UNIVERSIDAD NACIONAL "DANIEL ALCIDES CARRIÓN"

ESCUELA DE POSGRADO

MAESTRIA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN

"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN"

TESIS

PARA OPTAR EL GRADO ACADÉMICO DE MAESTRO

PRESENTADO POR:

Ing. Pit Frank ALANIA RICALDI

ASESOR

Mg. Zenon Manuel LOPEZ ROBLES

PASCO - PERÚ
2018

UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN"

TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAESTRO

PRESENTADO POR:

Ing. Pit Frank ALANIA RICALDI

SUSTENTADO Y APROBADO ANTE LOS JURADOS

_____

Msc. Hebert Carlos Castillos Paredes
Presidente

_____ _____

Mg. Percy Ramirez Medrano
Miembro

Mg. Oscar C. Campos Salvatierra
Miembro

PASCO - PERÚ
2018

DEDICATORIA

A mis padres, esposa e hijos, que siempre me dan sus apoyos. Gracias por la alegría que siento al estar con ustedes.

RECONOCIMIENTO

A mis colegas de la Escuela Profesional de Ingeniería de Sistemas y Computación de la UNDAC porque siempre me incentivaron para la elaboración de esta tesis.

A los docentes de la Escuela de Post Grado, que me inculcaron sus conocimientos, la innovación y el correcto camino académico.

A mi asesor y a los jurados, que con su amabilidad y buen juicio crítico me impulsaron en la consecución de este trabajo de investigación.

A mi esposa e hijos, que han sabido disculpar mi dedicación al trabajo y a esta tesis.

RESUMEN

La Minería de datos es usado para estudiar los datos disponibles en el cualquier campo y descubrir el conocimiento oculto en ella. Los métodos de clasificación como árboles de decisión, las reglas, red bayesiana, etc. se pueden aplicar a los datos académicos de una universidad para predecir el comportamiento de los estudiantes, el rendimiento en los exámenes, deserción estudiantil, etc. Esta predicción ayudará a las autoridades para identificar la deserción estudiantil y poder determinar la proyección de secciones y otras acciones. El algoritmo de árbol de decisión C4.5 (J48) se aplica en los datos de las notas finales semestrales de los estudiantes para predecir si abandona o no los estudios. El resultado del árbol de decisión predijo el número de estudiantes que son propensos a abandonar la carrera profesional. El resultado lo pueden utilizar a las autoridades para que puedan tomar las medidas para mejorar la toma de decisiones. Después de la evaluación con los datos originales se introducen un conjunto de datos de prueba en el sistema para analizar los resultados. El análisis comparativo de los resultados indica que la predicción ha ayudado determinar con mayor precisión el mejoramiento en el resultado. Para analizar la exactitud del algoritmo, se compara con el algoritmo Random Tree y se encontró que es tan eficiente en términos de precisión de los resultados académicos del estudiante y el tiempo tomado para crear el árbol.

Palabras clave: Deserción estudiantil, Predicción, Minería de Datos, CRISP DM, Redes Neuronales.

ABSTRACT

Data Mining is used to study the data available in any field and discover the knowledge hidden in it. Classification methods such as decision trees, rules, Bayesian network, etc. can be applied to the academic data of a university to predict the behavior of the students, the performance in the exams, student desertion, etc. This prediction will help the authorities to identify student desertion and determine the projection of sections and other actions. The decision tree algorithm C4.5 (J48) is applied in the students' semester final grade data to predict whether or not the studies are abandoned. The result of the decision tree predicted the number of students who are likely to drop out of the professional career. The result can be used to the authorities so that they can take measures to improve decision making. After the evaluation with the original data, a set of test data is introduced into the system to analyze the results. The comparative analysis of the results indicates that the prediction has helped to determine with greater precision the improvement in the result. To analyze the accuracy of the algorithm, it is compared to the Random Tree algorithm and found to be so efficient in terms of accuracy of the student's academic results and the time taken to create the tree.

Keywords: Student desertion, Prediction, Data Mining, CRISP DM, Neural Networks.

ÍNDICE

INTRODUCCIÓN

La deserción estudiantil se ha convertido en un problema social que afecta a muchas universidades en todo el Perú, reducir el número de estudiantes desertores es un tema álgido hoy en día. Que tienen muy presente cada uno de las universidades, donde las mismas planean implementar un plan estratégico para reducir el índice de estudiantes que deciden abandonar sus estudios.

Para contribuir con la solución del problema de la deserción estudiantil en la Universidad Nacional Daniel Alcides Carrión se plantea realizar un estudio comparativo de técnicas de minería de datos para predecir la deserción estudiantil universitaria en la región de Pasco.

El análisis predictivo es el proceso de tratar con variedad de datos y aplicar diversas fórmulas matemáticas para descubrir la mejor decisión para una situación dada. El análisis predictivo da a la universidad una ventaja competitiva. Para las universidades, cuyo objetivo es contribuir a la mejora de la calidad de la educación superior, el éxito de la creación de capital humano es el sujeto de un análisis continuo. Por lo tanto, la predicción de éxito de los estudiantes es fundamental para las instituciones de educación superior, porque la calidad del proceso de enseñanza es la capacidad de satisfacer las necesidades de los estudiantes. El análisis predictivo abarca una variedad de técnicas de minería de datos, las estadísticas, que

analizan los hechos actuales e históricos para hacer predicciones sobre eventos futuros. El término predictivo de minería de datos se aplica generalmente para identificar los proyectos de minería de datos con el objetivo de identificar un modelo de red neuronal o estadística o un conjunto de modelos que se puede utilizar para predecir algunas respuestas de interés. Por ejemplo, una universidad que desea, puede utilizar los datos con predicción minera, para obtener un modelo que pueda identificar rápidamente el índice de deserción universitaria.

PRIMERA PARTE: ASPECTOS TEÓRICOS

CAPITULO I

PROBLEMA DE INVESTIGACIÓN

1.1 IDENTIFICACIÓN Y DETERMINACIÓN DEL PROBLEMA

La deserción estudiantil universitaria es uno de los problemas que aborda la mayoría de las instituciones de educación superior, que afecta tanto a los estudiantes como a la institución que los acogen. Esto adquiere una importancia relevante según las cifras sobre deserción estudiantil expuestas por la Organización de Cooperación Económica y el Desarrollo (OECD) que agrupa a los principales

países desarrollados y que señala que las cifras de deserción en estos países vienen en franco crecimiento.

En el Perú, alcanza una tasa de deserción del 17% del problema de educación, según el portal del postulante Logros (Logros, 2011), donde se señala que cada año en el Perú se pierden más de 100 millones de dólares por el abandono de las aulas y además indica que en la próxima década se podrían perder más de 2 mil 100 millones de dólares si no se toman las medidas adecuadas.

Por otro lado, las universidades que imparten asignaturas de diferentes especialidades no pueden predecir de manera precisa la cantidad de alumnos que van a abandonar los estudios en un futuro cercano, lo cual impide determinar, entre otros aspectos, la apertura de grupos o secciones, así como, la correspondiente asignación docentes y de aulas a los grupos respectivos; esta deserción de alumnos es indeterminado por diferentes factores e incluye diversas causas.

CRISP-DM fue la metodología utilizada para la creación del modelo, la misma que es una de las más usadas en la actualidad para la generación de proyectos de Minería de datos, con ella se pretende obtener un modelo de análisis de datos, que con la ayuda de la implementación de algoritmos de Inteligencia Artificial, ya

incorporados en programas, se pueda predecir la probable deserción en las Instituciones educativas de educación superior y así tomar las medidas preventivas.

1.2 DELIMITACIÓN DE LA INVESTIGACIÓN

La delimitación del presente estudio estuvo enfocada en los siguientes aspectos:

1.2.1 Delimitación del espacio:

El estudio se realizó en la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión, ubicada en la Ciudad de Cerro de Pasco, distrito de Yanacancha, provincia de Pasco.

1.2.2 Delimitación del tiempo:

La investigación se realizó con los datos de los años 2013 al 2017.

1.2.3 Delimitación de población

El grupo objeto de estudio estuvo conformado por estudiantes de la Escuela de Formación Profesional de Sistemas y

Computación de la Universidad Nacional Daniel Alcides Carrión.

1.3 FORMULACIÓN DEL PROBLEMA

1.3.1 Problema General

¿Cómo puede predecirse la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión?

1.3.2 Problema Específico

¿Qué técnicas de minería de datos puede predecir la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión?

1.4 FORMULACION DE OBJETIVOS

1.4.1 Objetivo General

Predecir la deserción estudiantil mediante técnicas de minería de datos de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.

1.4.2 Objetivo Específico

Aplicar técnicas de minería de datos para predecir la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.

1.5 JUSTIFICACION DE LA INVESTIGACIÓN

En estos tiempos donde la tecnología ocupa un rol importante dentro de las actividades humanas, las técnicas de predicción de minería de datos dan solución a problemas de clasificación, agrupamientos, tendencias, etc. no solo en el ámbito empresarial sino también en el ámbito académico logrando realizar este proceso de manera razonable, no costosa, facilitando el análisis de los planes de trabajo o las políticas de operación. El trabajo propuesto es necesario, por la gran cantidad de datos que se manejan en el ámbito académico, es conocido que el aspecto educativo informático tiene uno de los más altos porcentajes de interés de las personas, por lo que es importante conocer sus tendencias en cuanto a la deserción estudiantil para las instituciones educativas superiores, para ello se cuenta con la existencia de diferentes tecnologías, técnicas y algoritmos que pueden ayudar a ser más eficiente las operaciones en este aspecto académico.

1.6 LIMITACIONES DE LA INVESTIGACIÓN

La investigación se realizó solo con datos de los alumnos de los años 2013 al 2017 de la Escuela de Profesional de Ingeniería de Sistemas y Computación de la Universidad Nacional Daniel Alcides Carrión.

CAPITULO II

MARCO TEORICO

2.1 ANTECEDENTES DE ESTUDIO

Para el desarrollo del proyecto como antecedentes de estudio se tienen trabajos de investigación relacionadas a la minería de datos y la deserción estudiantil la cuales son las siguientes:

2.1.1 A nivel nacional

a. En la Tesis intitulada **"Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la educación básica regular en la región de Lambayeque"**

de (Piscoya Ordonez, 2016), se arriba a las siguientes conclusiones:

Se realizó la selección de las técnicas predictivas de minería de datos, determinando que el modelo a utilizar sería uno de series de tiempo y redes neuronales, dada la naturaleza de los datos analizados en el datawarehouse, se realizó un breve análisis de las técnicas u algoritmos que intervenían en este tipo de modelo.

Se realizó el análisis comparativo de técnicas de minería de datos con lo cual se demostró que para esta investigación las de series temporales se ajusta a nuestro estudio, de acuerdo a los criterios de selección se obtuvo que para el presente trabajo de investigación las técnicas más adecuadas son ETS y redes neuronales. Siendo las redes neuronales auto regresiva el que mejor confiabilidad presenta, Tanto para el nivel primario y secundario con un 91% y 96% respectivamente. Podemos decir que Red neuronal auto regresiva obtuvo el nivel de confianza más elevado en comparación a ETS.

b. En la tesis intitulada: **“Aplicación de técnicas supervisadas de minería de datos para determinar la predicción de deserción académica”** de (Sulla Torres, 2015) se concluye que:

> Las diversas técnicas supervisadas de minería de datos se pueden aplicar de manera efectiva en los datos educativos prediciendo la deserción académica.
>
> Del análisis de las técnicas de clasificación se puede aplicar en los datos educativos para predecir el resultado de abandono o no del estudiante a los estudios y conocer la cantidad tentativa de grupos o secciones que se pueden asignar para su planificación.
>
> Las pruebas realizadas en el caso de estudio con el programa profesional de Ingeniería de Sistemas de la Universidad Católica de Santa María, en cuanto a la relación de atributos Abandono y Categoría: demuestran que la mayor cantidad de estudiantes que abandonan la carrera es de la categoría A. La relación de los atributos Abandono y Puntaje de Ingreso demuestran que la mayor cantidad de estudiantes que abandonan tiene un puntaje de ingreso promedio de

80.77. La relación de los atributos Abandono y cursos Aprobados demuestran que la mayor cantidad de estudiantes que abandonan tienen un promedio de 6.23 cursos Aprobados. La relación de los atributos Abandono y cursos Desaprobados demuestran que la mayor cantidad de estudiantes que abandonan la carrera tienen un promedio de 6.94 cursos desaprobados. La relación de los atributos Abandono y Total de cursos demuestran que la mayor cantidad de estudiantes que abandonan la carrera tienen un promedio de 11.25 cursos llevados. La relación de los atributos Abandono y Promedio Final demuestran que la mayor cantidad de estudiantes que abandonan la carrera tienen un promedio final de 7.84.

La eficiencia de los algoritmos de árboles de decisión C4.5 (J48) y RandomTree se puede analizar en función de su precisión comprobando que ambos algoritmos obtienen los mismos resultados.

c. En la tesis intitulada: "**Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela Profesional de Ingeniería**

de Computación y Sistemas, Universidad de San Martín de Porres, Lima-Perú" concluye que:

> Se logró predecir a través de técnicas estadísticos y minería de datos el rendimiento académico de los estudiantes ingresantes a la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres.
>
> Se identificaron los principales indicadores que permiten predecir el rendimiento académico de los estudiantes ingresantes y aplicaron pruebas que permitieron validar los resultados obtenidos como estadísticamente significantes.
>
> Se logró demostrar la influencia que tienen los diferentes factores de los ingresantes en el rendimiento académico en su primer ciclo de estudios. Las variables creadas a partir de la información extraída de la base de datos de la universidad permitieron crear perfiles que ayudaron a la identificación temprana de estudiantes que podrían encontrarse con dificultades en sus estudios.

Se aplicaron tres técnicas de minería de datos para realizar la predicción de rendimiento académico. El algoritmo C5.0 de árbol de decisiones es el que obtuvo los mejores resultados con una exactitud de predicción de 82.87%. Asimismo, la facilidad de interpretación de los resultados de la técnica de árbol de decisiones lo convierten en la mejor técnica a utilizar para este tipo de estudio.

2.1.2 A nivel internacional

a. En el artículo intitulado **"Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil"** de (Spositto, Etcheverry, Ryckeboer, & Bossero, 2010), la investigación se realizó aplicando el árbol de decisiones (j48) y el algoritmo FT sobre los datos de alumnos del período 2003-2008 para evaluar el rendimiento académico y la deserción de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas sobre los datos de los alumnos del periodo 2003 al 2008. Donde se obtuvieron como resultados con el algoritmo FT un 78,07 % mientras que con el algoritmo j48 72,53% llegando a la conclusión

que el algoritmo FT es mejor cuanto al rendimiento escolar es superior al algoritmo j48.

b. En el artículo intitulado "**Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos"** de (Valero, Salvador, & Garcia, 2010), en su investigación utilizaron las técnicas de minería de datos para poder predecir la deserción escolar en la Universidad Tecnológica de Izúcar de Matamoros, donde utilizaron los algoritmos tales como: C4.5 y el algoritmo de los k vecinos más cercanos. Obteniendo como resultado que las causas principales que desertan son: La edad, los ingresos familiares, El nivel de inglés. Llegando a la conclusión que con la propuesta planteada podrán determinar los factores de riesgos de manera oportuna.

Por otro lado (Timarán, Calderón, & Jiménez, 2013), en su investigación el objetivo fue la detección de patrones de deserción estudiantil partiendo a partir de los datos socioeconómicos, académicos, disciplinares e institucionales de los estudiantes de los programas de pregrado de la Universidad de Nariño e Institución Universitaria IUCESMAG, donde utilizando técnicas de

minería de datos su clasificación estuvo basada en árboles de decisión (j48), donde se seleccionaron los datos socio-económicos, académicos, disciplinares e institucionales de los estudiantes que ingresaron en los años 2004, 2005 y 2006. Obteniendo como resultado que la deserción en la Universidad de Nariño es estrictamente académico. Por lo que llegaron a la conclusión que aplicando las técnicas de clasificación y clustering sobre los datos de los estudiantes se ha obtenido un patrón común de deserción estudiantil, determinado por un promedio bajo y el tener materias perdidas en los primeros semestres de la carrera.

También (Silvaz Wanumen, 2010).En su investigación que hizo que lleva por nombre Minería de datos para la predicción de fraudes en tarjetas de crédito uso los algoritmos de árboles de clasificación (j48) y también uso las reglas de asociación (a priori), para la posible detección de fraudes a nivel de tarjetas de crédito.

Donde se compraron los dos algoritmos llegando a la conclusión que las reglas de asociación (a priori), fue menos efectiva que la de clasificación (j48).

2.2 BASES TEÓRICAS – CIENTÍFICAS

2.2.1. Deserción Estudiantil

(Bachman, Green, & Wirtanen, 1971), definen que la deserción escolar se ocasiona por aquellos estudiantes que interrumpen su asistencia a la escuela por varias semanas por diferentes razones, exceptuando aquellos por enfermedad.

(Morrow, 1985), define a la deserción cuando un estudiante el cual estuvo inscrito en la escuela, deja la misma por un largo periodo de tiempo y no se inscribió en otro colegio. Donde, no se toman en cuenta, a los estudiantes que estuvieron enfermos o fallecieron.

(Fitzpatrick & Yoels, 1992), se refieren a la deserción, cuando un estudiante deja la escuela sin graduarse, independientemente si regresan o reciben algún certificado equivalente.

(Lavado & Gallegos, 2005), elaboran su propia definición partiendo de las definiciones anteriores, donde llegaron a establecer que la deserción escolar se da siempre y cuando los individuos que habiendo asistido a la escuela el año anterior,

en el año actual o corriente no lo están haciendo, exceptuando solo a aquellos que han dejado de asistir por diversos motivos.

Por lo tanto la deserción escolar se define como aquel estudiante que realizo su matrícula o inscripción en un determinado año, y por causas determinadas deja inconclusa su preparación académica.

2.2.2. Minería de Datos

Según (Pérez & Santín, 2008), se refieren inicialmente a la minería de datos como un proceso de descubrimiento de nuevas y significativas relaciones, patrones al examinar grandes volúmenes de datos.

Por otro lado (Carrasco, 2011), expone que la minería de datos es el proceso de extracción de la información de interés partiendo de los datos, donde se entiende que solo el conocimiento es de interés siempre y cuando sea novedoso.

Según (Weiss & Indurkhya, 1998), Define a la minería de datos es la búsqueda de información valiosa en grandes volúmenes de datos. Se trata de un esfuerzo entre los humanos y las computadoras.

2.2.3. KDD: Proceso de Extracción de Conocimiento

Según (Usama & Wierse, 2002), refieren que el KDD es un proceso no trivial para poder identificar patrones válidos, novedosos, potencialmente útiles a partir de los datos.

Por otro lado (Guallart Romeu, 2010), contextualizan al KDD como al proceso de búsqueda y extracción de conocimiento partiendo de las bases de datos, mientras que la Minería de Datos es la parte de este proceso en la que se utilizan las técnicas de inteligencia artificial para obtener un modelo.

Hoy en día se puede confundir a la minería de datos con el proceso KDD. Donde la minería de datos forma parte del proceso de KDD como Se puede ver en la Figura. 1 (Guallart Romeu, 2010).

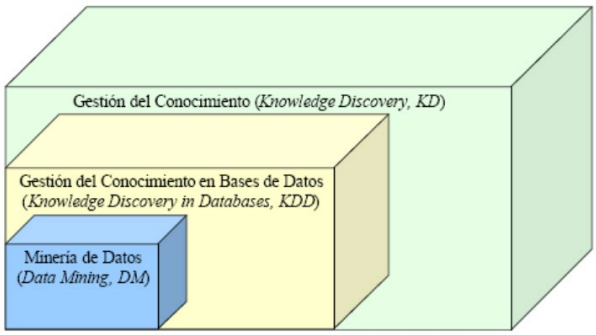


Ilustración 1: Comparación de los conceptos de minería de datos

El KDD forma parte de un área científica más amplia como es el descubrimiento de conocimiento que tiene otras muchas partes dentro de ella diferentes al KDD.

2.2.4. Fases de KDD

Según (Pernía & F., 2001) las fases KDD son:

1. Exploración del Dominio.
2. Recolección de los datos
3. Extracción de patrones en los datos
4. Inducir generalizaciones
5. Verificación del conocimiento
6. Transformación del conocimiento

Por otro lado (Brachman & Anand, 1996) define las fases así:

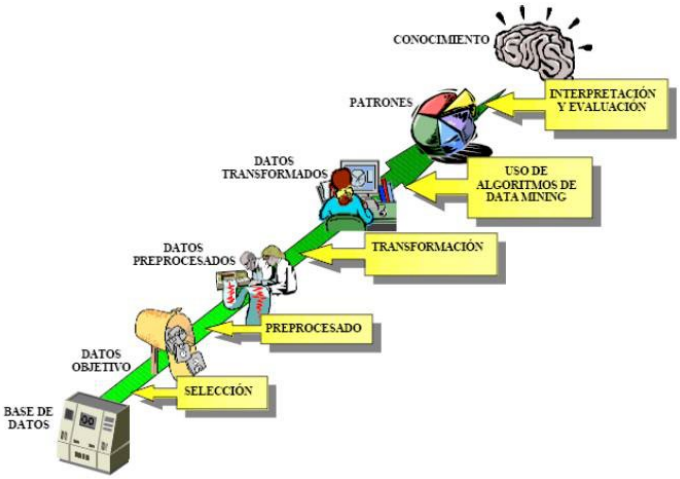


Ilustración 2: Proceso KDD

En (Hernández & Ferri, 2004), en su investigación expone las siguientes fases en el proceso de KDD:

1. Preparar los datos:

a. Especificar las fuentes de información las cuales puedan ser útiles.

b. Elaborar un esquema de almacén de datos (Data Warehouse) para poder unificar toda la información recogida.

c. Implantación del almacén de datos que permita la "navegación" y Visualización previa de sus datos, y asi poder diferenciar que atributos pueden ser interesantes para el estudio.

d. Selección, limpieza y transformación de los datos que se van a analizar. La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).

2. Minería de Datos:

a. Seleccionar y aplicar el método más apropiado.

b. Evaluación/Interpretación/Visualización.

c. Evaluar, interpretar, transformar y representar los patronos que se extraen.

d. Difundir y uso del nuevo conocimiento que se obtiene.

Según (WebMining Consultores, 2014), las etapas o fases del proceso KDD las divides en 5:

1. Selección de datos. En esta fase es determinar cuáles son las fuentes y el tipo de información que se va a utilizar. En esta fase los datos relevantes son extraídos desde la o las fuentes de datos.

2. Pre procesamiento. En esta fase se prepara y se limpia los datos que son extraídos desde las distintas fuentes de datos ya que van a ser necesario en las fases posteriores. En esta fase se emplean diversas estrategias para poder manejar datos faltantes, datos inconsistentes o que están fuera de rango, con

la finalidad de obtener una estructura adecuada para posteriormente transformarla.

3. Transformación. En esta fase consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables partiendo de las existentes con una estructura de datos apropiada. En esta fase se realizan las operaciones de agregación o normalización, donde se consolidan los datos de una forma necesaria para la fase siguiente.

4. Data Mining. Es la fase de modelamiento, en donde métodos inteligentes son aplicados con la finalidad de extraer patrones previamente desconocidos, validos, nuevos y potencialmente útiles que están contenidos u ocultos.

5. Interpretación y Evaluación. En esta fase es donde se identifican los patrones obtenidos y que son realmente interesantes, y que se basan en algunas medidas y se basándose en algunas medidas y se efectúa la evaluación de los resultados que se obtienen.

2.2.5. Clasificación de las Técnicas de Minería

En tanto la clasificación de la minería de datos entre autores se difiere:

Según, (Joshi, 1997), los componentes de la minería de datos son los siguientes:

1. Clustering: Donde se analizan los datos y se generan conjuntos de reglas que agrupan y clasifican los datos futuros.

2. Reglas de asociación: Son aquellas reglas o condiciones que presentan un grupo de objetos de una base de datos un ejemplo de regla de asociación o condición sería: “Un 30% de las transacciones que contienen toallitas de bebé, también contienen pañales; 2% de las transacciones contienen toallitas de bebé”. En el ejemplo antes mencionado el 30% es el nivel de confianza de la regla y 2% es la cantidad de casos que respaldan la regla.

3. Análisis de secuencias: Trata de descubrir patrones que suceden en una Secuencia determinada. Trabaja sobre datos que se presentan en distintas transacciones. “Muchos usuarios que han comprado X luego han comprado Y”.

4. Reconocimiento de patrones: Analiza la asociación de una señal de Información de entrada con aquella o aquellas con las que guarda mayor similitud, de entre las catalogadas por el sistema. Se usan para identificar causas de problemas o incidencias y buscar posibles soluciones, siempre y cuando se adecua a la base de información necesaria en donde buscar.

5. Predicción: Se busca determinar el comportamiento futuro de una variable o un conjunto de variables a partir de la evolución pasada y presente de las mismas o de otras de las que dependen. Las técnicas asociadas a estas herramientas tienen ya un elevado grado de madurez.

6. Simulación: Comparan la situación actual de una variable y su posible evolución futura.

7. Optimización: Resuelve el problema de la minimización o maximización de una función que depende de una serie de variables.

8. Clasificación: Permiten asignar a un elemento la pertenencia a un determinado grupo o clase. Se establece un perfil característico de cada clase y su expresión en términos de un algoritmo o reglas, en función de distintas variables. Se establece también el grado de discriminación o influencia de estas últimas. Con ello es posible clasificar un nuevo elemento una vez conocidos los valores de las variables presentes en él.

Mientras que para (Cabena, 1998), compone a la minería de datos en cuatro grandes operaciones soportadas por algunas técnicas comúnmente usadas

1. Modelización predictiva: Que usa las técnicas de:

a) Clasificación

b) Predicción de valores

2. Segmentación de bases de datos: Que usa técnicas de:

a) Clustering poblacional

b) Clustering por redes neuronales

3. Análisis de relaciones: Que utiliza las técnicas de:

a) Descubrimiento de asociaciones

b) Descubrimiento de secuencias de patrones

c) Descubrimiento de secuencias temporales similares

4. Detección de desviaciones:

a) Técnicas estadísticas

b) Técnicas de visualización

Según (Guallart Romeu, 2010) se puede clasificar las técnicas de aprendizaje de la siguiente manera:

1. Métodos inductivos:

Son aquellos que partiendo de los datos iniciales y del conocimiento generado son capaces de construir modelos que a partir de los datos generen los resultados.

2. Técnicas predictivas:

Interpolación: Es la generación de una función continua sobre varias dimensiones.

Predicción secuencial: Es cuando las observaciones están ordenadas en forma secuencial y se puede predecir el siguiente valor de la secuencia.

Aprendizaje supervisado: En éstas técnicas cada observación, compuesta por muchos valores de atributos, donde se interpone un valor de la clase a la que corresponde. Se genera un clasificador a partir de clases que se proporcionan. Es un caso particular de interpolación en el que la función genera un valor discreto en lugar de continuo

3. Técnicas descriptivas:

Aprendizaje no supervisado: Es el conjunto de observaciones las cuales no tienen algunas clases asociadas. Tiene como objetivo la detección regularidades en datos de cualquier tipo: agrupaciones de datos parecidos o próximos, contornos de delimitación de grupos, asociaciones o valores anómalos.

Métodos abductivos: Se pretende, partiendo de los valores generados y de las reglas, obtener los datos de origen. El

objetivo es la explicación de evidencia con respecto a los sucesos que se han producido, tal cual haría un investigador privado, que a partir de las consecuencias de los hechos y de ciertas reglas.

2.2.6. Técnicas de minería de Datos

2.2.6.1. Arboles de decisiones

Los árboles de decisión son una de las formas más populares de Minería de Datos porque tienen una representación sencilla de problemas con un número finito (y a ser posible reducido) de clases. Además son modelos comprensibles y proposicionales (Hernández & Ferri, 2004).

Un claro ejemplo de un árbol de decisión en (Guallart Romeu, 2010). Donde partir del valor de la variable X8, si el valor es menor de 3.2 se continuará la toma de decisiones por la rama izquierda y si es mayor o igual se continuará por la rama de la derecha. A partir de aquí cada rama tiene una variable separadora con un valor de separación, y así sucesivamente formando un árbol.

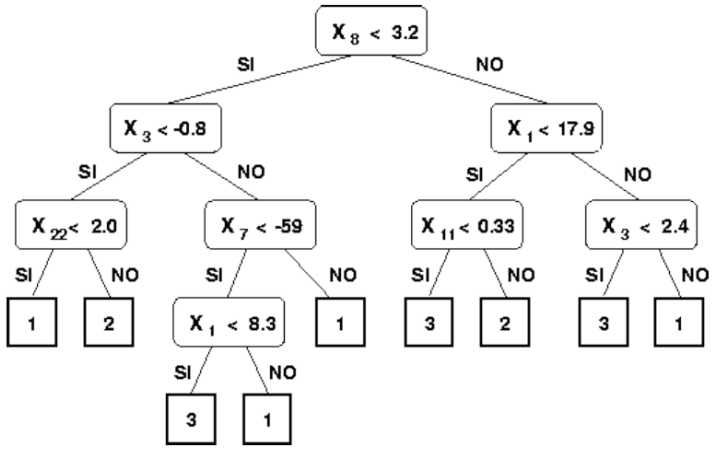


Ilustración 3: Árbol de decisiones

Donde también (Mazo & Bedoya, 2010), puntualizan que un árbol de decisión es una estructura en la cual cada nodo interno significa una prueba sobre uno o varios atributos, donde cada rama representa una salida de la prueba y los nodos hojas representan clases.

2.2.6.2. C4.5

Según (Quinlan, 1993), y su versión comercial C5.0 Es una extensión de ID3, el cual permite que se trabaje con valores continuos para los atributos, donde se separan los resultados en dos ramas: una para aquellos Ai=N y la otra para Ai>C4.5, donde es capaz de trabajar con ejemplos que contienen valores desconocidos y es tolerante a datos con ruido.

2.2.6.3. Métodos Bayesianos

Una de las características primordiales de los métodos bayesianos es el uso de distribuciones de probabilidad para cuantificar incertidumbre de los datos que se desea modelar. Estos métodos proporcionan una metodología práctica para la inferencia y predicción y, en última instancia, para tomar decisiones que involucran cantidades inciertas (Hernández & Ferri, 2004)

(Hernández & Ferri, 2004), dice que “es una de las que más se han utilizado en problemas de inteligencia artificial, con ello en el aprendizaje automático y minería de datos, ya que es un método práctico para realizar inferencias a partir de los datos, la misma que se basa en estimar la probabilidad de pertenecía (a una clase o grupo) mediante la estimación de las probabilidades, utilizando para ello el teorema de Bayes”.

2.2.6.4. Redes neuronales artificiales

Según (Hernández & Ferri, 2004) señala que las redes neuronales posee dos tipos de aprendizaje uno es el supervisado, en el mismo que se le proporciona un conjunto de datos de entrada y la respuesta correcta es útil en tareas de regresión y clasificación. Y el aprendizaje no supervisado solo se le da a la red un conjunto de datos de entrada y la red debe

auto-enseñarse para proporcionar una respuesta, este aprendizaje es útil para las tareas de agrupamiento.

Donde las redes neuronales han sido utilizadas en diversas áreas de estudio tal es el caso en la predicción de mercados financieros, control de robots, etc. (Guallart Romeu, 2010).

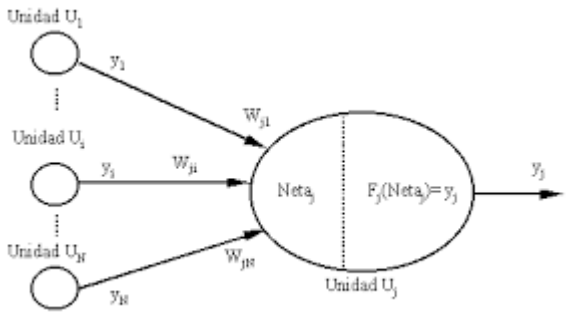


Ilustración 4: Esquema de una Neurona Artificial con sus principales elementos

En redes neuronales hay dos tipos principales de aprendizaje en RNA:

a) Aprendizaje supervisado: Estos algoritmos precisan que cada vector de entrada se empareje con su correspondiente vector de salida. Mientras que el entrenamiento se basa en la de mostrar un vector de entrada a la red, donde se calcula la salida de la red y después se compara con la salida deseada y por otro lado el error o diferencia resultante se emplea para realimentar la red y modificar los pesos de acuerdo con un algoritmo que tiende a minimizar el error. (Olabe Basogain, 2008)

b) Aprendizaje no supervisado: Son aquellos sistemas donde al aprendizaje solo se le da un determinado conjunto de datos de entrada y la red debe auto-enseñarse y así proporcionar una respuesta, donde este aprendizaje es de gran utilidad para tareas de agrupamiento. (Olabe Basogain, 2008).

2.2.6.5. K – means

Este algoritmo es uno de los más utilizados con lo que respecta al agrupamiento de datos, es el K-Medias o también conocido como K-Means por ser uno de los más veloces y eficaces. El algoritmo trabaja con un método de agrupamiento por vecindad, en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar sin etiquetar.

El propósito de K-Means es ubicar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares. (Moody & Darken, 1989).

Todo ejemplo nuevo, una vez que los prototipos han sido correctamente situados, es comparado con estos y asociado a aquel que sea el más próximo, en los términos de una distancia previamente elegida.

Normalmente, se utiliza la distancia euclidiana. El objetivo que se busca mediante el algoritmo K-Means es minimizar la varianza total intragrupo o la función de error cuadrático, para que el algoritmo pueda generar los mejores resultados.

2.2.6.6 Series de tiempo

Es aquel conocimiento que se obtiene a través de la recopilación de datos, la observación o el registro de intervalos de tiempos regulares, donde que a partir de ese conocimiento y con el supuesto de que no se producirán cambios, y así poder realizar predicciones. Algunas definiciones que se usan con esta técnica son:

A) Tendencia: Es aquel componente a largo plazo la cual representa la disminución o crecimiento en un amplio periodo de tiempo.

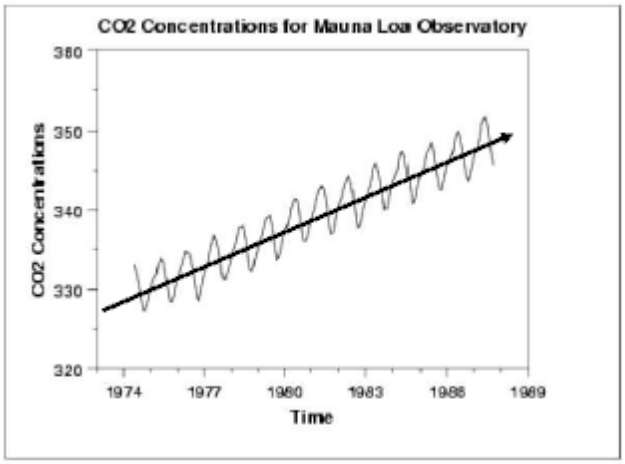


Ilustración 5: Gráfico de Tendencia de un conjunto de datos de los años 1974-1989

B) Estacionalidad: Es aquel elemento en el cual se presenta en series de frecuencia inferior a la anual, y se presume oscilaciones a un corto plazo regular, inferior al año y amplitud regular.

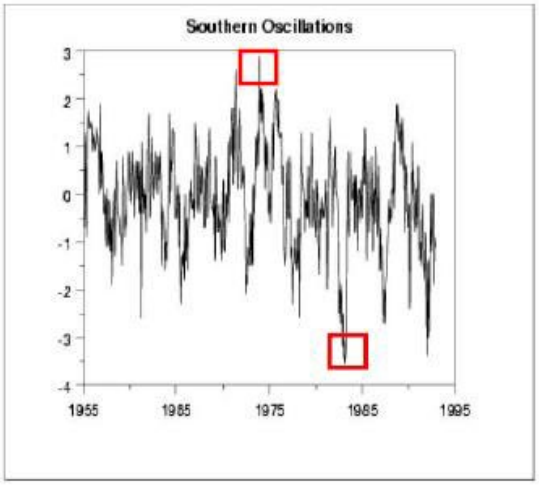


Ilustración 6: Gráfica de valores en el tiempo, donde se observa la estacionalidad

C) ETS (Exponential smoothing state)

(Hyndman R. J., 2014), Los Métodos de suavización exponencial han existido desde la década de 1950, y son los métodos de pronóstico más populares utilizados en los negocios y la industria. Recientemente, suavizado exponencial ha revolucionado con la introducción de un marco de modelización completa incorporando innovaciones modelos de estado espacio, cálculo de probabilidades, los intervalos de predicción y los procedimientos para la selección del modelo.

ETS (M, N, N) Suavización exponencial simple con errores multiplicativos: Según (Hyndman,2014) se puede especificar modelos con errores multiplicativos escribiendo los errores aleatorios de un solo paso como errores relativos:

$$\varepsilon_t = \frac{y_t - \hat{y}_{t|t-1}}{\hat{y}_{t|t-1}}$$

Entonces se puede escribir la forma multiplicativa del modelo de espacio de estados como se muestra:

$$y_t = \ell_{t-1}(1+\varepsilon_t)$$
$$\ell_t = \ell_{t-1}(1+\alpha\varepsilon_t).$$

D) Holwinters:

Es una variante, donde es conocida como alisado exponencial líneas con doble parámetro, donde consigue la eliminación del sesgo de la predicción de una serie de tendencia, a través de la inclusión en la media móvil de un componente de tendencia.

Por otro lado comprando con diversas técnicas, tal como ARIMA, donde el tiempo necesario para el cálculo en la predicción es considerablemente rápido.

De hecho, Holt-Winters es utilizado por diversas compañías para el pronóstico de la demanda a corto plazo siempre y cuando los datos de venta contengan tendencia y patrones estacionales de un modo subyacente.

2.2.7. Metodologías para la aplicación de minería de datos

a) CRISP – DM (Cross Industry Standard Process for Data Mining)

CRISP-DM organiza el desarrollo de un proyecto de Data Mining en una serie de fases o etapas, con tareas generales y específicas que permitan cumplir con los objetivos del proyecto. Estas fases funcionan de manera Cíclica e iterativa, pudiendo regresar desde alguna fase a otra anterior.

Se basa en función a un modelo jerárquico de procesos, donde se establece un ciclo de vida de los proyectos de explotación de información.

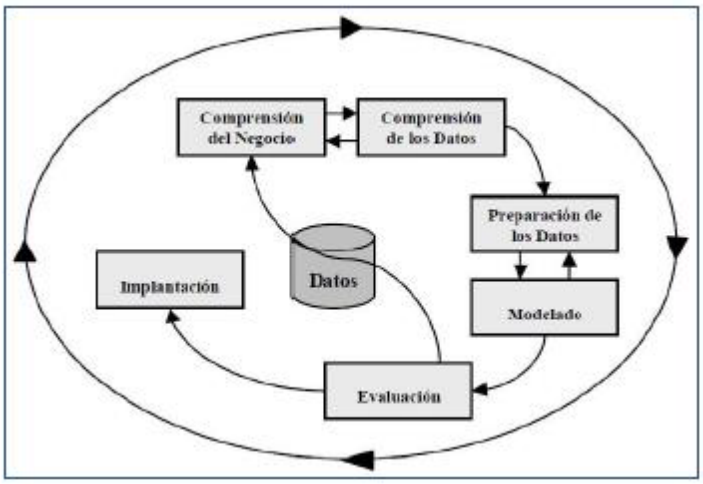


Ilustración 7: Fases del modelo CRISP – DM

Según (Orallo Hérnandez, 2015) las fases de la metodología CRISP – DM son las siguientes:

a. Comprensión del negocio: Es donde se infiere tanto como los objetivos y requerimientos del proyecto desde una perspectiva de negocio.

b. Comprensión de los datos: Se selecciona y adapta los datos, para poder identificar los problemas de calidad de datos y así obtener datos potenciales para poder analizar.

c. Preparación de los datos: Transformación de los datos. Se seleccionan los datos a utilizar y éstos pasan a una fase de limpieza, estructuración, integración y formateo.

d. Modelamiento y evaluación: Selección y aplicación de Data Mining e Interpretación y evaluación. Se selecciona la técnica a utilizar, construyendo el modelo, para luego ser sometido a diferentes pruebas y evaluaciones.

e. Despliegue del proyecto: Es donde se explota todo el potencial de los modelos y así intégralos en los procesos de toma de decisión de organización, y así difundir el conocimiento extraido, etc.

b) SEMMA

La metodología semma se caracteriza principalmente por la que toma su nombre de las etapas que esta metodología define para procesos de explotación de información, estas etapas son:

muestreo (sample), exploración (explore), modificación (modify), modelado (model) y valoración (assess).

La metodología semma fue desarrollada por la empresa SAS Institute Inc.,una de las mayores organizaciones relacionadas con el desarrollo con el software de inteligencia de negocios SEMMA esta desarrollada para aplicarla sobre la herramienta de minería de datos “SAS Entreprise Miner”.

2.2.8. Herramientas de Minería de datos

Para la aplicación de técnicas de minería de datos se clasificaría en dos Librerías y herramientas específicas:

Donde las librerías de Minería de datos son un conjunto de métodos donde se implementan funcionalidades y utilidades básicas como el acceso a datos, modelos de redes neuronales, métodos bayesianos, exportación de resultados Las librerías se encargan principalmente de facilitar el desarrollo de las tareas de Minería de Datos que son más complejas, como el diseño de experimentos. El problema de las librerías, es que es precisa la comprensión de conocimientos de programación.

Algunas de las Librerías más importantes son:

1. **Xelopes (Extended Library For Prudys Embedded Solution):** Es una librería bajo la licencia pública GNU

para el desarrollo de aplicaciones de Minería de Datos. Esta librería está implementada para que sea eficiente para la mayoría de los algoritmos de aprendizaje, por eso, es importante destacar que el usuario puede desarrollar aplicaciones particulares de Minería de Datos. Sus principales características son:

1. Acceso a datos
2. Modelos de redes neuronales
3. Métodos de agrupamiento
4. Métodos de reglas de asociación
5. Árboles lineales
6. Árboles no lineales

2. **MIc++ (Machine Learning Library In C++):** Es un conjunto de librerías que fueron desarrolladas por la Universidad de Standford. La mayoría de las versiones son bajo dominio de investigación, a excepción de la versión 1.3.x, que se distribuye bajo licencia de dominio público. Las principales características son:

   1. Acceso a datos.
   2. Transformaciones de datos

3. **Suites:** Posee las mismas capacidades que el procesamiento de datos, los modelos de análisis, el diseño de experimentos o el soporte gráfico para la visualización de resultados. En este caso, Suites destaca porque existe una interfaz que facilita la interacción entre el usuario y la herramienta.

4. **R-Project:** Es un entorno de trabajo basado en los entornos de programación S y S-PLUS desarrollados a principios de los años noventa del pasado siglo por Bill Venables y David M. Como señalan Venables et al. (2011), es un entono integrado de facilidades informáticas para la manipulación de datos, el cálculo y la generación de gráficos. R-Project pretende convertirse en un sistema internamente coherente que se caracterizaría por un desarrollo basado en la contribución relativamente altruista de la comunidad científica. (López Puga, 2010)

5. **SPSS Clementine:** Es uno de los sistemas de Minería de Datos más conocidos. Posee una herramienta visual desarrollada por ISL que tiene una arquitectura cliente/servidor. Este sistema se caracteriza por:

1. Acceso a datos.
2. Procesamiento de Datos.
3. Técnicas de Aprendizaje.
4. Técnicas de evaluación de modelos.
5. Visualización de resultados.
6. Exportación.

6. **Weka (Waikato Environment For Knowledge Analysis):** Es una herramienta visual de libre distribución desarrollada por los investigadores de la Universidad de Waikato en Nueva Zelanda. Sus principales características son:

   1. Acceso a los datos desde un archivo en formato ARFF.
   2. Pre procesado de datos.
   3. Modelos de Aprendizaje.
   4. Visualización del entorno.

7. **Kepler:** Sistema desarrollador y transformado en una herramienta comercial distribuida por Dialogis. Posee múltiples modelos de análisis. Sus principales herramientas de aprendizaje son:

1. Árboles de decisión.
2. Redes neuronales.
3. Regresión no lineal.
4. Aplicaciones estadísticas.

8. **Odms (Oracle Data Mining Suite):** Está diseñado sobre una arquitectura cliente servidor; ofrece una gran versatilidad en cuanto al acceso a grandes volúmenes de información. Se caracteriza principalmente por:

    1. Acceso a datos en diversos formatos: almacenes de datos, bases de datos relacionales como SQL, Oracle,etc.
    2. Pre procesado de datos: muestreo de datos, patrones de datos.
    3. Modelos de aprendizaje: redes neuronales, regresión lineal.
    4. Herramientas de visualización.

9. **Yale:** Herramienta de aprendizaje automático implementado en Java por la Universidad de Dormund. El sistema incluye operaciones para:

    1. Importación y pre-procesamiento de datos

2. Aprendizaje automático

3. Validación de modelos

2.3 DEFINICIÓN DE TÉRMINOS BÁSICOS

2.3.1. Método

Modo ordenado y sistemático de proceder para lograr un fin/conjunto de reglas (Getoor & Ben, 2007)

2.3.2. Metodología

Conjunto de métodos que se siguen en una disciplina científica/ciencia del método y de la sistematización científica. (Grudnitsky, 1992)

2.3.3. Predicción

Es la acción de aquello que supuestamente va ocurrir. Donde se puede predecir partiendo de conocimientos científicos, revelaciones o de algún tipo de indicios.

2.3.4. Deserción Estudiantil

(Bachman, Green, & Wirtanen, 1971), Refieren que la deserción institucional se origina siempre y cuando aquellos

estudiantes irrumpen su asistencia a la universidad por varias semanas.

2.3.5. Minería de Datos

Minería de datos es un conjunto de técnicas y tecnologías donde permitirían explorar grandes bases de datos, de manera automática, donde tiene como objetivo el encontrar patrones repetitivos, para así poder explicar el comportamiento de los datos en un contexto determinado.

2.4 FORMULACIÓN DE LA HIPÓTESIS

2.4.1 Hipótesis general

La deserción estudiantil puede predecirse mediante técnicas de minería de datos de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.

2.4.2 Hipótesis específica

Las técnicas de minería de datos pueden aplicarse para predecir la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.

2.5 IDENTIFICACIÓN DE VARIABLES

2.5.1 Variable independiente

Técnicas de minería de datos

2.5.2 Variable dependiente

Deserción estudiantil.

2.6 DEFINICIÓN OPERACIONAL DE VARIABLES E INDICADORES

Tabla 1: Operacionalización de variables e indicadores

Variables	Indicadores
Variable independiente Técnicas de minería de datos.	Técnicas
Variable dependiente Deserción estudiantil.	Predicción

CAPITULO III

METODOLOGÍA Y TÉCNICAS DE INVESTIGACIÓN

3.1 TIPO Y NIVEL DE INVESTIGACIÓN

3.1.1 Tipo de investigación

El tipo de la investigación del presente estudio es aplicada, ya que buscamos la aplicación de los conocimientos de Minería de datos para resolver problemas prácticos inmediatos.

3.1.2 Nivel de investigación

Por naturaleza del estudio el nivel de investigación es predictivo.

3.2 MÉTODOS DE INVESTIGACIÓN

El método utilizado en el desarrollo de la presente es el método deductivo que es utilizado en las áreas científicas, donde se recolectan datos de hechos y fenómenos para llegar a una hipótesis o teoría.

3.3 DISEÑO DE LA INVESTIGACIÓN

El diseño de la investigación es transeccional del tipo correlacional causal. Se busca describir la relación que existe entre la deserción estudiantil y las técnicas de minería de datos.

El diseño de la investigación es el siguiente:

En donde:

X1: Variable Indicador técnica de minería de datos

Y: Rendimiento académico de los ingresantes

3.4 POBLACIÓN Y MUESTRA

3.4.1 Población

La población objeto de investigación está constituida por 1810 alumnos matriculados en el periodo 2017-A de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.

3.4.2 Muestra

Para el presente estudio la muestra utilizada es no aleatoria, por selección intencionada o muestreo de conveniencia, constituida por 218 alumnos de Escuela Profesional de Sistemas y Computación de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.

En esta muestra se trabajó con las matrículas y notas de los alumnos desde el periodo 2013 al 2017-A.

3.5 TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

Se utilizó información histórica extraída de las bases de datos de la Oficina de Admisión y Oficina de Informática de la Universidad

Nacional Daniel Alcides Carrión relacionado con los alumnos y su rendimiento académico tales como: promedio ponderado por semestre, modalidad de ingreso, situación familiar y socio-económica, entre otros.

Se aplicaron técnicas de extracción, transformación y carga de datos para convertir la información extraída de una base de datos transaccional hacia un formato apropiado para la aplicación de las técnicas de minería de datos.

También se usaron las siguientes técnicas:

Análisis documental, Consiste en extraer la información de los diferentes, libros, papers, artículos, los cuales presentan una serie de teorías, técnicas, métodos que dan solución a determinados problemas. Todo servirá para limitar la investigación y caracterizar el modelo a estudiar, para analizar resultados obtenidos con las técnicas aplicadas.

Observación: Es el registro visual de lo que ocurre en una situación real, donde se clasifican los acontecimientos con algún esquema y dependiendo el problema que se estudia. En esta técnica es debido está atento para determinar de una forma adecuada todos los resultados confiables de las predicciones.

3.6 TÉCNICAS DE PROCESAMIENTO Y ANÁLISIS DE DATOS

Las técnicas utilizadas para el procesamiento y análisis de los datos obtenidos en el transcurso de la investigación, fueron los siguientes:

- Software de análisis estadístico IBM SPSS Statistics 25
- Software Weka 3.8.3: es una plataforma de software para el aprendizaje automático.

3.7 TRATAMIENTO ESTADÍSTICO

Las técnicas estadísticas que se utilizaron fueron de acuerdo al tipo y diseño de investigación, está es:

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados (GLM por sus siglas en inglés) que usa como función de enlace la función logit. Las probabilidades que describen el posible resultado de un único ensayo se modelan, como una función de variables explicativas, utilizando una función logística.

Se usó la regresión logística binaria.

SEGUNDA PARTE: DEL TRABAJO DE CAMPO O PRÁCTICO

CAPITULO IV

RESULTADOS Y DISCUSIÓN

4.1 DESCRIPCIÓN DEL TRABAJO DE CAMPO

Desarrollaremos la metodología de CRISP-DM siguiendo las fases y tareas que propone.

4.1.1 Comprensión del negocio.

Esta fase inicial se enfoca en entender los objetivos y requerimiento del proyecto convirtiendo esto en la definición del problema de minería de datos.

a) Determinación *de objetivos de negocio*

La Universidad Nacional Daniel Alcides Carrión es una comunidad integrada por profesores, estudiantes, y graduados. Se dedica al estudio, la investigación, la educación, la difusión del saber, la cultura, y a la extensión y proyección social.

Tiene como objetivos, entre otros, de docencia y de gestión integral:

- Programas profesionales acorde a la región.
- Mejora permanente de los servicios de apoyo académico.

b) Evaluación de la situación

En el ámbito universitario, la UNDAC adolece de la problemática de una tasa significativa de deserción estudiantil por diferentes motivos, no todos conocidos, lo que ocasiona que se enfrente a esa situación cada vez que se presente, afectando la planificación realizada por cada escuela profesional.

Los datos de los alumnos son capturados desde el momento en que ingresa a la Universidad, llenado sistema académico de la universidad donde figura los siguientes atributos: Nombre y Apellidos, Dirección de vivienda, Numero de hermanos,

De la Oficina de Admisión se recaban los atributos: Puntaje de Ingreso, Opción de ingreso.

Estos atributos han sido almacenados en diferentes fuentes de datos a lo largo del tiempo y se ha centralizado en la oficina de Informática de la Universidad.

Los recursos inicialmente disponibles son los siguientes:

- Materiales: Weka versión 3.8.3 como herramienta de Minería de datos y SQL Server Business Intelligence para el ETL.

- Humanos: El autor de la investigación, El Jefe de la Oficina de Informática,

- Datos del trabajo: Se dispone de una serie de datos académicos de los alumnos de la carrera profesional de Ingeniería de Sistemas de la UNDAC comprendida entre los años 2013 y 2017 que hacen un total de 218 registros de alumnos y 7238 registros de calificaciones.

c) Determinación de los objetivos de la Minería de datos

Objetivo de minería de datos: Dar soporte mediante técnicas de Minería de datos a los objetivos de la investigación

Objetivos específicos en la investigación:

- Realizar un estudio estadístico genérico de los datos.

- Encontrar la cantidad de alumnos que pertenecen a la categoría de abandono o no abandono.

Conocer estos objetivos permitirá determinar de una mejor manera la planificación en la apertura del año académico y distribución de aulas y secciones de la escuela profesional.

4.1.2. Comprensión de los datos

a) Recolección de los datos iniciales

Los datos iniciales fueron proporcionados por la Oficina de Informática de la UNDAC, los cuales se recogieron en diferentes momentos.

Estos datos fueron entregados en el formato de una hoja de cálculo Excel como sigue:

Pestaña ALUMNOS_SISTEMAS:

Campos:

- Código
- Apellido Paterno
- Apellido Materno
- Nombres
- Fecha de Nacimiento
- Sexo
- Civil
- Dirección
- Puntaje_Ingreso
- Modalidad_Ingreso
- Puesto_Ingreso
- Vive_Con

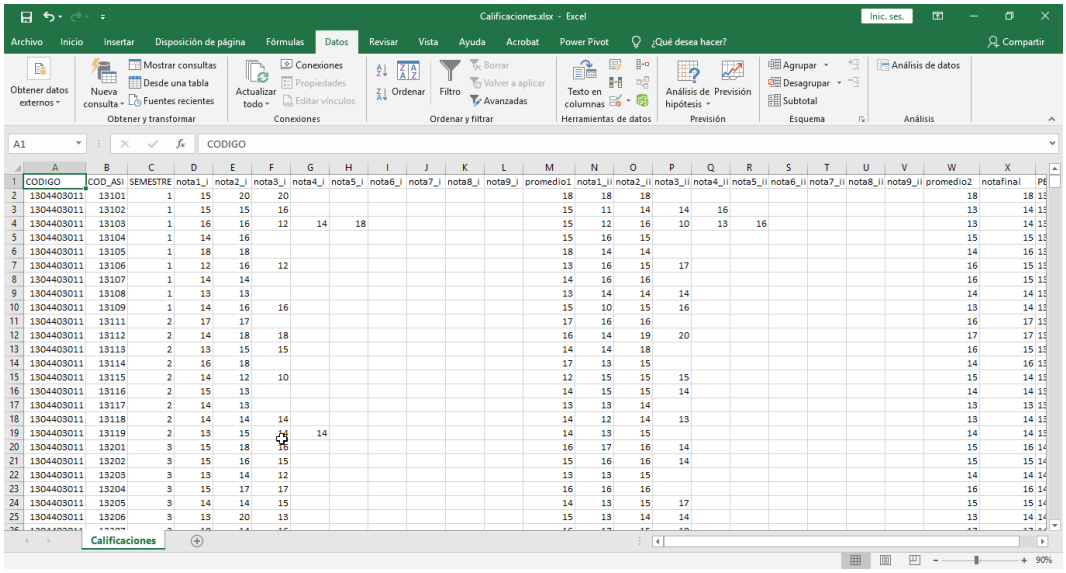


Ilustración 8: Ilustración 8 Alumnos de Sistemas 2013 al 2017

Pestaña CALIFICACIONES:

Campos:

- Código (Alumno)
- Cod_Asi (Código de la Asignatura)
- Semestre
- nota1_i
- nota2_i
- nota3_i
- nota4_i
- nota5_i
- nota6_i
- nota7_i
- nota8_i
- nota9_i

- promedio1
- nota1_ii
- nota2_ii
- nota3_ii
- nota4_ii
- nota5_ii
- nota6_ii
- nota7_ii
- nota8_ii
- nota9_ii
- promedio2
- notafinal
- PERIODO

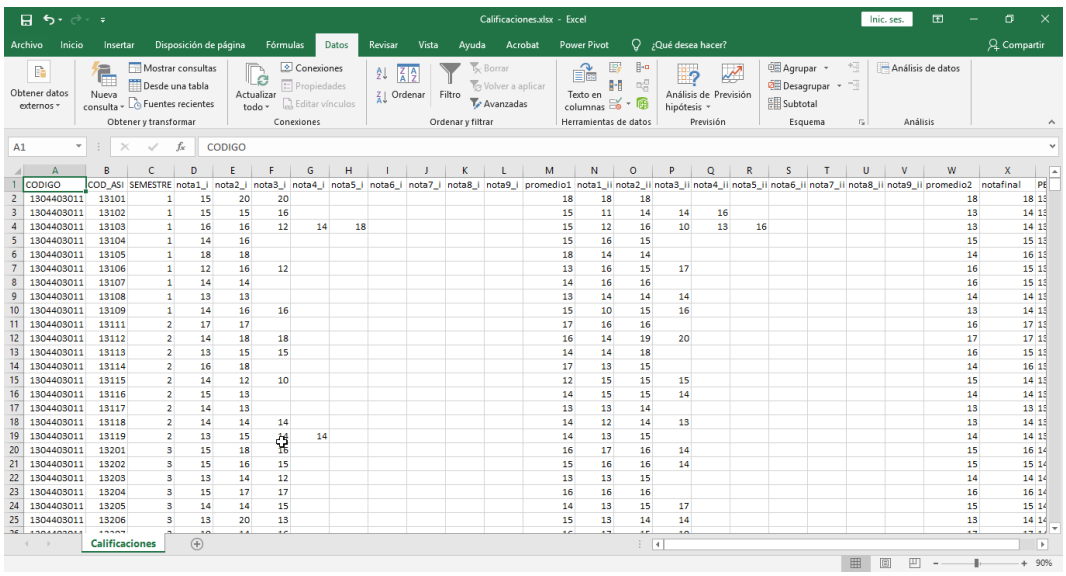


Ilustración 9: Calificaciones de los alumnos de Sistemas 2013 al 2017

b) Describir los datos

Los datos proporcionados por la Oficina de Informática, fue entregado en un archivo de hoja cálculo Excel correspondiente a los alumnos desde el año 2013 hasta el año 2017.

Todos los campos listados en el punto a) tienen un formato de campo del tipo TEXTO, ya que al parecer los datos solicitados a la Oficina de Informática fueron recabándose de diferentes tablas y se juntaron en una hoja de cálculo.

c) Explorar los datos.

Al realizar la exploración de los datos, se apreció que se necesitaba algunos campos que sirvan como referencia significativa del alumno y luego se pueda utilizar como fuente de entrada para los modelos en Weka, para ello se tuvo que realizar las siguientes actividades:

1. Transformar los campos necesarios del tipo Texto a su correspondiente tipo real, como por ejemplo el Promedio de cada asignatura convertirlo a tipo Numérico.
2. Generar nuevos campos como Promedio final de todos los cursos para cada alumno, asignaturas Aprobadas y asignaturas Desaprobadas.

d) Verificar la calidad de los datos

En esta tarea se observó que existían algunos registros de alumnos:

- No tenían notas ni promedio alguno por lo que se procedió a identificarlos y omitirlos.

- Algunas notas figuran con un símbolo NSP (Aplazados) que representa que no tienen nota, por tal motivo no se podían procesar numéricamente.

4.1.3. Preparación de datos

En esta etapa se convierte todas las actividades para construir la base de datos final (datos que alimentara a la herramienta de modelamiento). Para ello se procedió a crear una base de datos en Microsoft SQL Server y a migrar los datos de la Hoja de Cálculo a la nueva base de datos, además, la preparación tratará de:

- Eliminar valores anómalos, inconsistencias, valores ausentes, etc.

- Seleccionar datos a tratar.

- Modificar valores para su mejor tratamiento en función del algoritmo.

a) Selección de datos

La selección de los datos permitió distinguir: primero, que los campos que no tenían valor alguno (Dirección, Civil, Hijos, Vive_con) se omitan,

segundo, para aquellos registros que no están relacionados (alumnos que no tiene notas) no se consideren, de tal forma que solo se utilicen los datos que nos sirvan para el minado.

b) Limpieza de datos

Las tareas de limpieza de datos a realizar son:

- Conversión de tipos: El campo de Promedio se convirtió de tipo de dato Texto a numérico.
- Modificar datos erróneos: algunos símbolos especiales por el alfabeto utilizado en la fuente de datos.
- Añadir datos faltantes.
- Trimming: Eliminar espacios sobrantes antes y después de las cadenas.

Transformación ETL para la obtención de la Vista minable.

Para lograr ello se utilizó el proceso ETL (Extraer, Transformar y Cargar) que permite mover datos desde diferentes fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

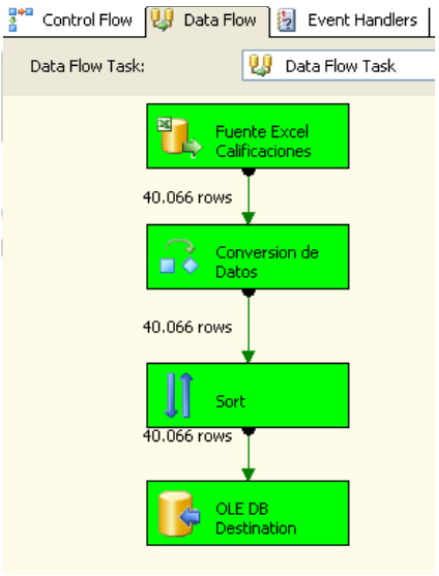


Ilustración 11: ETL para la conversión de datos.

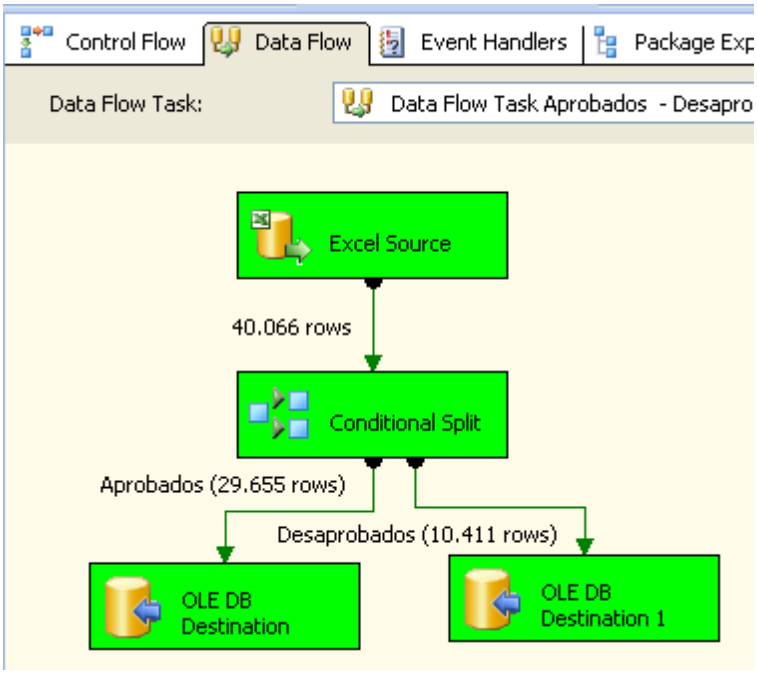


Ilustración 12: ETL para generar el Promedio Final y la cantidad de Aprobados y Desaprobados

c) . Construir datos

La construcción de los datos ha permitido generar las cuatro tablas que permitan organizar mejor lo datos quedando como sigue: ASIGNATURAS, CALIFICACIONES, ALUMNOS y ALUMNOS_DETALLE con sus respectivos atributos (campos).

Ilustración 13: Diagrama de la Base Datos Académico

La tabla ALUMNOS_DETALLE es la que se va a utilizar en la Herramienta de Minería de Datos WEKA, para ello se ha generado el correspondiente archivo .cvs (comma-separated values) con los siguientes campos: Codigo, OrdenMerito, Puntaje, ModalidadIngreso, Puntaje, Aprobados, Desaprobados, TotalCursos, CreditosAprobados, PromFinal, Abandono.

d) Integrar datos

La integración de los datos se ha realizado para obtener la tabla ALUMNOS_DETALLE que es la que finalmente se ha utilizado para aplicar la Minería de datos, en esta se encuentra los siguientes atributos: OrdenMerito (Orden de merito - ingreso), ModalidadIngreso (Modalida de ingreso), Puntaje (Puntaje de

ingreso), Aprobados (Asignaturas aprobadas), Desaprobados (Asignaturas desaprobadas), TotalCursos (Total de Asignaturas), CreditosAprobados (Total de créditos aprobados), PromFinal (Promedio Final), Abandono (Si / No)

e) Formatear datos

El formateo de los campos se ha realizado para tener una mejor comprensión en el aspecto sintáctico sin cambiar su significado, dándole nuevas denominaciones y etiquetas más comprensibles.

4.1.4. Modelado

a) Selección de la técnica de modelado

Antes de seleccionar el modelo apropiado, debemos de enfocarnos en nuestro objetivo: ¿Cuál el propósito de buscamos?, el propósito que buscamos es la predicción de la deserción académica.

A continuación decidimos el tipo de predicción más apropiado, que será el de clasificación que predice en que categoría o clase caerá el resultado, en nuestro caso será de Abandona y No abandona. Entonces nuestro modelo elegido será un árbol de decisión para la clasificación.

Para ello comparamos algunos clasificadores populares: J48 es una implementación open source en lenguaje de programación

Java del algoritmo C4.5 en la herramienta Weka de minería de datos. Luego también utilizaremos RandomTree para realizar la tarea comparativa.

b) Generación de la prueba de diseño

Se ha generado un archivo de prueba con registros adicionales, para las tareas de Minería de datos supervisadas como clasificación, por lo tanto se separa el conjunto de datos en datos originales y datos de prueba, se construye el modelo sobre el conjunto originales, y se evalúa el resultado del conjunto de prueba.

c) Construcción del modelo

Para llevar a cabo el objetivo de esta investigación, aplicaremos minería de datos usando el paquete de software Weka.

Weka es un acrónimo de Waikato Environment for Knowledge Analysis, es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario.

WEKA se distribuye como software de libre distribución desarrollado en Java.

Está constituido por una serie de paquetes de código abierto con diferentes técnicas de preprocesado, clasificación,

agrupamiento, asociación, y visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados.

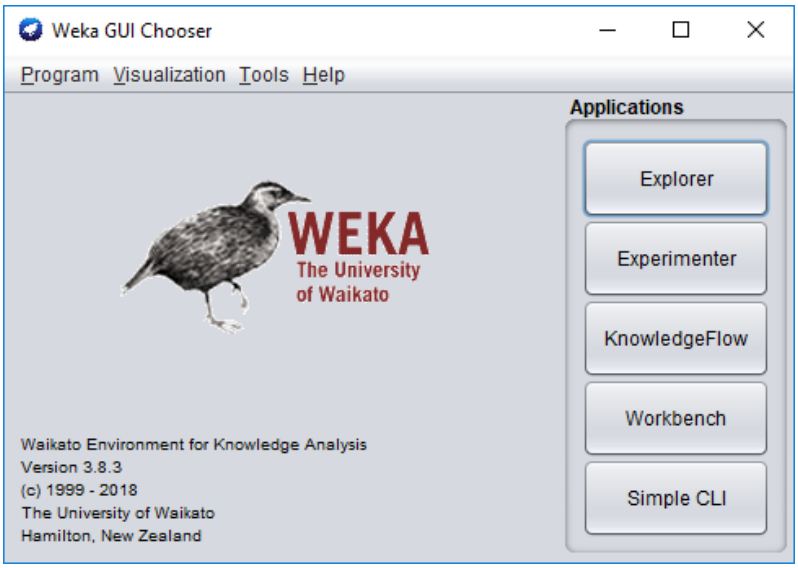


Ilustración 14: Ventana de inicio de Weka

Los datos de entrada a la herramienta, sobre los que operarán las técnicas implementadas, deben estar codificados en un formato específico, para nuestro caso fue .csv, que es un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en la que las columnas se separan por comas y las filas por saltos de línea.

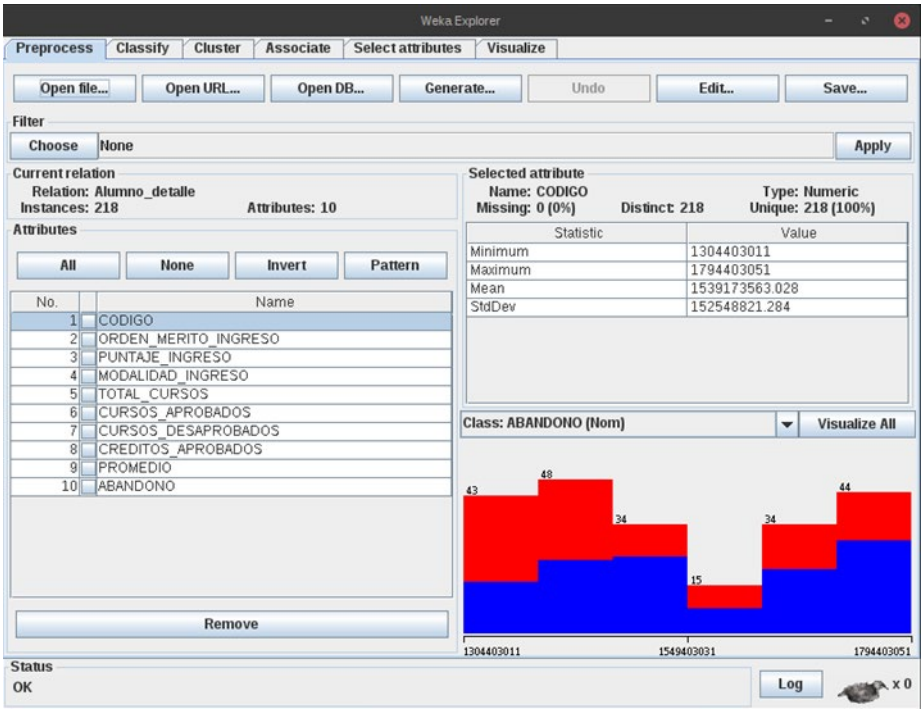


Ilustración 15: Archivo .csv con los atributos cargados en Weka

4.1.4. Pruebas

El principal objetivo del análisis predictivo es la mejora de la actividad académico del estudiante. El algoritmo de clasificación C4.5 (J48) se analiza usando los siguientes métodos (Kumar and Vijayalakshmi, 2011):

a) La precisión del algoritmo es medido usando la comparación de los datos originales con los de prueba de los estudiantes.

b) La eficiencia del algoritmo es medido comparando el algoritmo C4.5 (J48) con el algoritmo RandomTree

a) De la precisión del algoritmo

Los datos originales generados en el archivo .cvs se cargan en la herramienta de minería de datos Weka y luego se selecciona el algoritmo J48.

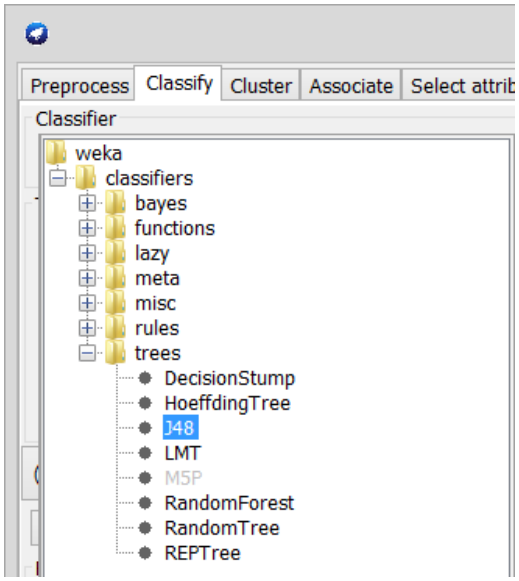
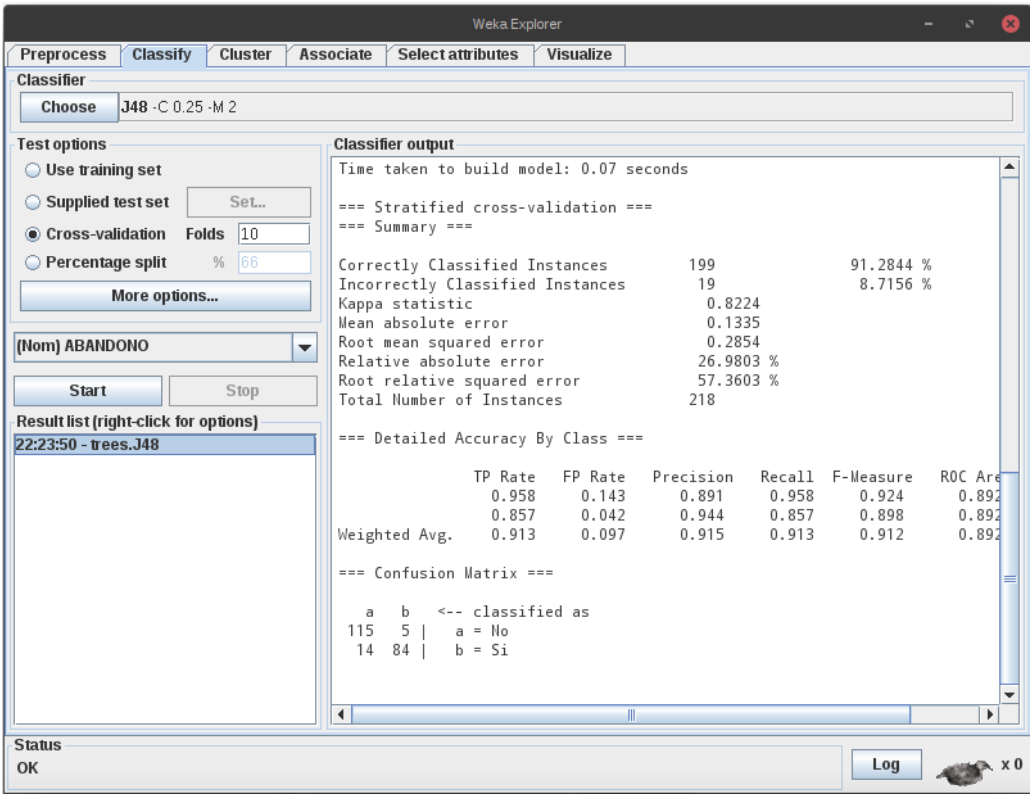


Ilustración 16: Selección del algoritmo J48 en Weka

Una vez seleccionado el algoritmo J48, se carga el archivo de Alumnos como se aprecia en la Figura.



El resultado de salida de la ejecución del algoritmo se muestra a continuación:

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Alumno_detalle

Instances: 218

Attributes: 10

CODIGO

Ilustración 17: Ejecución del clasificador J48 con los datos Académicos

ORDEN_MERITO_INGRESO

PUNTAJE_INGRESO

MODALIDAD_INGRESO

TOTAL_CURSOS

CURSOS_APROBADOS

CURSOS_DESAPROBADOS

CREDITOS_APROBADOS

PROMEDIO

ABANDONO

Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

CREDITOS_APROBADOS <= 32

| CODIGO <= 1644402053: Si (63.0/1.0)

| CODIGO > 1644402053

| | CURSOS_APROBADOS <= 0: Si (19.0)

| | CURSOS_APROBADOS > 0

| | | CODIGO <= 1694403059

| | | | TOTAL_CURSOS <= 24: Si (5.0)

| | | | TOTAL_CURSOS > 24: No (4.0/1.0)

| | | CODIGO > 1694403059: No (19.0)

CREDITOS_APROBADOS > 32: No (108.0/11.0)

Number of Leaves : 6

Size of the tree : 11

Time taken to build model: 0.07 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	199	91.2844 %
Incorrectly Classified Instances	19	8.7156 %
Kappa statistic	0.8224	
Mean absolute error	0.1335	
Root mean squared error	0.2854	
Relative absolute error	26.9803 %	
Root relative squared error	57.3603 %	
Total Number of Instances	218	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.958	0.143	0.891	0.958	0.924	0.892	No
	0.857	0.042	0.944	0.857	0.898	0.892	Si
Weighted Avg.	0.913	0.097	0.915	0.913	0.912	0.892	

=== Confusion Matrix ===

a	b	<-- classified as
115	5	a = No
14	84	b = Si

La matriz de confusión obtenida con el algoritmo J48:

Tabla 2: Matriz de confusión obtenido por el algoritmo J48

Instancias clasificadas como Abandona	Instancias clasificadas como No Abandona
115	5
14	84

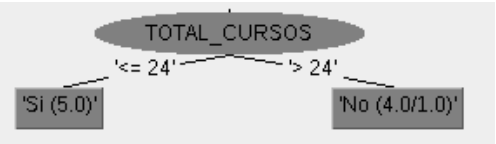


Ilustración 18: Arbol creado con el algoritmo J48

Las gráficas obtenidas muestran las diferentes relaciones del atributo Abandono con los demás atributos considerados: Modalidad, Puntaje de ingreso, Aprobados, Desaprobados, Total de cursos, y Promedio _final.

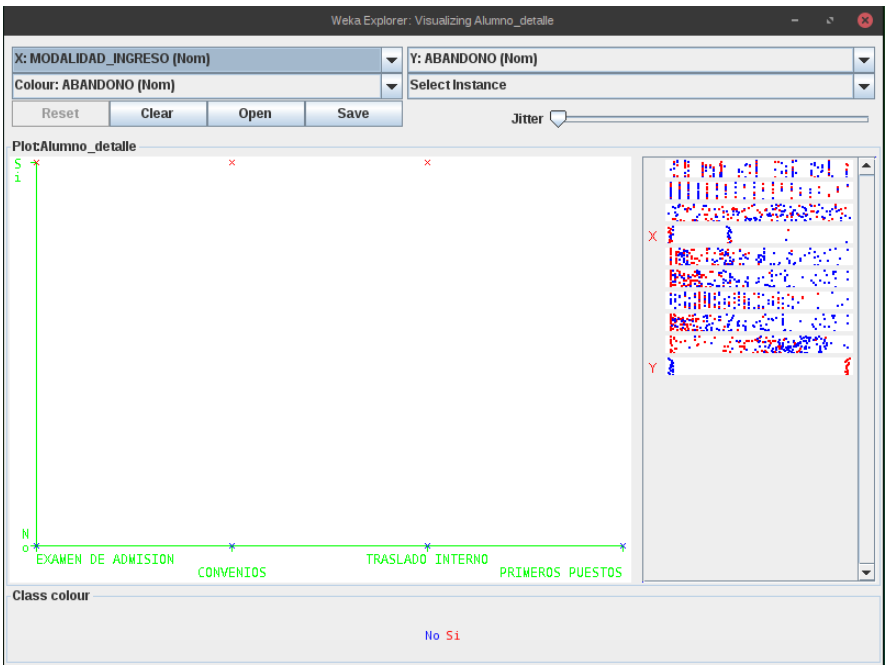


Ilustración 19: Abandono de estudiantes en relación al Total de cursos

4.2 PRESENTACION, ANALISIS E INTERPRETACIÓN DE RESULTADOS

4.2.1 PRESENTACIÓN DE RESULTADOS EN EL SPSS

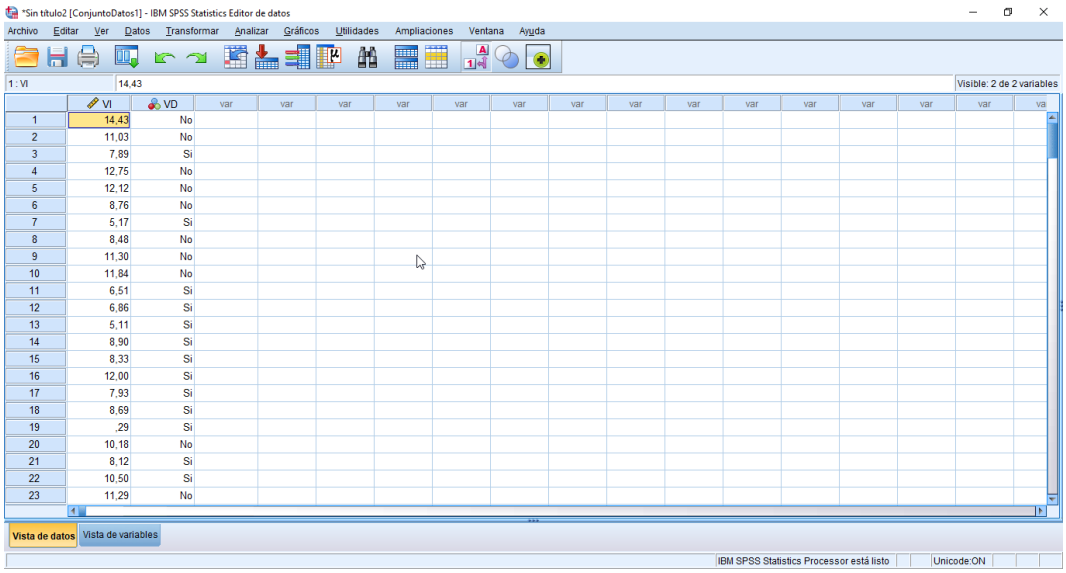


Ilustración 20: Ingreso de los datos al SPSS

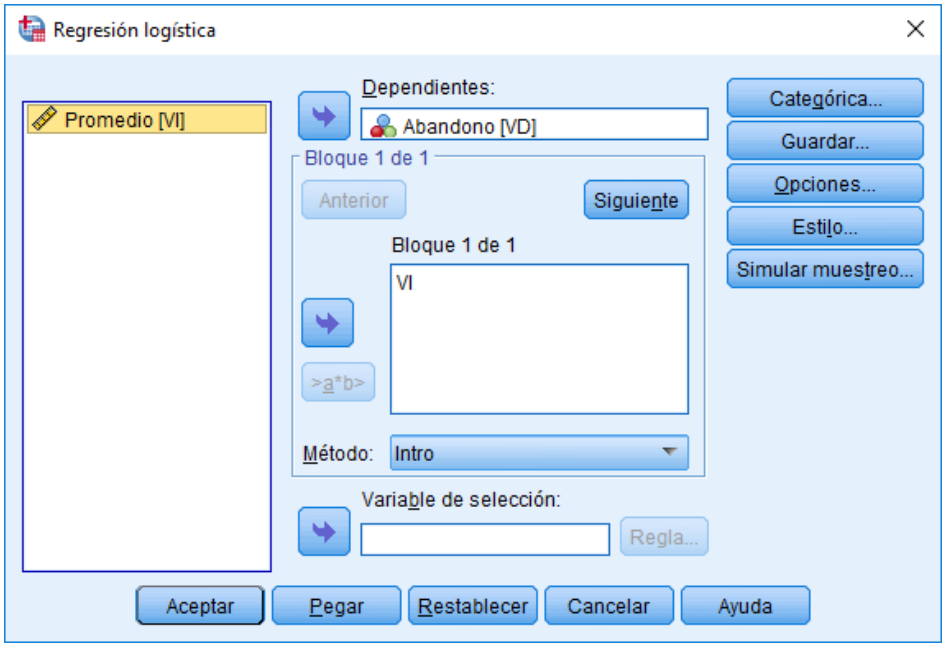


Ilustración 21: Aplicación de la Regresión Logística Binaria

4.3 PRUEBA DE HIPOTESIS

Hipótesis general

La deserción estudiantil puede predecirse mediante técnicas de minería de datos de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.

4.3.1 Prueba de Hipótesis en el SPSS

			Puntuación	gl	Sig.
Paso 0	Variables	Promedio	73,807	1	,000
	Estadísticos globales		73,807	1	,000

4.4 DISCUSION DE RESULTADOS

Existe una diferencia significativa entre la deserción estudiantil y promedio de notas, a un nivel de confianza del 95% y nivel de significancia del 5%. Por tanto, se concluye que **promedio de notas influye significativamente en la deserción estudiantil** de los alumnos de la escuela de formación profesional de Sistemas y Computación de la UNDAC.

CONCLUSIONES

Las diversas técnicas supervisadas de minería de datos se pueden aplicar de manera efectiva en los datos educativos prediciendo la deserción académica.

Del análisis de las técnicas de clasificación se puede aplicar en los datos educativos para predecir el resultado de abandono o no del estudiante a los estudios y conocer la cantidad tentativa de grupos o secciones que se pueden asignar para su planificación.

Las pruebas realizadas en el caso de estudio con en la carrera profesional de Ingeniería de Sistemas de la Universidad Nacional Daniel Alcides carrión, en cuanto a la relación de atributos Abandono y Categoría: demuestran que la mayor cantidad de estudiantes que abandonan la carrera es por las bajas notas. La relación de los atributos Abandono y cursos Aprobados demuestran que la mayor cantidad de estudiantes que abandonan son los han aprobado menos de 24 cursos Aprobados.

La eficiencia de los algoritmos de árboles de decisión C4.5 (J48) y RandomTree se puede analizar en función de su precisión comprobando que ambos algoritmos obtienen los mismos resultados.

En cuanto al tiempo tomado para derivar el árbol, se observa que el algoritmo RandomTree es ligeramente más veloz que el algoritmo J48 en la construcción del árbol de decisión.

RECOMENDACIONES

Ampliar otros atributos determinantes en el Sistema Académico de la UNDAC como los aspectos psicológicos y otros para realizar la una mejor predicción y obtener resultados más afinados.

Realizar la aplicación de otros algoritmos de clasificación para comparar los resultados y comprobar su eficiencia.

Utilizar predictores con diferentes variables de estudio, así como en otras muestras temporales para observar si se encuentran resultados similares a los de la presente muestra.

Incorporar criterios difusos a los atributos de tal forma que pueda considerar cierto grado de verdad.

BIBLIOGRAFÍA

Becerrra, O. (2012). *Guia para elaboración de instrumentos.* Obtenido de http://nticsaplicadasalainvestigacion.wikispaces.com/file/view/guia+para+el aboracion+de+instrumentos.pdf

E. Evans, E. (2013). Los entornos personales de aprendizaje en el marco de la educación permanente. *EDMETIC Educación Mediatica y TIC*, 94-110. doi:https://doi.org/10.21071/edmetic.v2i1

Hernández, R., Fernández, C., & Baptista, P. (2010). *Metodología de la Investigación.* México: McGrawHill.

Piscoya Ordonez, L. (2016). *Aplicación de técnicas de minería de datos para predecir la deserción estudiantil en la educación básica regular en la región de Lambayeque.* Pimentel.

SHAPIRO, S., & WILK, M. (1 de December de 1965). An analysis of variance test for normality (complete samples). *Biometrika*, 591-611. doi:https://doi.org/10.1093/biomet/52.3-4.591

Spositto, O., Etcheverry, M., Ryckeboer, H., & Bossero, J. (2010). Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. San Justo, Buenos Aires, Argentina.

Sulla Torres, J. (2015). *Aplicación de técnicas supervisadas de minería de datos para determinar la predicción de deserción académica.* Arequipa.

Valero, S., Salvador, A., & Garcia, M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Santiago Mihuacan, Puebla, Mexico.

TABLA DE ILUSTRACIONES

ANEXOS

MATRIZ DE CONSISTENCIA "APLICACIÓN DE TÉCNICAS DE MINERIA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN"

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES	INDICADORES	TECNICAS E INSTRUMENTOS	METODOLOGIA
GENERAL: ¿Cómo puede predecirse la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión?	**GENERAL:** Predecir la deserción estudiantil mediante técnicas de minería de datos de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.	**GENERAL:** La deserción estudiantil puede predecirse mediante técnicas de minería de datos de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.	**INDEPENDIENTE:** Técnicas de minería de datos.	Técnicas	- Información Histórica. - Análisis documental. - Observación.	**Tipo:** Aplicada **Nivel:** Predictivo **Método:** Deductivo **Diseño:** transeccional del tipo correlacional causal. **POBLACIÓN Y MUESTRA** **Población:** Constituida por 1810 alumnos matriculados en el periodo 2017-A de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión. **Muestra:** No aleatoria por selección intencionada o muestreo de conveniencia, constituida por 218 alumnos de Escuela Profesional de Sistemas y Computación de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.
ESPECÍFICO: ¿Qué técnicas de minería de datos puede predecir la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión?	**ESPECÍFICO:** Aplicar técnicas de minería de datos para predecir la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.	**ESPECIFICO** Las técnicas de minería de datos pueden aplicarse para predecir la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión.	**DEPENDIENTE:** Deserción estudiantil.	Predicción		