajo       Bajo       Bajo
rtas_docentes               Bajo       Bajo       Bajo
conversaciones_iniciadas    Bajo       Bajo       Bajo
tiempo_foro                 Alto       Bajo      Medio
tiempo_materia              Alto       Bajo      Medio
twitter                        0          0          0
facebook                       0          0          0
m_cursada                      1          0          1

=== Model and evaluation on training set ===
  Clustered Instances

  0      21 ( 31%)
  1      25 ( 37%)
  2      21 ( 31%)
```

Figura 10.7 – Algoritmo SimpleKMeans con 3 clusters.

Todos los clusters generados agrupan porcentajes muy similares de estudiantes.

El primer cluster representa a los estudiantes aprobados que tienen una participación pasiva alta en Moodle, una participación activa baja y no tienen comunicación con la cátedra mediante las redes sociales. Este cluster agrupa al 31% de los estudiantes.

El segundo de los clusters, está formado en su mayoría por estudiantes desaprobados con baja participación en el sistema Moodle y sin uso de las redes sociales para comunicación con las autoridades y compañeros de la materia. Agrupa el 37% de los estudiantes.

El tercero de los clusters, contiene en su mayoría estudiantes aprobados, con una participación pasiva media en Moodle y participación activa baja y también, como en los dos clusters anteriores, no hacen uso de las redes sociales para comunicación con la cátedra. Representa el 31% de los estudiantes de la materia seleccionada para este análisis.

La ejecución del algoritmo para generar cinco clusters arroja el siguiente resultado.

```
kMeans
======
Cluster centroids:
                        Cluster#
Attribute                     0          1          2          3          4
                            (10)       (24)       (15)       (10)        (8)
=============================================================================
participacion_foro        Medio       Bajo      Medio       Alto       Alto
rta_compañeros             Bajo       Bajo       Bajo       Bajo       Alto
rtas_docentes              Bajo       Bajo       Bajo       Bajo       Bajo
conversaciones_iniciadas  Medio       Bajo       Bajo       Bajo       Alto
tiempo_foro               Medio       Bajo      Medio       Alto       Alto
tiempo_materia            Medio       Bajo      Medio       Alto       Alto
twitter                       1          0          0          1          0
facebook                      0          0          0          0          0
m_cursada                     1          0          1          1          1

=== Model and evaluation on training set ===
 Clustered Instances
   0      10 ( 15%)
   1      24 ( 36%)
   2      15 ( 22%)
   3      10 ( 15%)
   4       8 ( 12%)
```

Figura 10.8 – Algoritmo SimpleKMeans con 5 clusters.

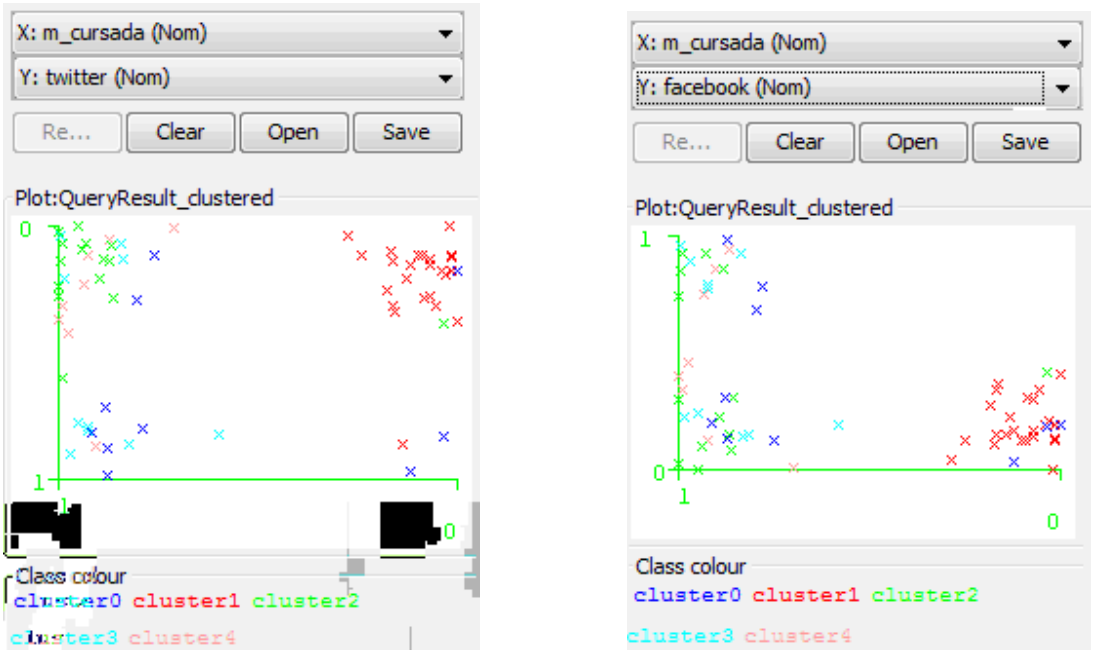
Vemos que no hay un grupo de los 5 formados, que se diferencie en cantidad de estudiantes y pueda llegar a ser un grupo de excepciones.

Algunos de los clusters formados en esta oportunidad sí se caracterizan por poseer en su mayoría estudiantes que hacen uso de las redes sociales para la comunicación con la cátedra. Uno es el caso del cluster 0 que contiene estudiantes aprobados, con participación pasiva media en el sistema Moodle; participación activa en las conversaciones iniciadas y uso de la red social Twitter. Luego tenemos al cluster 3 que contiene a estudiantes aprobados con alta participación pasiva en el sistema Moodle y baja participación activa, y se comunican con la cátedra también mediante Twitter.

Los otros tres clusters contienen, en su mayoría, a estudiantes que no se relacionan con la cátedra mediante las redes sociales de Facebook y Twitter, dos de las cuales contienen a estudiantes aprobados que tienen participación pasiva alta y media en el sistema Moodle y uno de los clusters contiene en su mayoría a estudiantes desaprobados con baja participación en el sistema Moodle.

10.5.3.1 Visualización de los grupos

En las siguientes figuras vemos gráficamente la distribución de los estudiantes en los diferentes clusters, teniendo en cuenta los atributos aprobados y los que los relacionan con el uso de las redes sociales en comunicación con la materia.



A - m_aprobado y Twitter.

B - m_aprobado y Facebook.

Figura 10.9 – Distribución en clusters según los atributos.

Vemos, en la Figura 10.9 (A) un agrupamiento bastante claro entre los estudiantes que siguen por Twitter la cuenta de la cátedra y los que no, con el resultado obtenido en la materia. Por ejemplo, los estudiantes que no siguen por Twitter a la cátedra y los que desaprueban, se encuentran mayoritariamente juntos, agrupados en este caso en el cluster 1 –parte superior derecha de la imagen –. De igual manera, observamos que se encuentran agrupados en su mayoría juntos, los estudiantes que sí siguen la cuenta de la cátedra por Twitter y aprueban la materia –cluster 0, parte inferior izquierda de la imagen –, y los estudiantes que siguen a la cátedra por Twitter y desaprueban –cluster 2, parte superior izquierda de la imagen –.

En la figura 10.9 (B), la separación de los clusters no es tan precisa. El cluster que mejor divide a los estudiantes, es el cluster 1 –parte inferior derecha de la imagen –, que agrupa mayoritariamente a los estudiantes desaprobados que no tienen relación con la cátedra mediante la red social Twitter.
Se observa también que no hay en ningún cluster, estudiantes que se relacionen con la cátedra mediante Facebook y tengan un resultado malo en la materia –parte superior derecha de la imagen –.

CONCLUSIONES Y TRABAJOS FUTUROS

Conclusiones

Se pudo realizar esta tesina de grado utilizando exclusivamente herramientas distribuidas bajo licencias de Software Libre. Se utilizaron diferentes herramientas a lo largo de las etapas del trabajo en las que se realizaron: la extracción de datos de los sistemas fuentes –los utilizados por la Facultad de Informática y las redes sociales –, el almacenamiento de los mismos en un ambiente local y su posterior preprocesamiento, realización de los diagramas de Entidad-Relación, generación de modelos y patrones para determinar la existencia de algún comportamiento común de los datos, y finalmente la realización de una aplicación Ad-Hoc que permitiera la exploración de los resultados obtenidos.

En este trabajo de grado, se aplicaron técnicas de minería de datos, sobre datos obtenidos de los estudiantes y egresados de la Facultad de Informática de la UNLP.
Primero, se trabajo con un solo data set, un conjunto reducido de datos pertenecientes a un único sistema utilizado en la Facultad de Informática. Se trabajó con datos extraídos del cubo 04 del sistema SIU-Guaraní.
Luego, se realizó una correlación de datos pertenecientes a diferentes sistemas utilizados por la Facultad, para lograr un análisis más integrado. Se trabajó con datos pertenecientes a los sistemas SIU-Guaraní, Moodle y Merán, siempre pensando en la utilización de los sistemas de software libre en la medida de lo posible.
Por último, se correlacionaron los datos seleccionados de los sistemas internos de la Facultad con datos de sistemas externos a la Facultad de Informática, como son las redes sociales, tan usadas actualmente por millones de personas alrededor del mundo. Para este análisis no se tuvo una visión tan gerencial de la información como se tuvieron en los análisis anteriores, ya que se trabajó sobre los datos de los estudiantes relacionados a una sola materia seleccionada.

Las correlaciones de los datos entre los diferentes sistemas fueron posibles ya que cada uno almacena al menos un dato sobre el estudiante ó egresado que lo identifica unívocamente, este dato puede ser el DNI ó el legajo. Si bien estos datos no se usan para realizar los análisis en sí, fueron muy útiles en las etapas de preprocesamiento y transformación.
Además de correlacionar datos de los estudiantes, fue necesario correlacionar datos almacenados en los diferentes sistemas sobre las materias dictadas en la Facultad de Informática; para esto se utilizó un código de materia almacenados en los diferentes sistemas y en ocasiones fue necesario la correlación mediante el nombre de la materia.

De las tareas realizadas sobre el conjunto de datos al que se tuvo acceso, pertenecientes al cubo 04 del sistema SIU-Guaraní, se puede observar que existe una relación entre el nivel de estudios de los padres de los estudiantes con la situación que los mismos tienen dentro de la Facultad, siendo los estudiantes más activos, aquellos cuyos padres tienen un nivel de estudios secundario o superior.
En cuanto a la variable que indica la procedencia del estudiante, se puede observar, que no tiene una correlación directa con la situación de los estudiantes en la carrera. Teniendo en cuenta los datos obtenidos, se puede destacar que, los estudiantes procedentes de las localidades que forman parte del grupo denominado "Gran La Plata", son en su mayoría pasivos. Bajo la dominación "Gran la Plata", se agrupan las localidades no limítrofes a la ciudad de La Plata, que se encuentran a una distancia NO mayor de 150 Km. de la misma. Para el resto de los valores posibles de esta variable, "La Plata y alrededores", "Interior de la provincia de Buenos Aires", "Interior de país", "Exterior" y "Sin datos", la correlación se encuentra ligada a la variable de nivel de estudio de los padres. Por lo expresado anteriormente, no podemos hacer una correlación directa entre las variables procedencia y la situación del estudiante en las carreras de la facultad.

Los histogramas de la sección *8.1.1 - Distribución de los datos*, muestran una distribución más equilibrada de estudiantes activos y pasivos, en los procedentes del interior de la provincia de Buenos Aires y del interior de país. En las demás procedencias, "La Plata y alrededores", "Gran La Plata", "Exterior" y "Sin datos", se observa una clara diferencia con más estudiantes pasivos que activos.

La variable que indica el género del estudiante, no tiene una correlación con la situación del mismo en la carrera, esto puede ser debido a la amplia diferencia que existe entre la cantidad de estudiantes masculinos sobre la cantidad de estudiantes femeninos que se encuentran cursando alguna carrera de la facultad; esta diferencia se refleja en el histograma de la sección *8.1.1 - Distribución de los datos*, donde se refleja que la cantidad de estudiantes de género masculino representan a más del 80% de los estudiantes de las carreras de la facultad.

De las tareas realizadas sobre los datos correlacionados de los diferentes sistemas de la Facultad de Informática, se puede observar que existe una relación entre el uso pasivo de los foros del sistema Moodle y el resultado que los estudiantes obtienen en las materias, cuanto más es el uso de los foros del sistema, más es la cantidad de estudiante aprobados que se tiene.

Las variables que indican el nivel del uso del sistema con una actividad específica como por ejemplo rtas_docentes, rta_compañeros ó conversaciones_iniciadas, son las variables menos significativas en el momento de realizar la clasificación de los estudiantes.
En los histogramas de la sección *8.2.1 Distribución de los datos*, se puede observar que cuanto mayor es el tiempo que se emplea en las actividades virtuales de las materias en general, –indicado por las variables tiempo_foro y tiempo_materia – y cuanta más participación tiene el alumno en el foro –indicado por la variable participacion_foro – , mayor es la cantidad de estudiantes aprobados.
Se observa también, que una participación activa en los foros, como los indican por ejemplo las variables conversaciones_iniciadas, rta_compañeros ó rtas_docentes, no influye en el resultado que obtiene el estudiante en la materia.

También se puede observar una correlación entre los estudiantes que hacen uso de bibliografía relacionada a las materias, proveniente de la biblioteca de la Facultad, con el resultado obtenido en las materias; siendo generalmente aprobados los estudiantes que hacen uso de bibliotecas. En el histograma correspondiente, se puede observar una amplia diferencia entre el uso y el no uso de la biblioteca por parte de los estudiantes, habiendo menos del 10% de estudiantes que hicieron uso de bibliografía, relacionada a alguna de las materias, pertenecientes a la biblioteca de la facultad. Esto se puede deber a la difusión de material de estudios por otros diversos medios.

Al incorporar la variable de uso de redes sociales, en la extensión de este trabajo, vemos que éstas no se correlacionan directamente con la variable dependiente que indica el resultado obtenido por el estudiante en la materia. Esto se puede deber a la poca cantidad de estudiantes, que por el momento optaron por estas herramientas como medio de comunicación con los responsables de la cátedra y sus compañeros. Siguen siendo en este análisis el tiempo total empleado en la materia y en los foros, las variables independientes que mejor clasifican a los estudiantes.

Cuando se estudian los datos almacenados para determinar patrones de comportamiento entre ellos, la completitud y consistencia de los mismos es importante para que los resultados obtenidos, tanto en los modelos y patrones como el porcentaje de precisión sea un reflejo fiel de la realidad.

Los resultados obtenidos, cumplen con los objetivos del proyecto. Fue posible correlacionar datos de diferentes sistemas utilizados por la Facultad de Informática para obtener información de los mismos. La información obtenida supera el porcentaje de precisión establecido.
Fue posible integrar los algoritmos aplicados junto con los resultados obtenidos, en una aplicación Ad-Hoc de fácil uso para los usuarios, que muestre de manera sencilla y amigable, el trabajo realizado con la herramienta Weka, en esta tesina.

Las prestaciones de la herramienta WEKA, permitieron la realización, de una manera fácil, del preprocesamiento de los datos, que incluyeron, entre otras, transformaciones como por ejemplo la discretización.
Fue posible realizar las diferentes tareas de clasificación y agrupamiento, utilizando diferentes parámetros para los mismos y comparando los resultados obtenidos; además de la visualización gráfica y su integración en un sistema Ad-Hoc propio.
Esta herramienta cuenta con una documentación completa, disponible en la Web, generada con ejemplos claros, que ayudaron a la resolución de cualquier inconveniente presentado durante su uso.

Trabajos futuros

La Facultad de Informática cuenta con otros sistemas de almacenamiento de datos además de los que se utilizaron para este trabajo; entre ellos podemos nombrar el sistema que registra los datos de los ingresantes, el sistema de egresados, postgrados y otros sistemas de aprendizaje utilizados por otras cátedras como son el WebUNLP, ó Moodle instanciado para otras cátedras, diferentes a las que se utilizaron para este trabajo; también el sistema SIU cuenta con otros datos de los estudiantes y egresados que quedaron fuera del alcance de este trabajo de grado.
De todos estos sistemas se pueden obtener datos que al ser integrados o correlacionados ayuden para poder hacer un seguimiento del comportamiento de los estudiantes en cada una de las etapas de su carrera universitaria y post universitaria.

La información obtenida sobre el uso que los estudiantes hacen de las redes sociales, puede ser más extensa que la usada en este trabajo de grado. Se podría determinar el comportamiento que los alumnos tienen sobre las redes sociales, el tipo de uso que hacen de ellas y correlacionar este uso con el uso que hacen de los sistemas de la Facultad de Informática.
También, se podría extraer datos de otras redes sociales como Linkedin, una red social de profesionales, para poder determinar el comportamiento de estudiantes y graduados en el mundo profesional.

En este trabajo de grado, se obtiene la información desde un punto de vista gerencial o más a alto nivel, ya que se tiene en cuenta el comportamiento general de la situación académica del estudiante en todos sus años de universidad, así como también se mira el comportamiento de los estudiantes en las materias desde una visión global teniendo en cuenta el comportamiento en todas las materias juntas.
Se puede realizar un análisis que no sea tan general, que realice un tracking de alumnos, teniendo en cuenta la situación académica de los estudiantes en las diferentes etapas vividas en la universidad. De la misma manera se puede tener un análisis propio de cada materia teniendo en cuenta que el comportamiento que los estudiantes tienen en cada una de ellas puede ser variado.

La aplicación presentada junto con este informe, se puede extender para que el usuario pueda seleccionar su propio conjunto de datos, inclusive conectándose directamente a la base de datos que contiene la información. También se podría pensar en una aplicación Web, aprovechando las facilidades que provee la herramienta WEKA para la creación de aplicaciones JAVA.

Apéndice A

SCRIPTS DE BASE DE DATOS

Scripts para el análisis social y demográfico

A continuación se muestran los scripts realizados para la creación de la vista minable para el análisis desde el punto de vista de factores sociales y demográficos.

1 Creación y llenado de las tablas en el ambiente local

```
#Creamos la base de datos:
CREATE DATABASE siu;
#Creación y llenado de la tabla fáctica ft_desgranamiento_persua
CREATE TABLE `FT_Desgranamiento_PersUA` (
  `AñoAcademico` int(10) DEFAULT NULL,
  `CodNivelActividad` int(11) DEFAULT NULL,
  `CodNivelActividadAcu` int(11) DEFAULT NULL,
  `Cohorte` int(11) DEFAULT NULL, `Sexo` varchar(2) DEFAULT NULL,
  `CodSitEst` varchar(10) DEFAULT NULL,
  `SituacionEstudiante` varchar(45) DEFAULT NULL,
  `CodUA` int(11) DEFAULT NULL, `CodCarrera` varchar(10) DEFAULT NULL,
  `CodNivelEstudios` int(11) DEFAULT NULL,
  `DescripcionEstudiosPadres` varchar(45) DEFAULT NULL,
  `CodTipoTituloSecundario` varchar(10) DEFAULT NULL,
  `CodColegio` int(11) DEFAULT NULL, `dni` varchar(45) DEFAULT NULL);
LOAD DATA INFILE 'cubo0004sql1.csv' INTO TABLE FT_Desgranamiento_PersUA
FIELDS TERMINATED BY '\;' LINES TERMINATED BY '\n';
#Creación y llenado de las tablas dimensiones
#Creación y llenado de la tabla  lt_sexos
CREATE TABLE `lt_sexos` (
  `CodSexo` int(11) NOT NULL, `DescSexo` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`CodSexo`));
LOAD DATA INFILE 'datos-sexo.csv' INTO TABLE lt_sexos FIELDS TERMINATED
BY '\;' LINES TERMINATED BY '\n';
#Creación y llenado de la tabla lt_situacionEstudiante
CREATE TABLE `lt_situacionestudiantes` (
  `CodSitEst` varchar(11) NOT NULL,
 `DescSitEst` varchar(45) DEFAULT NULL, PRIMARY KEY (`CodSitEst`));

LOAD DATA INFILE 'datos-situacionEstudiantes.csv' INTO TABLE
lt_situacionestudiantes FIELDS TERMINATED BY '\;' LINES TERMINATED BY
'\n';
#Creación y llenado de la tabla lt_carreras
CREATE TABLE `lt_carreras` (
  `CodCarrera` int(11) NOT NULL,
```

```
  `NombreCarrera` varchar(45) DEFAULT NULL, PRIMARY KEY (`CodCarrera`));
LOAD DATA INFILE 'datos-carreras.csv' INTO TABLE lt_carreras FIELDS
TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

#Creación y llenado de la tabla lt_nivelActividadAnual

```
CREATE TABLE `lt_nivelactividadanual` (
  `CodNivelActividad` int(11) NOT NULL,
  `DescNivelActividad` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`CodNivelActividad`));
LOAD DATA INFILE 'datos-nivelActividadAnual.csv' INTO TABLE
lt_nivelactividadanual FIELDS TERMINATED BY '\;' LINES TERMINATED BY
'\n';
```

#Creación y llenado de la tabla lt_nivelActividadAcumulado

```
CREATE TABLE ` lt_nivelactividadacumulado` (
  `CodNivelActividadAcu` int(11) NOT NULL,
  `DescNivelActividadAcu` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`CodNivelActividadAcu`));

LOAD DATA INFILE 'datos-nivelActividadAcumulado.csv' INTO TABLE
lt_nivelactividadacumulado FIELDS TERMINATED BY '\;' LINES TERMINATED BY
'\n';
```

#Creación y llenado de la tabla lt_colegios

```
CREATE TABLE `lt_colegios` (
  `CodColegio` int(11) NOT NULL,
  `NombreColegio` varchar(100) DEFAULT NULL,
  `NombreLocalidad` varchar(100) DEFAULT NULL,
  `NombreProvincia` varchar(100) DEFAULT NULL,
  `nombrepais` varchar(100) DEFAULT NULL,
  `Procedencia` varchar(100) DEFAULT NULL, PRIMARY KEY (`CodColegio`));
LOAD DATA INFILE 'datos-colegios.csv' INTO TABLE lt_colegios FIELDS
TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

#Creación y llenado de la tabla lt_localidades_colsec

```
CREATE TABLE `lt_localidades_colsec` (
  `CodLocalidad` int(11) NOT NULL,
  `NombreLocalidad` varchar(45) DEFAULT NULL,
  `CodProvincia` int(11) DEFAULT NULL, PRIMARY KEY (`CodLocalidad`));
LOAD DATA INFILE 'datos-localidades_colsec.csv' INTO TABLE
lt_localidades_colsec FIELDS TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

#Creación y llenado de la tabla lt_provincias

```
CREATE TABLE `lt_provincias` (
  `CodProvincia` int(11) NOT NULL,
  `NombreProvincia` varchar(45) DEFAULT NULL,
  `CodPais` int(11) DEFAULT NULL, PRIMARY KEY (`CodProvincia`));
LOAD DATA INFILE 'datos-provincias.csv' INTO TABLE lt_provincias FIELDS
TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

#Creación y llenado de la tabla lt_paises

```
CREATE TABLE `lt_paises` (
  `CodPais` int(11) NOT NULL, `NombrePais` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`CodPais`));
LOAD DATA INFILE 'datos-paises.csv' INTO TABLE lt_paises FIELDS
TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

#Creación y llenado de la tabla lt_estudiosPadres

```
CREATE TABLE `lt_estudiospadres` (
  `CodNivelEstudios` int(11) NOT NULL,
  `DescNivelEstudios` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`CodNivelEstudios`));
LOAD DATA INFILE 'datos-estudiosPadres.csv' INTO TABLE lt_estudiospadres
FIELDS TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

#Creación y llenado de la tabla lt_tipoTitSec

```
CREATE TABLE `lt_tipotitsec` (
  `CodTipoTituloSecundario` int(11) NOT NULL,
  `NombreTipoTituloSecundario` varchar(45) DEFAULT NULL,
  PRIMARY KEY (`CodTipoTituloSecundario`));
LOAD DATA INFILE 'datos-tipoTitSec.csv' INTO TABLE lt_tipotitsec FIELDS
TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

*/*Exportación e importación de los datos de los egresados */*

```
SELECT  t.nro_inscripcion, t.titulo, t.carrera, t.legajo, t.fecha_egreso,
       t.duracion_carrera, c.periodo_inscripcio, p.nro_documento
FROM sga_titulos_otorg t, sga_carrera_aspira c, sga_personas p
WHERE t.unidad_Academica = c.unidad_academica
AND t.carrera = c.carrera AND t.nro_inscripcion = c.nro_inscripcion
AND c.unidad_academica = p.unidad_academica
AND c.nro_inscripcion = p.nro_inscripcion AND c.periodo_inscripcio >1999
AND t.carrera IN ('LI', 'LS') AND t.titulo IN ('LI', 'LS')
AND c.situacion_asp IN ('AC', 'IL', 'IC');
```

#Creación y llenado de la tabla lt_egresados.

```
CREATE TABLE `LT_Egresados` (
  `nro_inscripcion` int(11) NOT NULL, `titulo` varchar(45) DEFAULT NULL,
  `carrera` varchar(45) DEFAULT NULL, `legajo` varchar(45) DEFAULT NULL,
  `fecha_egreso` varchar(45) DEFAULT NULL,
  `duracion_carrera` int(10) DEFAULT NULL,
  `periodo_inscripcion` int(10) DEFAULT NULL,
  `dni` varchar(45) DEFAULT NULL, PRIMARY KEY (`nro_inscripcion`),
  KEY `dni_inx` (`dni`));
LOAD DATA INFILE 'egresados.csv' INTO TABLE LT_Egresados FIELDS
TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

2 Creación y llenado de la tabla

FT_Desgranamiento_PersUA_Licenciatura

```
CREATE TABLE FT_Desgranamiento_Persua_Licenciatura
SELECT * FROM FT_Desgranamiento_PersUA WHERE codCarrera IN ('LI','LS');
```

3 Atributo Procedencia

#Creación del atributo procedencia

```
ALTER TABLE `FT_Desgranamiento_Persua_Licenciatura` ADD COLUMN
`Procedencia` VARCHAR(255) DEFAULT NULL ;
```

#Registros con la procedencia "La Plata y Alrededores"

```
UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'La Plata y Alrededores'
WHERE codColegio IN
(SELECT cod_colegio
 FROM  LT_Colegios col, LT_Localidades_ColSec loc
 WHERE col.codLocalidad = loc.codLocalidad
   AND loc.NombreLocalidad IN ('La Plata', 'Berazategui', 'Berisso',
'Brandsen', 'Ensenada', 'Florencio Varela', 'Magdalena', 'San Vicente'));
```

#Registros con la procedencia "Gran La Plata"

```
UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'Gran La Plata'
WHERE codColegio IN
(SELECT codColegio
 FROM LT_Colegios col, LT_Localidades_ColSec loc
 WHERE col.codLocalidad = loc.codLocalidad
```

```
   AND loc.NombreLocalidad IN ('Almirante Brown', 'Campana','Castelli',
'Chascomús', 'Escobar', 'General Belgrano', 'Lobos', 'Lomas de Zamora',
'Lujan', 'Monte', 'Pila', 'Quilmes', 'Vicente López'));

UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'Gran La Plata'
WHERE codColegio IN
(SELECT codColegio
 FROM LT_Colegios col, LT_Localidades_ColSec loc, LT_Provincias prov
 WHERE col.codLocalidad = loc.codLocalidad
   AND loc.codProvincia = prov.codProvincia
   AND p.NombreProvincia = 'Capital Federal');
```

#Registros con la procedencia "Interior de la provincia de Buenos Aires"

```
UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'Int. de la Pcia. de Bs.As.'
WHERE prodecencia IS NULL AND codColegio IN
(SELECT codColegio
 FROM LT_Colegios col, LT_Localidades_ColSec loc, LT_Provincias prov
 WHERE col.codLocalidad = loc.codLocalidad
   AND loc.codPovincia = prov.codProvincia
   AND loc.nombreLocalidad != 'Indeterminado'
   AND prov.NombreProvincia = 'Buenos Aires');
```

#Registros con la procedencia "Interior del país"

```
UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'Interior del país'
WHERE procedencia IS NULL AND codColegio IN
(SELECT codColegio
  FROM LT_Colegios col, LT_Localidades_ColSec loc, LT_Provincias prov,
LT_Paises pais
WHERE col.codLocalidad = loc.codLocalidad
  AND loc.codPovincia = prov.codProvincia
  AND prov.codPais = pais.CodPais AND pais.NombrePais = 'Argentina');
```

#Registros con la procedencia "Exterior"

```
UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'Exterior'
WHERE codColegio IN
(SELECT codColegio
FROM LT_Colegios col, LT_Localidades_ColSec loc, LT_Provincias prov,
LT_Paises pais
WHERE col.codLocalidad = loc.codLocalidad
  AND loc.codPovincia = prov.codProvincia AND prov.codPais = pais.CodPais
  AND pais.NombrePais = 'Otros');
```

#Registros con la procedencia "Sin Datos"

```
UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'Sin Datos'
WHERE codColegio NOT IN (SELECT codColegio FROM LT_Colegios col);

UPDATE FT_Desgranamiento_Persua_Licenciatura
SET procedencia = 'Sin Datos'
WHERE prodecencia IS NULL AND codColegio IN
(SELECT codColegio
  FROM LT_Colegios col, LT_Localidades_ColSec loc, LT_Provincias prov
  WHERE col.codLocalidad = loc.codLocalidad
    AND loc.codPovincia = prov.codProvincia
    AND loc.nombreLocalidad = 'Indeterminado'
    AND prov.NombreProvincia = 'Buenos Aires');
```

4 Atributo Situación del Estudiante

#Creación del atributo Situación del Estudiante

```
ALTER TABLE `FT_Desgranamiento_Persua_Licenciatura`
ADD COLUMN `SituacionDelEstudiante`
VARCHAR(255) DEFAULT NULL;
```

*/*Estudiantes pasivos "P - Pasivo". Nunca tuvieron actividad, tuvieron actividad solamente el primer año de la carrera, ó tienen más años sin actividad que con actividad. */*

#Alumnos sin actividad

```
UPDATE FT_Desgranamiento_Persua_Licenciatura
SET situacionDelEstudiante = 'P - Pasivo' WHERE codsitest = 'A';

UPDATE ft_desgranamiento_persua_licenciatura
SET situaciondelestudiante = 'P - Pasivo'
WHERE dni IN (SELECT dni FROM (
SELECT DISTINCT(dni)
FROM ft_desgranamiento_persua_licenciatura ft1
WHERE situaciondelestudiante IS NULL
  AND NOT EXISTS (SELECT 1
                  FROM ft_desgranamiento_persua_licenciatura ft2
                  WHERE ft2.dni = ft1.dni
                    AND ft2.codnivelactividad <> 0)) AS t);
```

#Alumnos con actividad solamente en el primer año de carrera.

```
UPDATE ft_desgranamiento_persua_licenciatura
SET situaciondelestudiante = 'P - Pasivo'
WHERE dni IN (SELECT dni FROM(
SELECT dni
FROM ft_desgranamiento_persua_licenciatura ft1
WHERE situaciondelestudiante IS NULL
  AND ft1.añoacademico = ft1.cohorte AND codnivelactividad <> 0
  AND cohorte <> 2012 AND NOT EXISTS
 (SELECT 1 FROM ft_desgranamiento_persua_licenciatura ft2
  WHERE ft2.dni = ft1.dni AND ft2.cohorte <> ft2.añoacademico
   AND ft2.codnivelactividad <> 0)) AS t);
```

#Alumnos que tienen más años sin actividad que con actividad.

```
UPDATE ft_desgranamiento_persua_licenciatura
SET situacionDelEstudiante = 'P - Pasivo'
WHERE dni IN (
SELECT DISTINCT(dni) FROM(
SELECT añoacademico, dni, cohorte, codnivelactividad,
  (SELECT count(*)FROM ft_desgranamiento_persua_licenciatura ft2
    WHERE ft2.dni = ft1.dni AND codnivelactividad = 0) AS sin_actividad,
  (SELECT count(*) FROM ft_desgranamiento_persua_licenciatura ft2
    WHERE ft2.dni = ft1.dni AND codnivelactividad <> 0) AS con_actividad
FROM ft_desgranamiento_persua_licenciatura ft1
WHERE situaciondelestudiante IS NULL
HAVING sin_actividad > con_actividad) AS t);
```

#Estudiantes activos "A – Activo"

```
UPDATE ft_desgranamiento_persua_licenciatura
SET situaciondelestudiante = 'A - Activo'
WHERE dni IN (SELECT dni FROM (
SELECT DISTINCT(dni)
FROM ft_desgranamiento_persua_licenciatura ft1
WHERE situaciondelestudiante IS NULL
  AND NOT EXISTS
(SELECT 1 FROM ft_desgranamiento_persua_licenciatura ft2
WHERE ft2.dni = ft1.dni AND
(ft2.codnivelactividad = 0 OR ft2.codnivelactividad = 1))) AS t);
```

```
#Alumnos Egresados
/*Cantidad de años en que el alumno culmina la carrera de licenciatura. */
CREATE  TABLE tmp
SELECT ft.dni,ft.cohorte,e.fecha_egreso,
       CONVERT(SUBSTRING(e.fecha_egreso,7),UNSIGNED INTEGER) - ft.cohorte
        AS cantidad_años
FROM FT_Desgranamiento_Persua_Licenciatura ft, lt_egresados e
WHERE ft.dni = e.dni
ORDER BY ft.dni,ft.añoacademico;
/*Actualización de los registros de la tabla FT_Desgranamiento_Persua_Licenciatura */
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET situacionDelEstudiante = 'ET - Egreso en término'
WHERE dni IN (SELECT dni FROM tmp WHERE cantidad_años <= 7);
/*Egresados fuera de  término (mayor a 7 años).*/
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET situacionDelEstudiante = 'EFT - Egreso fuera de término'
WHERE dni IN (SELECT dni FROM tmp WHERE cantidad_años >= 8);
```

5 Atributo Nivel Estudio Padres

```
#Creación del atributo Nivel Estudio Padres
ALTER TABLE FT_Desgranamiento_PersUA_Licenciatura
ADD COLUMN NivelEstudiosPadres VARCHAR(45) NULL;
#Registros con nivel estudios padres “Sin estudios”
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET NivelEstudiosPadres = 'Sin estudios'
WHERE codnivelestudios = 1 OR codnivelestudios = 2;
#Registros con nivel estudios padres “Estudios Primarios”
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET NivelEstudiosPadres = 'Estudios Primarios'
WHERE codnivelestudios = 3 OR codnivelestudios = 4;
#Registros con nivel estudios padres “Estudios Secundarios”
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET NivelEstudiosPadres = 'Estudios Secundarios'
WHERE codnivelestudios = 5 OR codnivelestudios = 6;
#Registros con nivel estudios padres “Estudios Universitarios”
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET NivelEstudiosPadres = 'Estudios Universitarios'
WHERE codnivelestudios = 7 OR codnivelestudios = 12;
#Registros con nivel estudios padres “Sin Datos”
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET NivelEstudiosPadres = 'Sin datos'
WHERE codnivelestudios = 0 OR codnivelestudios = 13;
```

6 Atributo Género

```
#Creación del atributo Género.
ALTER TABLE `siu`.`ft_desgranamiento_persua_licenciatura`
ADD COLUMN `Genero` VARCHAR(10) NULL DEFAULT NULL;
#Registros con genero ‘F’
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET genero = 'F' WHERE sexo = 2;
#Registros con genero ‘M’
```

```
UPDATE FT_Desgranamiento_PersUA_Licenciatura
SET genero = 'M' WHERE sexo = 1;
```

Scripts para el análisis de la participación social en las materias

A continuación se muestran los scripts realizados para la creación de la vista minable para el análisis de la participación social de los estudiantes en las materias.

7 Importación de la base de datos cátedras al ambiente local

```
/**Importación de la base de datos Moodle a un ambiente local.*/
#Creamos la base de datos:
CREATE DATABASE catedras;
#Ahora le damos la ubicación de nuestro archivo .sql
SOURCE C:\Users\Claudia\Desktop\catedas.sql
```

8 Recursos más usados por las materias en Moodle

```
/**Usos de cada  recurso de Moodle en las materias*/
SELECT course, module, count(*) as cant
FROM mdl_log
WHERE course IN (SELECT id FROM mdl_tesis_course_fac_informatica)
  AND module in ('forum', 'survey', 'blog', 'discussion', 'message',
      'workshop', 'quiz', 'chat', 'wiki', 'openmeetings', 'twitter')
GROUP BY course, module ORDER BY course, cant DESC;
```

9 Tablas del SIU-Guaraní para el análisis de participación social en las materias

```
# lt_materias
CREATE TABLE `lt_materias` (
  `codMateria` int(10) NOT NULL, `descMateria` varchar(255) DEFAULT NULL,
  `descCortaMateria` varchar(45) DEFAULT NULL);
LOAD DATA INFILE 'datos-materias.csv' INTO TABLE lt_materias FIELDS
TERMINATED BY '\;' LINES TERMINATED BY '\n';
# lt_alumnosMaterias
CREATE TABLE `lt_alumnosMaterias` (
  `legajoAlumno` int(10) NOT NULL, `codMateria` int(10) DEFAULT NULL,
  `fecha` timestamp DEFAULT NULL);
LOAD DATA INFILE 'datos-alumnos_materias.csv' INTO TABLE
lt_alumnosMaterias FIELDS TERMINATED BY '\;' LINES TERMINATED BY '\n';
```

10 Creación y llenado de la taba mdl_tesis_course_fac_informatica

```
#Creación de la tabla  mdl_tesis_course_fac_informatica
CREATE TABLE `mdl_tesis_course_fac_informatica` (
  `id` INT(10) UNSIGNED NOT NULL AUTO_INCREMENT,
  `category` INT (10) UNSIGNED NOT NULL DEFAULT '0',
  `sortorder` INT (10) UNSIGNED NOT NULL DEFAULT '0',
  `fullname` VARCHAR(254) NOT NULL DEFAULT '',
```

```
  `shortname` VARCHAR (255) NOT NULL DEFAULT '',
  `idnumber` VARCHAR (100) DEFAULT NULL, `summary` TEXT,
  `summaryformat` TINYINT (2) UNSIGNED NOT NULL DEFAULT '0',
  `format` VARCHAR (10) NOT NULL DEFAULT 'topics',
  `showgrades` SMALLINT(2) UNSIGNED NOT NULL DEFAULT '1',
  `modinfo` LONGTEXT,
  `newsitems` SMALLINT (5) UNSIGNED NOT NULL DEFAULT '1',
  `startdate` INT (10) UNSIGNED NOT NULL DEFAULT '0',
  `numsections` SMALLINT (5) UNSIGNED NOT NULL DEFAULT '1',
  `marker` INT (10) UNSIGNED NOT NULL DEFAULT '0',
  `maxbytes` INT (10 UNSIGNED DEFAULT '0',
  `legacyfiles` SMALLINT (4) UNSIGNED NOT NULL DEFAULT '0',
  `showreports` INT (4) UNSIGNED DEFAULT '0',
  `groupmode` INT (4) UNSIGNED DEFAULT '0',
  `groupmodeforce` INT (4) UNSIGNED DEFAULT '0',
  `defaultgroupingid` bigint(10) UNSIGNED NOT NULL DEFAULT '0',
  `lang` VARCHAR (30) NOT NULL DEFAULT '',
  `theme` VARCHAR (50) DEFAULT NULL,
  `visible` INT (10) UNSIGNED NOT NULL DEFAULT '1',
  `visibleold` TINYINT (1) UNSIGNED NOT NULL DEFAULT '1',
  `hiddensections` INT (2) UNSIGNED NOT NULL DEFAULT '0',
  `timecreated` INT (10) UNSIGNED NOT NULL DEFAULT '0',
  `timemodified` INT (10) UNSIGNED NOT NULL DEFAULT '0',
  `requested` INT (10) UNSIGNED NOT NULL DEFAULT '0',
  `restrictmodules` INT (11 NOT NULL DEFAULT '0',
  `enablecompletion` TINYINT (1) UNSIGNED NOT NULL DEFAULT '0',
  `completionnotify` TINYINT (1) UNSIGNED NOT NULL DEFAULT '0',
  `completionstartonenrol` TINYINT (1) UNSIGNED NOT NULL DEFAULT '0',
  PRIMARY KEY (`id`),  KEY `idnumber` (`idnumber`),
  KEY `mdl_cour_sho_ix` (`shortname`));
```

#Correlación entre shortname y cod_materia.

```
INSERT INTO mdl_course_fac_informatica
SELECT * FROM mdl_course
WHERE shortname IN (SELECT cod_materia FROM siu.lt_materias);
```

#Cursos pertenecientes a la categoría 14 – SIU-Guaraní.

```
INSERT INTO mdl_course_fac_informatica
SELECT * FROM mdl_course
WHERE substring_index(shortname, '-',1) IN (SELECT cod_materia
                                              FROM siu.lt_materias)
  AND category IN (14);
```

#Se actualiza el shortname.

```
UPDATE mdl_course_fac_informatica
SET shortname = substring_index(shortname, '-',1)
WHERE category = 14;
```

#Cursos pertenecientes a la categoría 21 – Cursos 2011. Cuyo shortname no se corresponde con el cod_materia..

```
INSERT INTO mdl_tesis_course_fac_informatica
SELECT * FROM mdl_course
WHERE substring_index(shortname, '-',1) IN (SELECT cod_materia
                                              FROM siu.lt_materias)
 AND id NOT IN (SELECT id FROM mdl_course_fac_informatica)
 AND category IN (21);
UPDATE mdl_tesis_course_fac_informatica
SET shortname = substring_index(shortname, '-',1)
WHERE category = 21;
```

#Eliminación de los cursos pertenecientes a las sedes de Tres Arroyos ó Las Flores

```
DELETE FROM mdl_tesis_course_fac_informatica
```

```
WHERE fullname LIKE '%Tres Arroyos%' OR fullname LIKE '%Las Flores%'
   OR fullname LIKE '%3 Arroyos%';
```

11 Creación y llenado de la taba mdl_tesis_courses_students

#Creación de la tabla auxiliar mdl_tesis_couses_students

```
CREATE TABLE `mdl_tesis_courses_students` (
  `userid` INT(10) DEFAULT NULL, `username` VARCHAR(100) DEFAULT NULL,
  `firstname` VARCHAR(100) DEFAULT NULL,
  `lastname` VARCHAR(100) DEFAULT NULL,
  `legajo` VARCHAR(100) DEFAULT NULL, `courseid` INT(10) DEFAULT NULL,
  `fullname` VARCHAR(254) DEFAULT NULL,
  `shortname` VARCHAR(150) DEFAULT NULL,
  `assignyear` INT(10) DEFAULT NULL,
  `coursestartyear` INT(10) DEFAULT NULL,
  `coursetimecreated` INT(10) DEFAULT NULL,
  `coursetimemodified` INT(10) DEFAULT NULL,
  `success` INT(2) DEFAULT NULL, KEY `legajo_idx` (`legajo`));
```

#Alumnos que cursaron cada materia

```
INSERT INTO mdl_tesis_courses_students
SELECT usr.id AS userid, usr.username, usr.firstname, usr.lastname,
       data.data AS legajo, c.id AS courseid, c.fullname, c.shortname,
       FROM_UNIXTIME(rol.timemodified, '%Y') AS assignyear,
       FROM_UNIXTIME(c.startdate, '%Y') AS coursestartyear,
       FROM_UNIXTIME(c.timecreated, '%Y') AS coursecreatedyear,
       FROM_UNIXTIME(c.timemodified, '%Y') AS coursemodifiedyear,0
  FROM mdl_context ctx, mdl_tesis_course_fac_informatica c,
       mdl_role_assignments rol, mdl_user usr, mdl_user_info_data data
  WHERE ctx.instanceid = c.id
    AND ctx.contextlevel = 50 #50 es el contexto de cursos
    AND rol.contextid = ctx.id
    AND rol.roleid = 5 #5 es el id del rol Estudiante
    AND rol.userid = usr.id
    AND data.userid = usr.id
    AND data.fieldid = 5; #5 es el id del field legajo
```

12 Creación y llenado de la taba mdl_tesis_course_students_note

#Creación de la tabla mdl_tesis_course_students_note

```
CREATE TABLE `mdl_tesis_course_students_note` (
  `courseid` INT(10) DEFAULT NULL, `fullname` VARCHAR(254) DEFAULT NULL,
  `shortname` VARCHAR(150) DEFAULT NULL, `userid` INT(10) DEFAULT NULL,
  `legajo` VARCHAR(45) DEFAULT NULL,`año_academico` INT(10) DEFAULT NULL,
  `m_cursada` INT(1) DEFAULT NULL, `año_ final` INT(11) DEFAULT NULL,
  KEY `cs_usercourse` (`courseid`,`userid`), KEY `cs_user` (`userid`),
  KEY `cs_course` (`courseid`));
```

#Inserto todos los alumnos que aprobaron la materia que cursaron.

```
INSERT INTO mdl_course_students_note
SELECT courseid, fullname, shortname,userid, legajo, assignyear,1,null
```

```
FROM mdl_tesis_courses_students cs, iu.lt_alumnos_materias am
WHERE cs.legajo = am.legajo_alumno AND cursada = 1
  AND FIND_IN_SET(am.cod_materia,cs.shortname) > 0
  AND assignyear != 1969 AND assignyear = año_academico
  GROUP BY courseid, legajo ORDER BY assignyear,courseid;
#Inserto todos los alumnos que desaprobaron la materia que cursaron.
INSERT INTO mdl_course_students_note
SELECT courseid, fullname, shortname,userid, legajo, assignyear,0,null
  FROM mdl_courses_students cs, siu.lt_alumnos_materias am
WHERE cs.legajo = am.legajo_alumno AND cursada = 1
  AND FIND_IN_SET(am.cod_materia,cs.shortname)>0
  AND assignyear != 1969
  AND assignyear != año_academico AND assignyear < año_academico
  GROUP BY courseid, legajo ORDER BY courseid,legajo;
/**Los alumnos-materias que no están en LT_Alumnos_Materias, también se los considera
desaprobados.*/
INSERT INTO mdl_course_students_note
SELECT courseid, fullname, shortname, userid, legajo, assignyear,0,null
  FROM mdl_courses_students cs
WHERE NOT EXISTS (SELECT 1 FROM siu.lt_alumnos_materias am
                  WHERE am.legajo = cs.legajo
                    AND FIND_IN_SET(am.cod_materia, cs.shortname) > 0);
```

13 Participación de los alumnos en los foros

```
/**Creación de la tabla que contiene la participación total de los alumnos en los foros */
CREATE TABLE mdl_tesis_participacion_total_foros
SELECT cs.userid, cs.courseid,
       (SELECT count(log.action)
          FROM mdl_log log
          WHERE log.userid = cs.userid AND log.course = cs.courseid
          AND (log.module = 'forum' OR log.module = 'discussion')
          ) AS cant_participacion_foro
FROM mdl_tesis_course_students_note cs GROUP BY userid, courseid;
#Agrego la columna para discretizar el atributo cant_participacion_foro.
ALTER TABLE `mdl_tesis_participacion_total_foros`
ADD COLUMN `participación_foro` VARCHAR(45) NULL DEFAULT NULL;
#Participación baja en los foros.
UPDATE mdl_tesis_participacion_total_foros
SET participación_foro = 'Bajo' WHERE cant_participacion_foro <= 8;
#Participación media en los foros.
UPDATE mdl_tesis_participacion_total_foros
SET participación_foro = 'Medio'
WHERE cant_participacion_foro > 8 AND cant_participacion_foro <= 63;
#Participación alta en los foros.
UPDATE mdl_tesis_participacion_total_foros
SET participación_foro = 'Alto' WHERE cant_participacion_foro > 63;
```

```
/**Tabla con la cantidad de respuestas de alumnos a compañeros.*/
CREATE TABLE mdl_tesis_respuesta_post_compañeros
SELECT cs.userid, cs.courseid,
       (SELECT count(*)
          FROM mdl_forum_posts post, mdl_forum_posts post2,
               mdl_forum_discussions discussion
          WHERE post.userid = cs.userid
            AND post.discussion = discussion.id
            AND discussion.course = cs.courseid
          AND post.parent = post2.id
          AND post2.userid IN (
              SELECT co_st.userid
              FROM mdl_tesis_course_students_note co_st
              WHERE co_st.courseid = discussion.course)) AS
                                          cant_respuestas_post_compañeros
FROM mdl_tesis_course_students_note cs GROUP BY userid,courseid;
#Agrego la columna para discretizar el atributo cant_respuestas_posts_compañeros.
ALTER TABLE ` mdl_tesis_respuesta_post_compañeros `
ADD COLUMN `rta_compañeros` VARCHAR(45) NULL DEFAULT NULL;
#Baja cantidad de  respuestas a compañeros.
UPDATE mdl_tesis_respuesta_post_compañeros
SET rta_compañeros = 'Bajo' WHERE cant_respuestas_post_compañeros = 0;
#Cantidad media de  respuestas a compañeros.
UPDATE mdl_tesis_respuesta_post_compañeros
SET rta_compañeros = 'Medio' WHERE cant_respuestas_post_compañeros = 1;
#Alta cantidad de  respuestas a compañeros.
UPDATE mdl_tesis_respuesta_post_compañeros
SET rta_compañeros = 'Alto' WHERE cant_respuestas_post_compañeros > 1;

/**Tabla con cantidad de respuestas de alumnos a posts de docentes*/
CREATE TABLE mdl_tesis_respuesta_post_docentes
SELECT cs.userid, cs.courseid,
       (SELECT count(*)
          FROM mdl_forum_posts post, mdl_forum_posts post2,
               mdl_forum_discussions discussion
          WHERE post.userid = cs.userid
          AND post.discussion = discussion.id
          AND discussion.course = cs.courseid
          AND post.parent = post2.id
          AND post2.userid IN (
               SELECT a.userid
                 FROM mdl_context  con, mdl_role_assignments  a
                WHERE con.contextlevel = 50
                  AND con.instanceid = discussion.course
                  AND a.contextid = con.id
                  AND       a.roleid       IN       (2,3,4,7)))       AS
cant_respuestas_post_docentes
FROM mdl_tesis_course_students_note cs GROUP BY userid,courseid;
#Agrego la columna para discretizar el atributo cant_respuestas_posts_docentes.
```

```
ALTER TABLE  `mdl_tesis_respuesta_post_docentes`
ADD COLUMN `rtas_docentes` VARCHAR(45) DEFAULT NULL;

#Baja cantidad de  respuestas a docentes.
UPDATE mdl_tesis_respuesta_post_docentes
SET rtas_docentes = 'Bajo' WHERE cant_respuestas_post_docentes = 0;
#Cantidad media de  respuestas a docentes.
UPDATE mdl_tesis_respuesta_post_docentes
SET rtas_docentes = 'Medio' WHERE cant_respuestas_post_docentes = 1;
#Alta cantidad de  respuestas a docentes.
UPDATE mdl_tesis_respuesta_post_docentes
SET rtas_docentes = 'Alto' WHERE cant_respuestas_post_docentes > 1;

/**Conversaciones iniciadas por cada estudiantes en las materias*/
CREATE TABLE mdl_tesis_conversaciones_iniciadas
SELECT cs.userid, cs.courseid,
      (SELECT count(*)
         FROM mdl_forum_discussions discussion
        WHERE discussion.course = cs.courseid
          AND discussion.userid = cs.userid
          ) AS cant_conversaciones_iniciadas
FROM mdl_tesis_course_students_note cs GROUP BY userid,courseid;
#Agrego la columna para discretizar el atributo cant_conversaciones_iniciadas.
ALTER TABLE `mdl_tesis_conversaciones_iniciadas`
ADD COLUMN `conversaciones_iniciadas` VARCHAR(45) DEFAULT NULL;
#Baja cantidad de conversaciones iniciadas.
UPDATE mdl_tesis_conversaciones_iniciadas
SET conversaciones_iniciadas = 'Bajo'
WHERE cant_conversaciones_iniciadas = 0;
#Cantidad media de conversaciones iniciadas.
UPDATE mdl_tesis_conversaciones_iniciadas
SET conversaciones_iniciadas = 'Medio'
WHERE cant_conversaciones_iniciadas = 1;
#Alta cantidad de conversaciones iniciadas.
UPDATE mdl_tesis_conversaciones_iniciadas
SET conversaciones_iniciadas = 'Alto'
WHERE cant_conversaciones_iniciadas > 1;

/**Acciones de los estudiantes en las materias que cursan*/
CREATE TABLE mdl_tesis_tiempos
SELECT log.userid, log.course, log.module, log.action, log.time
  FROM mdl_log log, mdl_tesis_course_students_note cs
 WHERE log.userid = cs.userid AND log.course = cs.courseid
ORDER BY log.userid, log.course, log.time desc;

/**Creación y llenado de la tabla mdl_tesis_time_summarization*/
# Agrego una columna para transfromar a date la columna tiempo
```

```
ALTER TABLE mdl_tesis_tiempos ADD COLUMN fh_tiempo DATE;

#Paso a date la columna tiempo
UPDATE mdl_tesis_tiempos
SET fh_tiempo = to_date(time, 'YYYY-MM-DD HH24:MI:SS');
#Algoritmo para llenar los datos de la tabla mdl_time_summarization
/**Cursor que calcula el tiempo en minutos que dura cada una de las acciones.*/
DECLARE
CURSOR tiempos (p_usuario number) IS
      SELECT usuario, materia, module, action, fh_tiempo, next_fh_action,
             (next_fh_action  -  fh_tiempo)  *  (24  *  60  *  60)  AS
             TIEMPO_DE_LA_ACCION
        FROM (SELECT usuario, materia, module, action, fh_tiempo,
                     lead (fh_tiempo,1) over (ORDER BY usuario,
                     fh_tiempo   asc)  AS next_fh_action
                FROM mdl_forum_log
               WHERE usuario = p_usuario);
id_alumno number(10);
tiempo tiempos%rowtype;

PROCEDURE agregar_tiempo_materia(tiempo tiempos%rowtype) IS
existe_registro number; tiempo_maximo number;
BEGIN
  IF(tiempo.TIEMPO_DE_LA_ACCION > 10) THEN
    tiempo_maximo := 10; #Se toma un máximo de 10 minutos por acción
  ELSE
    tiempo_maximo := tiempo.TIEMPO_DE_LA_ACCION;
  END IF;

  SELECT count(*) INTO existe_registro
    FROM mdl_time_summarization
   WHERE usuario = tiempo.usuario AND materia = tiempo.materia;

  IF (existe_registro > 0) THEN
  BEGIN
     UPDATE mdl_time_summarization
     SET TOTAL_TIME_MATERIA = TOTAL_TIME_MATERIA + tiempo_maximo
     WHERE usuario = tiempo.usuario AND materia = tiempo.materia;

     IF (tiempo.module = 'forum') THEN
      UPDATE mdl_time_summarization
      SET TOTAL_TIME_FORUM = TOTAL_TIME_FORUM + tiempo_maximo
      WHERE usuario = tiempo.usuario AND materia = tiempo.materia;
     END IF;
   END;
  ELSE
    IF (tiempo.module = 'forum') THEN
     INSERT INTO  mdl_time_summarization
     VALUES (tiempo.usuario, tiempo.materia, tiempo_maximo,
             tiempo_maximo);
```

```
    ELSE
       INSERT INTO  mdl_time_summarization
       VALUES (tiempo.usuario, tiempo.materia, tiempo_maximo, 0);
    END IF;
  END IF;
END agregar_tiempo_materia;

BEGIN
  #Se itera por cada uno de los usuarios
  FOR id_alumno IN (SELECT distinct(usuario) FROM mdl_tesis_tiempos)
  LOOP
     FOR tiempo IN tiempos(id_alumno.usuario)
     LOOP
       IF (tiempo.TIEMPO_DE_LA_ACCION IS NOT NULL
           AND tiempo.action != 'logout')
       THEN
         agregar_tiempo_materia(tiempo);
       END IF;
     END LOOP;
  END LOOP;
END;
/
/*Agrego las columnas para discretizar los atributos total_tiempo_materia y
total_tiempo_foro.*/
ALTER TABLE `mdl_tesis_time_summarization`
ADD COLUMN `tiempo_materia` VARCHAR(45) DEFAULT NULL,
ADD COLUMN `tiempo_foro` VARCHAR(45) DEFAULT NULL;
#Participación baja en las materias.
UPDATE mdl_tesis_time_summarization
SET tiempo_materia = 'Bajo' WHERE total_tiempo_materia <= 488;
#Participación media en las materias.
UPDATE mdl_tesis_time_summarization
SET tiempo_materia = 'Medio'
WHERE total_tiempo_materia > 488 AND total_tiempo_materia <= 2493;
#Participación alta en las materias.
UPDATE mdl_tesis_time_summarization
SET tiempo_materia = 'Alto' WHERE total_tiempo_materia > 2493;
#Participación baja en los foros.
UPDATE mdl_tesis_time_summarization
SET tiempo_foro = 'Bajo' WHERE total_tiempo_foro <= 47;
#Participación media en los foros.
UPDATE mdl_tesis_time_summarization
SET tiempo_foro = 'Medio'
WHERE total_tiempo_foro > 47 AND total_tiempo_foro <= 444;
#Participación alta en los foros.
UPDATE mdl_tesis_time_summarization
SET tiempo_foro = 'Alto' WHERE total_tiempo_foro > 444;
```

14 Uso de biblioteca

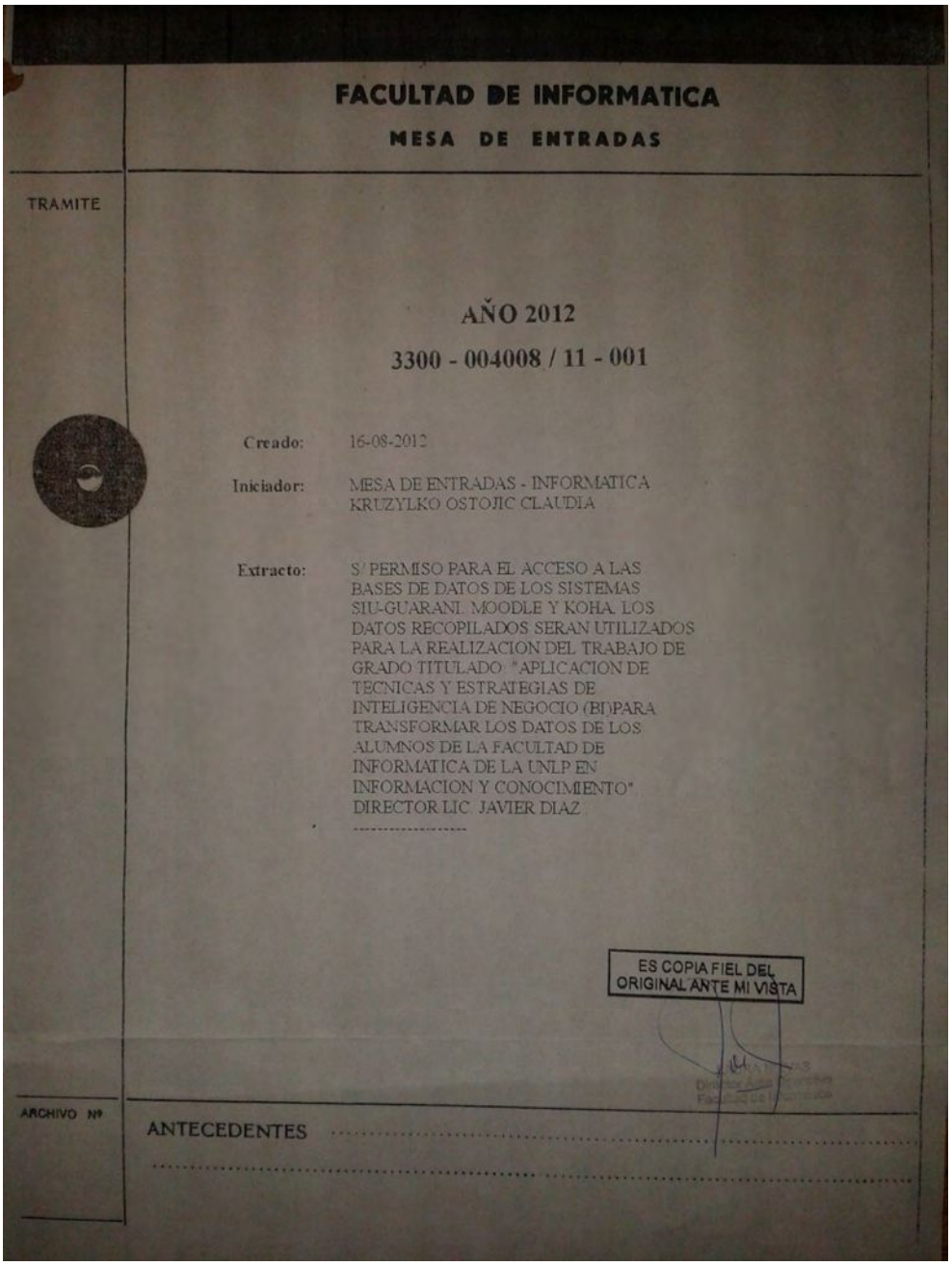
```
/*No tengo en cuenta las materias que no están en los estantes virtuales**/
DELETE FROM mdl_tesis_course_students_note
WHERE shortname NOT IN
      (SELECT cod_materia FROM tesis_estante_virtual_materia)
/*Estudiantes que hicieron  uso de la biblioteca el año en que curso la materia **/
CREATE TABLE tesis_uso_bibliotecas
SELECT usr.id as userid, cn.courseid, 1 as uso_biblioteca
FROM mdl_user usr, mdl_user_info_data usr_d, meran.tesis_datos td,
     meran.tesis_estante_virtual_materia ev,
     mdl_tesis_course_students_note cn
WHERE usr.id = usr_d.userid
AND usr_d.fieldid = 2 #El field 2 contiene el dato dni
AND td.dni = usr_d.data
AND td.cod_estante_virtual = ev.id_estante_virtual
AND cn.shortname = ev.cod_materia AND cn.año_academico = td.año_academico
AND cn.userid = usr.id
GROUP BY usr.id, cn.courseid;
/*Estudiantes que NO  hicieron  uso de la biblioteca el año en que curso la materia **/
INSERT INTO tesis_uso_bibliotecas
SELECT userid, courseid, 0
FROM mdl_tesis_course_students_note2 cn
WHERE NOT EXISTS (SELECT 1 FROM tesis_uso_bibliotecas b
                  WHERE b.userid = cn.userid
                      AND b.courseid = cn.courseid)
GROUP BY userid, courseid;
```

15 Vista minable

```
/**Creación de la vista minable que contiene los datos para el análisis de la participación
social de los alumnos en las materias.*/
CREATE TABLE mdl_tesis_vista_minable
SELECT ptf.participacion_foro, rpc.rta_compañeros,
       rpd.rtas_docentes, cci.conversaciones_iniciadas, sum.tiempo_foro,
       sum.tiempo_materia, b.uso_biblioteca, csn.m_cursada
FROM mdl_tesis_course_students_note csn,
     mdl_tesis_participacion_total_foros ptf,
     mdl_tesis_respuesta_post_compañeros rpc,
     mdl_tesis_respuesta_post_docentes rpd,
     mdl_tesis_conversaciones_iniciadas cci,
     mdl_tesis_time_summarization sum, tesis_uso_bibliotecas b
WHERE csn.userid = ptf.userid AND csn.courseid = ptf.courseid
  AND csn.userid = rpc.userid AND csn.courseid = rpc.courseid
  AND csn.userid = rpd.userid AND csn.courseid = rpd.courseid
  AND csn.userid = cci.userid AND csn.courseid = cci.courseid
  AND csn.userid = sum.userid AND csn.courseid = sum.courseid
  AND csn.userid = b.userid AND csn.courseid = b.courseid;
```

Apéndice B
RESOLUCIONES

En este apéndice se adjuntan las notas presentadas y las copias del expediente en donde se explicita el permiso para extraer los datos de las bases de datos de los sistemas de la Facultad de Informática – UNLP, con el objetivo de la realización de este trabajo de grado.



FACULTAD DE INFORMATICA

MESA DE ENTRADAS

TRAMITE

AÑO 2012

3300 - 004008 / 11 - 001

Creado: 16-08-2012

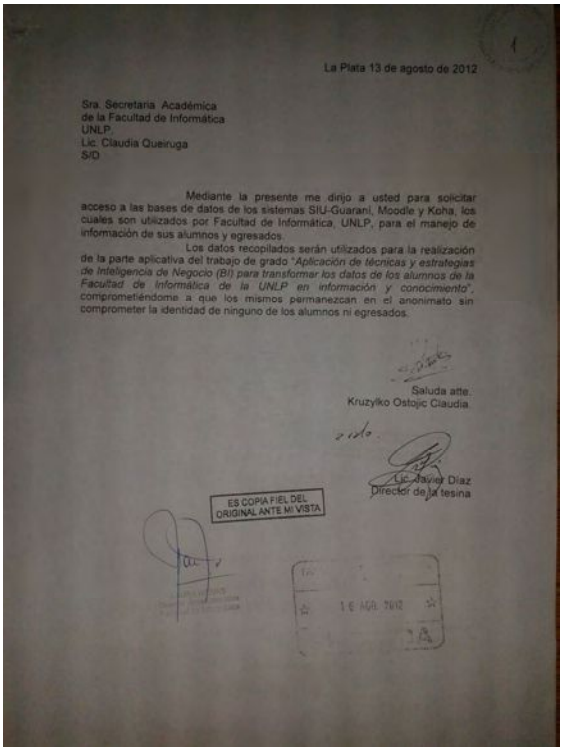
Iniciador: MESA DE ENTRADAS - INFORMATICA
KRUZYLKO OSTOJIC CLAUDIA

Extracto: S/ PERMISO PARA EL ACCESO A LAS BASES DE DATOS DE LOS SISTEMAS SIU-GUARANI, MOODLE Y KOHA. LOS DATOS RECOPILADOS SERAN UTILIZADOS PARA LA REALIZACION DEL TRABAJO DE GRADO TITULADO "APLICACION DE TECNICAS Y ESTRATEGIAS DE INTELIGENCIA DE NEGOCIO (BI)PARA TRANSFORMAR LOS DATOS DE LOS ALUMNOS DE LA FACULTAD DE INFORMATICA DE LA UNLP EN INFORMACION Y CONOCIMIENTO" DIRECTOR LIC. JAVIER DIAZ

ES COPIA FIEL DEL ORIGINAL ANTE MI VISTA

ARCHIVO Nº

ANTECEDENTES



La Plata 13 de agosto de 2012

Sra. Secretaria Académica
de la Facultad de Informática
UNLP
Lic. Claudia Queiruga
S/D

Mediante la presente me dirijo a usted para solicitar acceso a las bases de datos de los sistemas SIU-Guaraní, Moodle y Koha, los cuales son utilizados por Facultad de Informática, UNLP, para el manejo de información de sus alumnos y egresados.

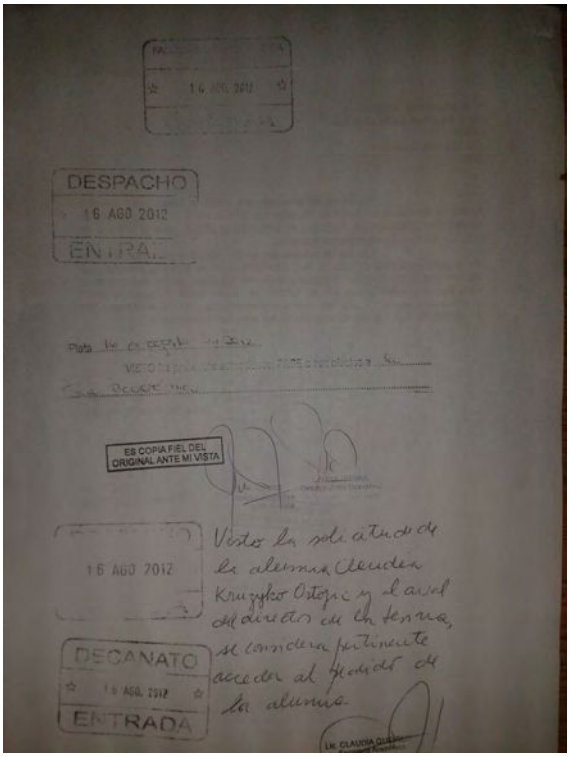
Los datos recopilados serán utilizados para la realización de la parte aplicativa del trabajo de grado "*Aplicación de técnicas y estrategias de Inteligencia de Negocio (BI) para transformar los datos de los alumnos de la Facultad de Informática de la UNLP en información y conocimiento*", comprometiéndome a que los mismos permanezcan en el anonimato sin comprometer la identidad de ninguno de los alumnos ni egresados.

Saluda atte.
Kruzylko Ostojic Claudia

Lic. Javier Diaz
Director de la tesina

ES COPIA FIEL DEL ORIGINAL ANTE MI VISTA

16 AGO 2012



16 AGO 2012

DESPACHO
16 AGO 2012
ENTRADA

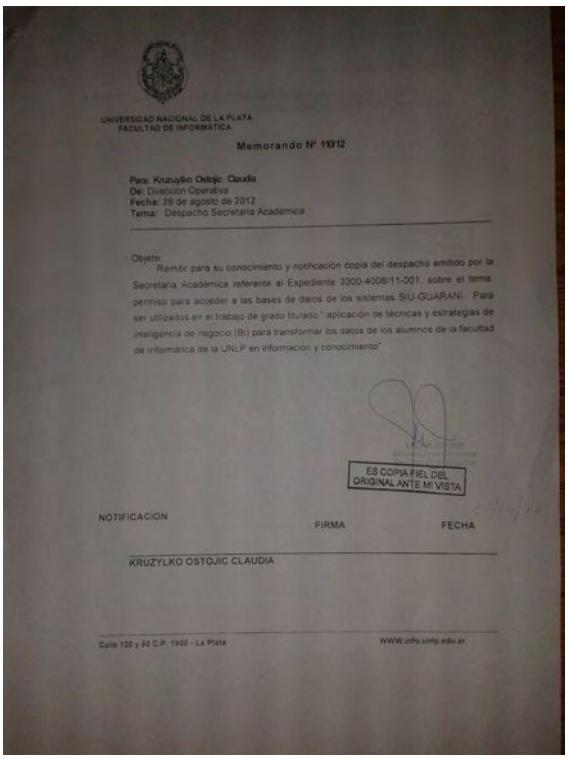
ES COPIA FIEL DEL ORIGINAL ANTE MI VISTA

16 AGO 2012

DECANATO
16 AGO 2012
ENTRADA

Vistas las solicitudes de la alumna Claudia Kruzylko Ostojic y el aval del director de la tesina, se considera pertinente acceder al pedido de la alumna.

Lic. CLAUDIA QUEIRUGA



UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE INFORMATICA

Memorando N° 1012

Para: Kruzylko Ostojic Claudia
De: Dirección Operativa
Fecha: 29 de agosto de 2012
Tema: Despacho Secretaría Académica

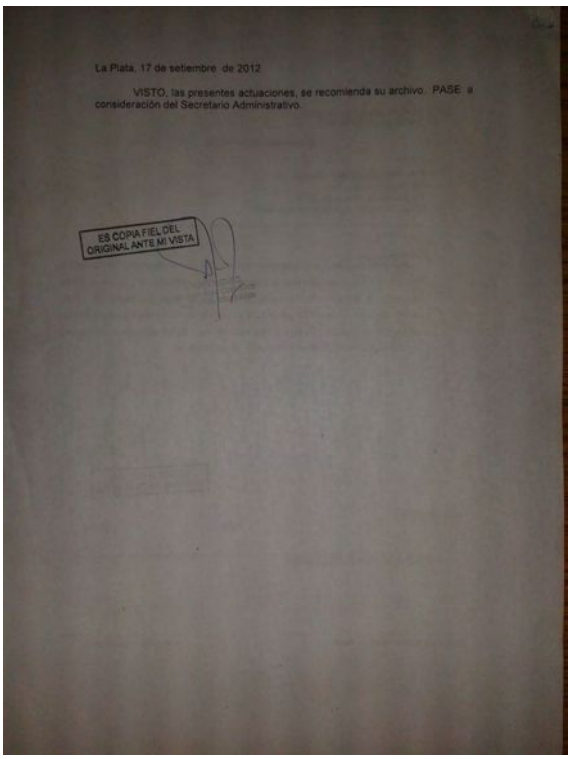
Objeto:
Remitir para su conocimiento y notificación copia del despacho emitido por la Secretaria Académica referente al Expediente 3300-4008/11-001, sobre el tema: permiso para acceder a las bases de datos de los sistemas SIU-GUARANI. Para ser utilizados en el trabajo de grado titulado " aplicación de técnicas y estrategias de inteligencia de negocio (BI) para transformar los datos de los alumnos de la facultad de informática de la UNLP en información y conocimiento"

ES COPIA FIEL DEL ORIGINAL ANTE MI VISTA

NOTIFICACION FIRMA FECHA

KRUZYLKO OSTOJIC CLAUDIA

Calle 120 y 60 C.P. 1900 - La Plata WWW.info.unlp.edu.ar



La Plata, 17 de setiembre de 2012

VISTO, las presentes actuaciones, se recomienda su archivo. PASE a consideración del Secretario Administrativo.

ES COPIA FIEL DEL ORIGINAL ANTE MI VISTA

REFERENCIAS BILIOGRÁFICAS

[Agrawal, Imielinski & Swami 1993] Agrawal Rakesh, Imielinski Tomasz and Swami Arun. "Mining association rules between sets of items in large databases". ACM SIGMOD Conference on Management of Data, pages 207-216, 1993.

[Agrawall & Srikant 1995] Agrawall Rakesh, Srikant Ramakrishnan. "Mining Sequential Patters" in *Proceedings of the Eleventh International Conference of Data Engineering, ICDE*, Taipei, Taiwan, pp. 3-14, 1995.

[Agrawall & Srikant 1996] Agrawall Rakesh, Srikant Ramakrishnan. "Mining Sequential Patters: Generalization and Performance Improvements" in *Proceedings of the Eleventh International Conference on Extending Database Technology*, Lecture Notes in Computer Science, Springer Verlag, Vol. 1057, pp. 3-17, 1996.

[AllAnalitics] http://www.allanalytics.com. Fecha de consulta: Agosto de 2013.

[Antunes 2010] Antunes Claudia. "Anticipating Students' Failure As Soon As Possible". En *Romero C., Ventura S., Pechenizkiy M. and Baker R. S. J. d. (Eds.) Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press. Capítulo 25. pp. 353-364. 2010.

[Arthur & Vassilvitskii 2007] Arthur, D. and Vassilvitskii, S. "*k*-means++: the advantages of careful seeding" in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035, 2007.

[Baracosa & Antunes 2011] Barracosa Joana., Antunes Claudia. "Anticipating Teachers' Performance". In *KDD 2011 Workshop: Knowledge Discovery in Educational Data*, 17th ACM SIGKDD Conference on Knowlodge Discovery and Data Mining. San Diego, CA, 21 – 24 de Agosto, 2011. https://pslcdatashop.web.cmu.edu/KDD2011/

[Barbenabeu, 2010] Bernabeu Ricardo Darío. *HEFESTO V2.0*. Córdoba, Argentina – Lunes 19 de Julio de 2010.

[Bhardwaj & Pal 2011] Bharadwaj B.K., Pal S. “Data Mining: A prediction for performance improvement using classification”. In *IJCSIS - International Journal of Computer Science and Information Security*. Vol. 9, No. 4, pp. 136-140, 2011.

[Bi-argentina] Inteligencia de Negocios para Empresas. Business Intelligence Open Source en Argentina. ¿Qué es Business Intelligence? Fecha de consulta: Octubre de 2012. Disponible en: http://www.bi-argentina.com.ar/que-es-business-intelligence/.

[Brachman & Anand, 1994] Brachman Ronald J., Anand Tej. “The Process of Knowledge Discovery in Databases: A First Sketch”. In KDD Workshop, Seattle, Washington, USA, pp. 1-12. 1994.

[Breiman et. al. 1984] Breiman, L., Friedman J. H., Olshen R. A., Stone C. J. “Classification and regression trees”. Monterey, Calif., U.S.A.: Wadsworth, Inc. 1984.

[Cabena et al., 1998] Cabena Peter, Hadjinian Pablo, Stadler Rolf, Verhees Jaap, Zanasi Alessandro. “Discovering Data Mining: From Concept to Implementation”. Prentice Hall, Upper Saddle River, NJ, 1998.

[Calders & Pechenizkiy 2011] Calders Toon, Pechenizkiy Mykola. “Introduction to The Special Section on Educational Data Mining”. Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), Diciembre de 2011. Volume 13, Issue 2.

[Chen et al. 2008] Chen Ruey-Shun, Tsai Yung-Shun, Yeh K.C., Yu D.H., Bak-Sau Yip. “Using Data Mining to Provide Recommendation Service”. In *World Scientific and Engineering Academy and Society (WSEAS)*. Jornada TRANSACTIONS on INFORMATION SCIENCE & APPLICATIONS. Issue 4, Volume 5, April 2008.

[DBMiner Technology Inc.] DBMiner Technology Inc. Fecha de consulta Octubre de 2012. Disponible en http://www.dbminer.com/.

[Dekker, Pechenizkiy & Vleeshouwers 2009] Dekker Gerben W., Pechenizkiy Mykola, Vleeshouwers Jan M. “Predicting Students Drop Out: A Case Study”. In *Proceedings of the Second International Conference on Educational Data Mining*. Córdoba, España, Julio 1–3, 2009, pp. 38–47.

[Dimic, et al. 2010] Dimic Gabrijela, Prokin Gragana, Kuk Kristijan, Spalevic Petar. “The Use of Data Mining Methods for Analyzing and Evaluating Course Quality in the Moodle System”. In *UNITECH – International Scientific Conference, Gabrovo, Bulgaria*. November 19 – 20 2010.

[Duda & Hart 1973] Duda R. O., Hart P. E. “Pattern Classification and scene analysis”. John Wiley and Sons. 1973.

[Eclipse] http://www.eclipse.org/.

[Falakmasir & Habibi 2010] Falakmasir Mohammad Hassan, Habibi Jafar. “Using Educational Data Mining Methods to Study the Impact of Virtual Classroom in E-Learning”. In *Proceedings of the third International Conference on Educational Data Mining*. Pittsburgh, PA, USA, Junio 11–13, 2010, pp. 241-248.

[Fayyad et al. 1996] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., “The KDD process for extracting useful knowledge from volumes of data”. Communications of The ACM 39, November 1996., pp. 27-34.

[Fisher 1987] Fisher, D. “Knowledge acquisition via incremental conceptual clustering” *Machine Learning*, 2, pp 139-172, 1987.

[Frawley et.al, 1992] Frawley William J., Piatetsky-Shapiro Gregory, Matheus Christopher J. “Knowledge Discovery in Databases: An Overview”.

[Friedman 1997] Friedman, J.: “Data mining and statistics: what is the connection?” Keynote Address, 29th Symposium on the Interface: Computing Science and Statistics, 1997.

[Gartner Group IT Glossary] Gartner IT Glossary; Definning the IT Industry. Fecha de consulta: octubre de 2012. Disponible en: http://www.gartner.com/it-glossary/.

[Gennari et al. 1990] Gennari, J. H., Langley, P., Fisher, D. “Models of incremental concept formation”. Artificial Intelligence, 40, 11–61. 1990.

[Godoy & Amandi 2010] Godoy Daniela y Amandi Analía. “Link Recommendation in E-Learning Systems Based on Content-Based Student Profiles”. In *C. Romero, S. Ventura, M. Pechenizkiy and R. S. J. d. Baker (Eds.) Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press, capítulo 19. pp. 273-286. 2010.

[Han, et al 2012] Han Jiawei, Kamber Micheline, Pei Jian. “Data Mining Concepts and Techniques. Third Edition”. Morgan Kaufmann Publishers is an imprint of Elsevier. 225 Wyman Street, Waltham, MA 02451, USA, 2012.

[Hand et. al. 2001] Hand David, Mannila Heikki, Smyth Padhraic. “Principles of Data Mining”. A Bradford Book The MIT Press, Institute of Technology, Cambridge, Massachusetts London England, ISBN 0-262-08290-X, 2001. ISBN: 026208290x.

[Hand 1998] Hand David. J., “Data mining: statistics and more?” The American Statistician, 52, pp. 112-118. 1998.

[Howson, 2007] Howson Cindi. “Successful Business Intelligence: Secrets to Making BI a Killer App”. Ed. McGraw-Hill Osborne Media. 2007. ISBN-P: 978-0-07-149851-7
ISBN-W: 0-07-149851-6

[IBM - SPSS] IBM – SPSS software - Predictive analytics software and solutions. Fecha de consulta Octubre de 2012. Disponible en http://www-01.ibm.com/software/analytics/spss/.

[IEDMS 2012] International Educational Data Mining Society. Fecha de consulta: segundo semestre de 2012. Disponible en http://www.educationaldatamining.org/.

[Kopanakis et.al, 2003] Kopanakis I. and Theodoulidis B.. “Visual data mining modeling techniques for the visualization of mining outcomes”. Journal of Visual Languages and Computing, 14(6):543–589, 2003.

[Kovacic 2010] Kovacic Zlatko J. “Early Prediction of Student Success: Mining Students Enrolment Data”. In *Proceedings of the Informing Science and Information Technology Education Conference*. Southern, Italia, Junio 19-24, 2010.

[Langley et. al. 1992] Langley P., Iba W., Thompson K. “An análisis of bayesian classifiers”. In *Proceeding of the 10th National Conference on Artificial Intelligence*. 223-223. 1992.

[Laurinen 2006] Laurinen Perttu. “A TOP-DOWN APPROACH FOR CREATING AND IMPLEMENTING DATA MINING SOLUTIONS”. Academic Dissertation to be presented with the assent of the Faculty of Technology, University of Oulu, for public discussion in the Auditorium TS101, Linnanmaa, on June 22nd, 2006.

[López, et al. 2012] López Manuel Ignacio, Romero Cristóbal, Ventura Sebastián y Luna J.M. “Classification via clustering for predicting final marks starting from the student participation in Forums”. In *Proceedings of the Fifth International Conference on Educational Data Mining*. Panorama Hotel, Chania, Greece, Junio 19 – 21, 2012, pp 148-151.

[MacQueen 1967] MacQueen James. "Some Methods for classification and Analysis of Multivariate Observations", in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*", Berkeley, University of California, Press 1, 281-297, 1967.

[Maimon & Last, 2000] Maimon O., Last M. "Knowledge Discovery and Data Mining: The Info-Fuzzy network (IFN) methodology". Kluwer Academic Publishers, 2000.

[Maimon & Rokach 2010] Maimon Oded, Rokach Lior. "Data Mining and Knowledge Discovery Handbook". 2nd ed. Springer 2010, ISBN 978-0-387-09822-7.

[Márquez-Vera et al. 2011] Márquez-Vera Carlos, Romero Cristóbal, Ventura Sebastián. "Predicting School Failure Using Data Mining". *In Proceedings of the 4th International Conference on Educational Data Mining*. Eindhoven, the Netherlands Julio 6 – 8, 2011, pp 271-276.

[Nagata et al. 2009] Nagata Ryo, Takeda Keigo, Suda Koji, Kakegawa Junichi, Morihiro Koichiro. "Edu-mining for Book Recommendation for Pupils". In *Proceedings of the Second International Conference on Educational Data Mining*. Córdoba, España, Julio 1–3, 2009, pp. 91–100.

[Obsivac et. al. 2012] Obsivac Tomas, Popelinsky Lubos, Bayer Jaroslav, Geryk Jan y Bydzovska Hana. "Predicting drop-out from social behavior of students". In *Proceedings of the Fifth International Conference on Educational Data Mining*. Panorama Hotel, Chania, Greece, Junio 19 – 21, 2012, pp 103-109.

[online-behavior] http://online-behavior.com/analytics/definition. Fecha de consulta: Agosto de 2013.

[Pazzani et.al, 1997] Pazzani Michael J., Mani Subramani, Shankle Rodman W., "Comprehensible Knowledge-Discovery in Databases". In Cognitive Science Conference, University of California, Berkeley. 1997.

[Pechenizkiy et. al 2012] Pechenizkiy Mykola, Trcka Nikola, De Bra Paul, Toledo Pedro. "CurriM: Curriculum Mining". In *Proceedings of the Fifth International Conference on Educational Data Mining*. Panorama Hotel, Chania, Greece, Junio 19 – 21, 2012, pp 216-217. Open Job of the Information Systems group of the Computer Science Department, TU/e Technische Universiteit Eindhoven University of Technology. http://www.win.tue.nl/is/doku.php?id=jobs:start

[Purple Insight - MineSet] Purple Insight, Gloucester – MineSet. Fecha de consulta Octubre de 2012. Disponible en http://www.algorithmic-solutions.com/leda/projects/mineset.htm.

[prudsys XELOPES] prudys | the Realtime Analytics Company – XELOPES. Fecha de consulta Octubre de 2012. Disponible en http://www.prudsys.de/en/technology/xelopes/.

[Quinlan 1993] Quinlan, J.R. "C4.5: Programs for Machine Learning". Morgan Kaufmann Publishers, San Mateo, California, EE.UU. 1993.

[Quinlan 1986] Quinlan J.R.. "Induction of Decision Trees". In *Machine Learning*. Capítulo 1, p.81-106. Morgan Kaufmann, 1986.

[RapidMiner] Rapid-I, Report the feature. Disponible en http://rapid-i.com/content/view/181/190/. Fecha de consulta Junio 2013.

[Reinartz, 2002] Reinartz, T.: *"*A Unifying View on Instance Selection*",* In *Data Mining and Knowledge Discovery*, Vol. 6, No. 2, pp. 191–210, 2002.

[Romero, et al. 2008] Romero Cristóbal, Ventura Sebastián, Espejo Pedro G., Hervás César. "Data Mining Algorithms to Classify Students". In *Proceedings of the First International Conference on Educational Data Mining*. Montreal, Québec, Canadá. Junio 20 – 21, 2008, pp. 8-17.

[Romero & Ventura 2007] Romero Cristóbal, Ventura Sebastián. "Educational data mining: A survey from 1995 to 2005". Expert Systems with Applications. Volume 33, Issue 1, July 2007, Pages 135-146.

[RuleQuest Research See5 / C5.0] RULEQUEST RESEARCH data mining tools … helping you transform data into knowledge. Data Mining Tools See5 and C5.0. Fecha de consulta: Octubre de 2012. Disponible en: http://www.rulequest.com/see5-info.html.

[Salford Systems - CART] Salford Systems – CART, Classification and Regression Trees. Fecha de consulta: Octubre de 2012. Disponible en: http://www.salford-systems.com/en/products/cart.

[SAS - Enterprise Miner] SAS - Enterprise Miner. Fecha de consulta: Octubre de 2012. Disponible en http://www.sas.com/technologies/analytics/datamining/miner/.

[Schepes, 2008] Schepes Swain. “Business Intelligence for Dummies”. Ed. Wiley Publishing, Inc. 2008.

[SDM] Search Data Management. Data Analytic (DA). Fecha de consulta. Julio 2013. Disponible en: http://searchdatamanagement.techtarget.com/definition/data-analytics

[Sgi – MLC++] sgi – MLC ++. Fecha de consulta Octubre de 2012. Disponible en http://www.sgi.com/tech/mlc/.

[Tang & McCalla 2010] Tang Tiffany Y., McCalla Gordon I. “Data mining for contextual educational recommendation and evaluation strategies”. En *Romero C., Ventura S., Pechenizkiy M. and Baker R. S. J. d. (Eds.) Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press. Capítulo 18. pp. 257-272. 2010.

[Technopedia] http://www.techopedia.com/definition/26418/data-analytics. Fecha de consulta: Agosto de 2013.

[Thai-Nghe, Janecek & Haddawy 2007] Thai-Nghe Nguyen, Janecek Paul, Haddawy Peter. “A comparative analysis of techniques for predicting academic performance”. In *Proceedings of 37th ASEE/IEEE Frontiers in Education (FIE 2007), IEEE Xplore*, pp. 7-12.

[Timarán Pereira 2009] Timaran Pereira Silvio Ricardo. “La minería de datos en el descubrimiento de perfiles de deserción estudiantil en la Universidad de Nariño”. En *Revista Universidad Y Salud, Nariño, Colombia*. *ISSN:* 0124-7107 *ed.:* Editorial Universidad De Nariño. Año 9 *fasc.*11 *v.*1 pp. 151 - 158 ,2009.

[Trcka et al. 2011] Trcka Nikola, Pechenizkiy Mykola, Aalst Wil van der. “Process Mining from Educational Data”. In *C. Romero, S. Ventura, M. Pechenizkiy and R. S. J. d. Baker (Eds.) Handbook of Educational Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, CRC Press, 2011. Capítulo 9. pp. 123-142.

[Vialardi, Bravo, et al. 2009] Vialardi Sacin Cesar, Bravo Agapito Javier, Shafti Leila, Ortigosa Alvaro. “Recommendation in Higher Education Using Data Mining Techniques”. In *Proceedings of the Second International Conference on Educational Data Mining*. Córdoba, España, Julio 1–3, 2009, pp. 190–199.

[Ward System Group, Inc] Ward System Group, Inc. Artificial Intelligence Software for Science and Business. Fecha de consulta: Octubre de 2012. Disponible en: http://www.wardsystems.com/.

[WEKA – The University of Waikato] WEKA – The University of Waikato, Machine Learning Group at University of Waikato. Fecha de consulta Octubre de 2012. Disponible en http://www.cs.waikato.ac.nz/ml/weka/.

[White, 2009] White, Tom. "Hadoop: The Definitive Guide". s.l. : O'Reilly, 2009. ISBN: 978-0-596-52197-4.

[Witten et. al 2011] Witten Ian H., Frank Eibe, Hall Mark A. "Data Mining Practical Machine Learning Tools and Techniques". 3rd Edition. Morgan Kaufmann Publishers - Elsevier Inc. 2011.

[Xu & Recker 2011] Xu Beijie, Recker Mimi. "Understanding Teacher Users of a Digital Library Service: A clustering approach". In *JEDM - Journal of Educational Data Mining* (ISSN 2157-2100) Volume 3, Issue 1, (in press 2011).

[Yadav & Pal 2012] Yadav Surjeet Kumar, Pal Saurabh. "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification". In *WCSIT - World of Computer Science and Information Technology Journal (ISSN: 2221-0741).* Vol. 2, No. 2, pp. 51-56, 2012.