



UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA
FACULTAD DE INGENIERÍA
SECRETARIA DE INVESTIGACIÓN Y POSGRADO
INGENIERÍA EN COMPUTACIÓN

DESARROLLO DE SISTEMAS DE PREDICCIÓN CON
REGRESIÓN DEL DESEMPEÑO ACADÉMICO DE ALUMNOS DE
NUEVO INGRESO EN UNIVERSIDADES TECNOLÓGICAS

PRESENTA

JOSÉ ALFREDO DE LA CRUZ NAVA

CHIHUAHUA, CHIH.

Diciembre/2016



UNIVERSIDAD AUTÓNOMA DE CHIHUAHUA
FACULTAD DE INGENIERÍA
INGENIERÍA EN COMPUTACIÓN

DESARROLLO DE SISTEMAS DE PREDICCIÓN CON
REGRESIÓN DEL DESEMPEÑO ACADÉMICO DE ALUMNOS DE
NUEVO INGRESO EN UNIVERSIDADES TECNOLÓGICAS

APROBADO:

Dr. Ricardo Blanco Vega, Director
Asesora: Dra. Alejandra Castro Cuesta.

M.S.I & M.A Manuel Gardea Meraz sinodal

M.I Ricardo Torres Knight sinodal

CHIHUAHUA, CHIH.

15/12/16

Derechos reservados
© José Alfredo De La Cruz Nava
Calle Lauro Gomez No. 13, Col. Misael Nuñez,
Chihuahua, Chihuahua
2016

Copyright ©

por

José Alfredo De La Cruz Nava

2016

6 de diciembre de 2016

ING. JOSÉ ALFREDO DE LA CRUZ NAVA

Presente

En atención a su solicitud relativa al trabajo de tesis para obtener el grado de Maestro en Ingeniería en Computación, nos es grato transcribirle el tema aprobado por esta Dirección, propuesto y dirigido por el director Dr. Ricardo Blanco Vega para que lo desarrolle como tesis, con el título: **"Desarrollo de sistemas de predicción con regresión del desempeño académico de alumnos de nuevo ingreso en universidades tecnológicas"**.

ÍNDICE

Agradecimientos

Resumen

Índice de Tablas

Índice de Figuras

Capítulo 1. Introducción

- 1.1 Métricas de desempeño
- 1.2 Herramientas y datos históricos
- 1.3 Trabajos relacionados
- 1.4 Objetivos

Capítulo 2. Justificación

Capítulo 3. Fundamentación teórica

- 3.1 Bases de datos
- 3.2 Minería de datos
- 3.3 Metodología KDD
- 3.4 Reducción de la dimensionalidad

Capítulo 4. Validación experimental

- 4.1 Entendimiento del negocio
- 4.2 Pre procesamiento y transformación de los datos
- 4.3 Actividades de minería de datos
- 4.4 Resultados

Facultad de Ingeniería
Circuito No.1, Campus Universitario 2
Chihuahua, Chih. C.P. 31125
Tel. (614) 442-85-00 www.fing.uach.mx





Capítulo 5. Conclusiones

5.1 Conclusiones del objetivo 1

5.2 Conclusiones del objetivo 2

5.3 Conclusiones del objetivo 3

Bibliografía

Solicitamos a Usted tomar nota de que el título del trabajo se imprima en lugar visible de los ejemplares de las tesis.

ATENTAMENTE
"naturam subiecit aliis"

EL DIRECTOR

M.I. JAVIER GONZALEZ CANTÚ

FACULTAD DE
INGENIERÍA
UACH



DIRECCIÓN

EL SECRETARIO DE INVESTIGACIÓN
Y POSGRADO

DR. FERNANDO RAFAEL ASTORGA
BUSTILLOS

Facultad de Ingeniería

Circuito No.1, Campus Universitario 2

Chihuahua, Chih. C.P. 31125

Tel. (614) 442-95-00 www.fing.uach.mx



Dedicatoria

Para mis padres y hermanos que siempre me apoyaron en mi educación dando lo mejor de sí para mejorar mi futuro y creyeron en mí, a mi director de tesis que me apoyó en lo que necesité para la elaboración de este trabajo compartiendo sus conocimientos.

Agradecimientos

Agradezco primeramente a Dios por permitirme llegar a estos momentos de desarrollo profesional, por darme la fuerza y las habilidades para terminar esta tesis.

A mi familia por siempre apoyarme en mis nuevas metas, alentándome a conseguirlas y siempre tener fe en que yo lo podía conseguir.

También agradezco a mis sinodales que con paciencia y vocación me orientaron a desarrollar este trabajo.

A mi asesor de tesis por sus consejos y motivaciones para seguir adelante en cada momento de la carrera.

Resumen

En México es de gran importancia la educación de nivel superior ya que se necesitan profesionistas para alcanzar un mayor desarrollo, sin embargo no todos los estudiantes logran terminar su carrera. Se han realizado sistemas de predicción para pronosticar deserción o bajo rendimiento en los alumnos en general, pero no se identifican a los alumnos de nuevo ingreso con bajo rendimiento para otorgarles atención personalizada, y de esta manera ayudarles a terminar su carrera, es por eso que mediante minería de datos se realizará un sistema de predicción basado en algoritmos de regresión que arroje información sobre los estudiantes de nuevo ingreso de la Universidad Tecnológica de Chihuahua con mayor probabilidad de fracasar en su primer cuatrimestre. Se buscará el mejor algoritmo a partir de datos recabados en la Universidad, basándose en conceptos de bases de datos y minería de datos. Las actividades que se realizarán en este proyecto tendrán una metodología propia, basándose inicialmente en la metodología KDD (*Knowledge Discovery in Databases*), cada una de estas actividades están planificadas en los dos años de duración de la maestría.



Índice de Contenido

Agradecimientos	x
Resumen.....	xii
Índice de Tablas	xiv
Índice de Figuras	xv
Capítulo 1: Introducción	1
1.1 Métricas de desempeño	1
1.2 Herramientas y datos históricos.....	2
1.3 Trabajos relacionados.....	4
1.4 Objetivos	6
Capítulo 2: Justificación.....	7
Capítulo 3: Fundamentación Teórica	8
3.1 Bases de datos	8
3.2 Minería de datos	10
3.3 Metodología KDD.....	15
3.4 Reducción de la dimensionalidad.....	16
Capítulo 4: Validación experimental.....	17
4.1 Entendimiento del negocio.....	17
4.2 Pre procesamiento y transformación de los datos	18
4.3 Actividades de minería de datos.....	20
4.4 Resultados	23
Capítulo 5: Conclusiones	38



5.1	Conclusiones del objetivo 1	39
5.2	Conclusiones del objetivo 2	40
5.3	Conclusiones del objetivo 3	41
Bibliografía		42



Índice de Tablas

Tabla 4 1: Tabla comparativa de cada modelo de regresión	28
--	----

Índice de Figuras

Figura 3 1: Ejemplo del modelo entidad relación	9
Figura 3 2: Ejemplo del modelo relacional	10
Figura 3 3: Ejemplo de ajuste lineal en el modelo de regresión	13
Figura 3 4: Ejemplo de ajuste lineal en el modelo de regresión cuadrática.....	13
Figura 3 5: Ejemplo de ajuste lineal en el modelo de regresión logarítmica	14
Figura 3 6: Ejemplo de ajuste lineal en el modelo de regresión exponencial.....	14
Figura 4 1: Vista de los atributos del conjunto de datos en Weka.....	18
Figura 4 2: Forma para abrir el conjunto de datos del CENEVAL	19
Figura 4 3: Conjunto de datos del EXANI.....	19
Figura 4 4: Actividades del objetivo 1	20
Figura 4 5: Actividades del objetivo 2	21
Figura 4 6: Actividades del objetivo 3	22
Figura 4 7: Ejemplo de entendimiento de los datos utilizando datos del EXANI de Abril del 2015 en Weka	23
Figura 4 8: Archivo completo en ARFF de promedios con EXANI.....	24
Figura 4 9: Archivo ARFF de trayectorias con EXANI, mostrando algunos de los campos...	25
Figura 4 10: Lista de atributos con su respectivo coeficiente de correlación y la selección del mejor del conjunto con promedio y EXANI.	26
Figura 4 11: Lista de atributos con su respectivo coeficiente de correlación y la selección del mejor del conjunto con trayectorias y EXANI.....	27
Figura 4 12: Modelos de regresión con su coeficiente de determinación promedios con EXANI.....	28
Figura 4 13: Modelos de regresión con su coeficiente de determinación trayectorias con EXANI.....	29



Figura 4 14: Atributos ordenados con base a su coeficiente de correlación	30
Figura 4 15: Atributos ordenados con base a su coeficiente de correlación	31
Figura 4 16: Explicación del mejor atributo y la clase	32
Figura 4 17: Primera vista del sistema de predicción.....	34
Figura 4 18: Metodología propia.....	37
Figura 5 1: Método Delphi.....	40

Capítulo 1: Introducción

En los últimos años, en diferentes partes del mundo la educación se ha convertido en algo muy influyente en cuanto al desarrollo de un país, Un ejemplo de ello está en Chihuahua México donde se sigue trabajando para mejorar los índices de educación en todos los niveles de educación, ya que es de suma importancia de que cada año mejoren (SEP, 2015) Sin embargo gran parte de las personas que comienzan sus estudios en una universidad no logran terminar su carrera. En América Latina menos del 10% de los adultos que tienen 25 años han logrado terminar su educación universitaria (Fiegehen, 2005). Existen varios factores que afectan a que la deserción en universidades sea tan alta. Uno de estos factores tiene que ver con la crisis personal de los jóvenes que entran a realizar sus estudios, las cuales se atribuyen a sus problemas académicos, esta crisis se puede desarrollar durante su estancia en la carrera y puede requerir ayuda especializada, también algunos estudiantes simplemente pierden interés en la escuela al darse cuenta que están fallando en sus exámenes, estos estudiantes desarrollan problemas de depresión, otros desarrollan problemas de estrés o crisis en sus actividades diarias, pero el factor que del que más se tienen datos históricos en las instituciones y el más recomendado para trabajar es con respecto la desempeño académico.

1.1 Métricas de desempeño

En la actualidad muchas universidades aplican diferentes exámenes para conocer el nivel académico de los futuros aspirantes, uno de estos exámenes es el psicométrico, el cual ayuda a medir el nivel de éxito de un estudiante durante su estadía en su carrera, se mide con psicología científica, que depende mucho de la lógica, en otras instituciones se realizan entrevistas y exámenes de conocimiento. El examen más común en México para la entrada a las instituciones de nivel superior es el EXANI (Examen Nacional de Ingreso), el cual

tiene como objetivo otorgar información de los estudiantes con mayores probabilidades de éxito en la educación de nivel superior (CENEVAL, 2014).

1.2 Herramientas y datos históricos

La información obtenida de cada estudiante que ingresa a las instituciones de nivel superior va generando datos históricos, los cuales pueden ser útiles para resolver problemas a futuro. La información que se va recopilando en la institución se almacena en una base de datos que se define como el conjunto de información estructurada en registros y grabada en un soporte técnico legible por un ordenador (Yunta, 2001). Dicha información es útil cuando necesitamos predecir ciertos eventos que podemos aprovechar para resolver un problema.

Para predecir el desempeño de un alumno se necesitan utilizar ciertas características asociadas a él que lo describen y definen. También es necesario contar con herramientas que son de utilidad para buscar y analizar datos que se han ido recolectando. Es por eso que se aplica la minería de datos, la cual se define como un campo en la ciencia de la computación referido al proceso de buscar patrones en grandes volúmenes de conjuntos de datos (Rokach, 2010).

La institución educativa donde se realizará este proyecto tiene que haber visualizado este problema, y también tiene que contar con datos históricos que ayuden a una toma de decisiones que puedan llegar a la solución del problema mencionado, este proceso es algo común cuando hablamos de minería de datos (Moxons, 1996), ya que su proceso consiste en descubrir y extraer la información, los resultados son arrojados mediante programas simples que dan una predicción con una buena probabilidad de éxito (Moxons, 1996), otro punto importante por el cual se utilizará minería de datos es por la facilidad que se tiene para encontrar patrones o tendencias, todo esto localizado en datos históricos (Arenas, 1996).

Un sistema de predicción es un punto importante en la ingeniería de sistemas, debido a que puede indicar una trayectoria esperada de hechos para poder adelantarnos a tomar una decisión adecuada (Yi-sheng, Chao, & Zhi-bin, 2010). Para realizar este tipo de sistemas se pueden elegir diferentes medios o caminos para poder llevarlo a cabo, en este trabajo se utilizará el modelo de regresión lineal, el cual consiste en tomar diferentes variables de resultados anteriores y encontrar las variables dependientes para poder aplicarlas a una función (Devore, 2009). Este modelo se utiliza con diversas transformaciones, entre las cuales encontramos regresión cuadrática, exponencial, logarítmica y potencial, las cuales se utilizarán para encontrar mejores resultados.

La regresión puede dividirse en regresión simple y regresión múltiple, la primera es una regresión lineal que ajusta la ecuación de la recta solo de una variable independiente “ X_1 ” a la variable dependiente “ Y ”.

$$Y=B_0+B_1X_1$$

Mientras que la regresión lineal múltiple obtiene una relación entre un conjunto de variables independientes “ X_n ”, con una variable dependiente “ Y ”.

$$Y=B_0+B_1X_1+\dots+B_nX_n.$$

Una derivación es el sistema lineal jerárquico, el cual consiste en ir tomando los datos que se van presentando de manera jerárquica (Luna, 2007).

La regresión lineal utiliza diferentes transformaciones para poder generar modelos con mayor exactitud, entre las transformaciones utilizadas en este documento se encuentran la cuadrática, exponencial y logarítmica, con las cuales se puede realizar una comparación para encontrar el mejor modelo.

1.3 Trabajos relacionados

En Brasil se diseñó un sistema de redes neuronales para predecir la posible deserción de un grupo de estudiantes de sus clases (Valquíria R. C. Martinho, 2013), sabemos con esta información que si es posible predecir el comportamiento de un alumno teniendo los datos que han sido recolectados con anterioridad, sin embargo el trabajo se realizó a grupos de alumnos y no en toda la escuela, esto mediante respuestas en cuestionarios proporcionados por el profesor donde se evaluaba el desempeño y datos socioeconómicos. A diferencia de ese trabajo el alcance de este proyecto es basándose en datos históricos ya recolectados para predecir su calificación durante el primer cuatrimestre de la universidad, comparando algoritmos de regresión con una muy alta precisión.

En la Universidad Nacional de Noroeste en Argentina se desarrolló un sistema de predicción, los datos fueron tomados de reportes que realizaban los profesores sobre sus alumnos, estos reportes contenían información sobre el rendimiento de los alumnos en las clases, de esta manera se diseñó un sistema de regresión lineal que arrojó resultados para medir el posible rendimiento de los alumnos que entraban a los primeros semestres de esta Universidad (Porcel, Dapozo, & Victoria, 2009). Los resultados en este trabajo fueron válidos y aceptados, ofreciendo una gran ayuda a la toma de decisiones de esta Universidad. Otro sistema de regresión lineal, con el mismo objetivo, pero en este caso de forma jerárquica se realizó en España, donde se tomaban datos a partir de las aptitudes de los estudiantes, auto concepto académico y los atributos casuales, como el comportamiento y su manera de ser en la escuela. Para capturar estos datos se aplicaron exámenes por parte de colaboradores, aquí también se aplicó la regresión múltiple para poder tener mayor precisión, los resultados fueron relacionados a las aptitudes y forma de pensar de los estudiantes, aquellos que se fijan objetivos y cuentan con buen rendimiento en estudios anteriores, son los que logran terminar los cursos (Miñano Pérez, Cantero Vicente, & Castejón Costa, 2008).

En el estado de Chihuahua se realizó un estudio sobre el perfil que define al alumno en su rendimiento escolar. Realizado en el Instituto Tecnológico de Chihuahua II se dio a conocer información a partir de una muestra de alumnos sobre las características que tienen los diferentes perfiles de rendimiento, tomando en cuenta factores como la escolaridad de los padres, las edades, las personas con las que vive, el semestre, las materias que lleva reprobadas entre otros (Chavez, Ibarra, De La Garza, & Blanco, 2012).

En el año 2014 en la región de Tuzla, Bosnia, se seleccionó un grupo de estudiantes del año 2011 a 2012 para poder predecir los casos de éxito con diferentes algoritmos de minería de datos como árboles de decisión, Naïve Bayes y multilayer perceptron, teniendo buenos resultados. (OSMANBEGOVIĆ, AGIĆ, & SULJIĆ, 2014).

En Cuba se desarrolló un sistema basado en regresión logística ordinal para detectar el posible bajo rendimiento escolar de los alumnos en las materias difíciles, para que el profesor pudiera enfocarse en los alumnos que necesiten más ayuda, aplicando diversos métodos de enseñanza. (Heredia, Rodriguez, & Villalta, 2014)

El enfoque que se tiene en este trabajo de investigación basado en sistemas de predicción, a diferencia de los proyectos mencionados anteriormente, es que el análisis se realiza por medio de un método de regresión lineal simple con transformación en el dato más importante maximizando su relación. En este caso se tendrán como entradas datos del resultado en el examen de admisión a la institución y el examen psicométrico. Los datos recabados se analizarán para determinar un posible bajo rendimiento que el alumno pueda llegar a tener a lo largo de su estancia en la escuela. No se han encontrado trabajos iguales a lo que se presenta en este proyecto, pero sí con cierta similitud como se vio anteriormente.

1.4 Objetivos

Desarrollar un sistema de predicción basado en algoritmos de regresión. El modelo de regresión tiene un coeficiente de determinación superior o igual al 90%. El sistema identifica a los alumnos de nuevo ingreso que podrían llegar a tener bajo rendimiento escolar. El caso de estudio son los alumnos de primer cuatrimestre de educación superior de la Universidad Tecnológica de Chihuahua..

1.4.1 Objetivos específicos

1. Búsqueda del mejor algoritmo para crear modelos de regresión de alta calidad ($R^2 \geq 90\%$) para estimar el desempeño de cada estudiante de nuevo ingreso.
2. Validar la utilidad y la novedad de los modelos aprendidos utilizando el juicio de expertos, buscando la adaptación y mejora.
3. Crear una metodología propia para el desarrollo de este tipo de proyectos. Se basa en el proceso KDD.



Capítulo 2: Justificación

Uno de los principales objetivos de las Instituciones de Educación Superior es lograr formar profesionales que contribuyan al desarrollo del país en el cual van a ejercer.

En México es de gran importancia que se tengan egresados preparados para enfrentar los retos de su profesión debido a que es un país en desarrollo, es por eso que la deserción escolar o el bajo rendimiento del alumnado de una Institución de Educación Superior es muy preocupante al momento en que la institución tiene que mostrar resultados, es por eso que se necesita identificar a los alumnos de nuevo ingreso que podrían tener un bajo rendimiento durante su estancia en la escuela.

Este proyecto ayudará a la Universidad Tecnológica de Chihuahua a identificar aquellos alumnos que posiblemente tendrán un bajo rendimiento escolar y les otorgará una atención más personalizada buscando disminuir el problema de deserción escolar. Debido a que se implementará un sistema de predicción por regresión, Esto se logrará implementando un sistema de regresión basado en datos históricos que serán proporcionados por la institución. Anteriormente ya se han realizado proyectos similares en otras Instituciones sobre deserción del alumnado, sin embargo este proyecto apoyará en la toma de decisiones dentro de la institución para planes futuros en el proceso de ingreso al primer cuatrimestre..

Capítulo 3: Fundamentación Teórica

Para la realización de este proyecto se tomarán diferentes conceptos de los cuales se obtiene el conocimiento adecuado para poder ser llevado a cabo, conceptos de bases de datos con sus respectivos programas, minería de datos, regresión lineal, análisis de los sistemas de predicción, conceptos de programación, entre otros.

3.1 Bases de datos

Las bases de datos pueden ser comparadas a una biblioteca, la cual tiene un conjunto de datos que están acomodados en un orden apropiado, de esta manera cuando se requiere acceder a los datos solamente se selecciona el tipo que queremos. Una base de datos es precisamente un conjunto de datos estructurados para que podamos manipularlos a nuestra conveniencia y poder hacer diferentes operaciones como ingresar datos, borrar, consultar entre otras. (Silberschatz, Korth, & Sudarshan, 2002).

3.1.1 *Programas gestores de bases de datos*

Existen diferentes programas que ayudan a realizar la manipulación de las bases de datos, los más comunes son Oracle, Microsoft SQL Server y MySQL que ayudan a realizar bases de datos de forma gráfica y conceptual, separando los datos en diferentes tablas para su almacenamiento, a pesar de que las instrucciones son similares cada uno tiene ventajas y desventajas, por ejemplo MySQL se distribuye de forma gratuita, sin embargo Oracle es más distribuido debido a su facilidad.

3.1.2 *Tipos de modelos de bases de datos*

Para el manejo de la estructura de los datos se utilizan los modelos de datos, los cuales son herramientas que sirven de ayuda para poder describir los datos, aquí se describen dos tipos de modelos, el modelo entidad relación y el modelo relacional.

3.1.2.1 *Modelo entidad relación*

Una manera de representar los datos es mediante un modelo conceptual donde se enlazan una cierta cantidad de atributos que caracterizan a una entidad, por lo tanto en una

base de datos se puede estructurar diferentes entidades, todas las entidades que tengamos representadas se van a relacionar con una asociación que tenga cada par de identidades. En este caso para representarlo gráficamente se utilizan recuadros para mostrar las identidades con sus respectivos campos o atributos en rectángulos adjuntos, , para mostrar las relaciones se utilizan líneas con un conector que representan el tipo de relación con la llave maestra o llaves foraneas como se ve en la figura 3.1, la cual está basada en un diagrama indentidad relación de SQL WORKBENCH en la cual se puede ver la representación de relaciones entre las tablas de ventas, detalle de ventas, clientes y productos, , se realiza una relación entre cada una de las entidades para lograr realizar la operación, se pueden observar los ejemplos de cómo utilizar cada uno de los tipos de relaciones y su representación, hay relaciones de uno a uno, y de uno a muchos. (Silberschatz, Korth, & Sudarshan, 2002).

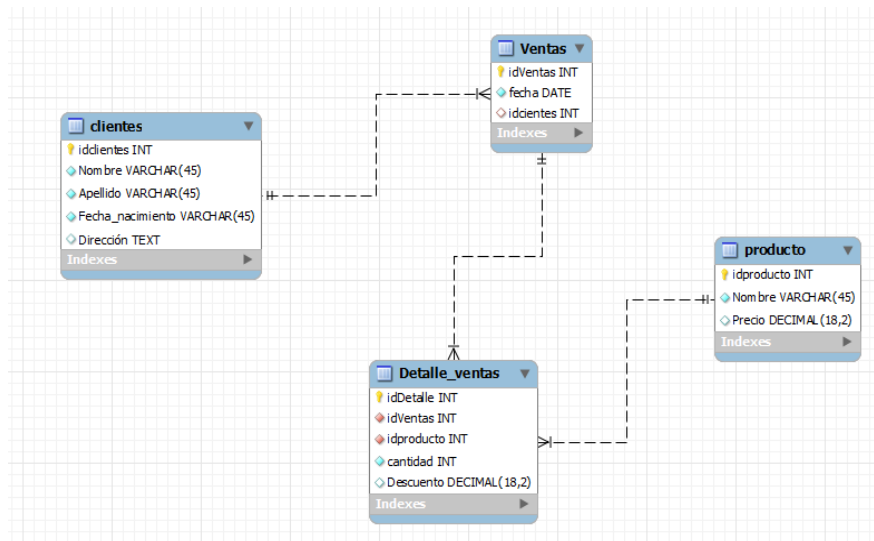


Figura 3 1: Ejemplo del modelo entidad relación

3.1.2.2 Modelo relacional

En este modelo se encuentran los datos estructurados en tablas, estas tablas también representan la relación entre ellos, tienen diferentes columnas las cuales tienen un nombre único, como se ve en la figura 3.2, se puede observar que cada tabla tiene diferentes atributos que tienen relación con otras tablas, es por eso que se puede ver señalizada con una flecha la relación. Este es un modelo basado en registros, donde cada uno de los identificadores que contienen tiene un grado relacional, es decir una interacción jerárquica conforme en los atributos que son más usados, con otros tipos de tablas que se encuentren dentro de las bases de datos (Silberschatz, Korth, & Sudarshan, 2002).

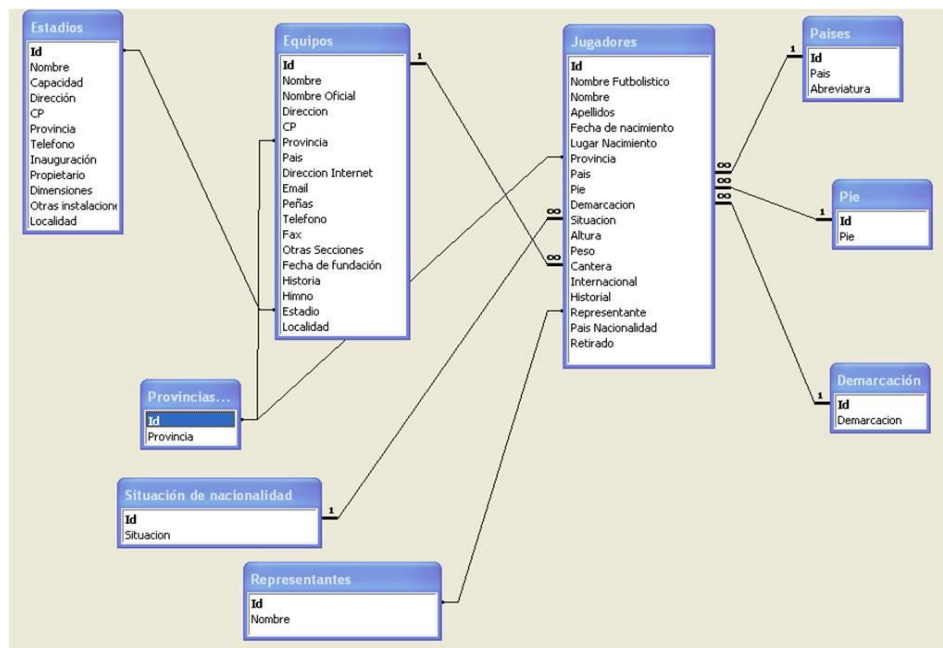


Figura 3 2: Ejemplo del modelo relacional

3.2 Minería de datos

Se utilizarán técnicas basadas en la minería de datos para darle solución al problema planteado sobre el bajo desempeño académico en los alumnos del primer cuatrimestre La minería de datos se define como el proceso donde se pueden descubrir, extraer y almacenar información guardada en diferentes bases de datos, para esto se utilizan diferentes

aplicaciones programadas exclusivamente para realizar todas estas operaciones, de esta manera se aprovecha todo el valor de la información que es obtenida de la institución (Moxons, 1996).

Uno de los retos más grandes de la minería de datos es trabajar con grandes volúmenes de datos de los cuales se pueda obtener información importante sobre el estudio o la tarea que se va a realizar, por lo tanto la minería de datos consiste en el proceso de poder extraer información útil a partir de datos históricos. (Orallo, 2007)

3.2.1 WEKA

Desarrollado en Java por la universidad de Waikato, el programa WEKA es usado para la minería de datos, ya que contiene diferentes herramientas para visualización y algoritmos que nos ayudan a realizar un modelo predictivo. Contiene variadas técnicas de predicción, su uso es fácil y tiene una licencia pública general de GNU (Witten & Eibe Frank, 1999). El ambiente y librerías de WEKA se utilizan en la realización de éste proyecto.

3.2.2 Técnicas de predicción

En la minería de datos se utilizan diferentes tareas para poder lograr una predicción, de las cuales las más importantes son de clasificación y regresión. En clasificación se puede tomar un conjunto de valores de los ejemplos que se presentan por medio de una función, la función aprendida será capaz de determinar la clase para cada nuevo ejemplo sin clasificar. En regresión hacemos algo similar que en clasificación, sin embargo los valores que obtenemos y con los que trabajamos son valores numéricos y enteros. Las dos principales tareas presentadas anteriormente tienen diferentes métodos con los cuales podemos realizar el sistema de predicción (Orallo, 2007).

Las técnicas basadas en redes neuronales artificiales aprenden un modelo mediante diferentes pasos o experiencias de nodos o neuronas, tienen diferentes topologías, la más conocida es retro propagación.

Los modelos basados en núcleo y máquinas de soporte vectorial intentan maximizar el margen entre todos los tipos de clases que se tienen formadas, utilizan transformaciones que llegan a aumentar la dimensión.

Existen otras técnicas de predicción que son híbridas entre las anteriores, sin embargo las aquí presentadas son las de mayor importancia (Orallo, 2007).

3.2.3 Modelo de regresión lineal

En este trabajo se utilizarán las técnicas de predicción conforme al modelo de regresión la cual toma valores cuantitativos, la regresión se puede realizar de diferentes formas, pero la más conocida y más utilizada es la regresión lineal, en la cual se toman diferentes valores o conjuntos de datos que siguen una línea que se ajusta conforme a los valores de respuesta, por ejemplo un conjunto de valores determinados (k), que toman un valor para "X" X_k ($k = 1, \dots, K$), por lo tanto la regresión lineal es explicada como la participación de múltiples variables que van participando de forma aditiva y constante para todo el dominio observado, se puede simplificar a $Y = \beta_1 + \beta_2 X$ (Orallo, 2007), en la figura 3.3 se pueden observar los puntos que representan los valores obtenidos en "x" procedimientos y como se distribuyen conforme a la recta de regresión.

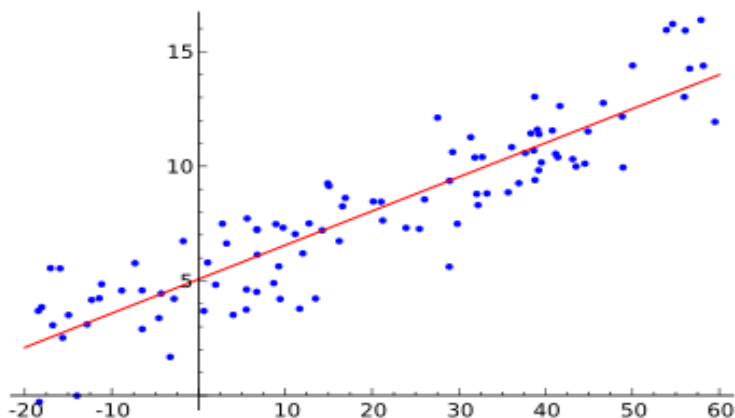


Figura 3 3: Ejemplo de ajuste lineal en el modelo de regresión.

3.2.4 Transformaciones de regresión lineal

En una transformación que se aplica a las variables de ecuación lineal se pueden obtener diferentes formas por medio de fórmulas para que la línea se ajuste a los datos. Algunos ejemplos son:

- Cuadrática $y = A + Bx^2 + Cx + z$

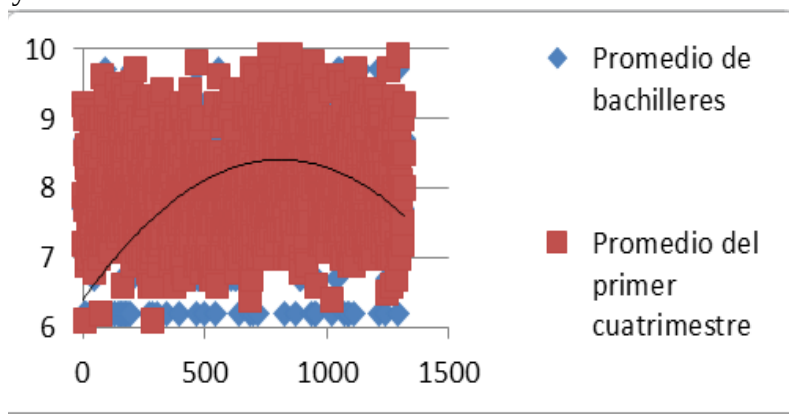


Figura 3 4: Ejemplo de ajuste lineal en el modelo de regresión cuadrática utilizando el conjunto de datos del promedio de los alumnos en el primer cuatrimestre con respecto a su promedio de bachilleres

- Logarítmica $y = \ln(x) - z$

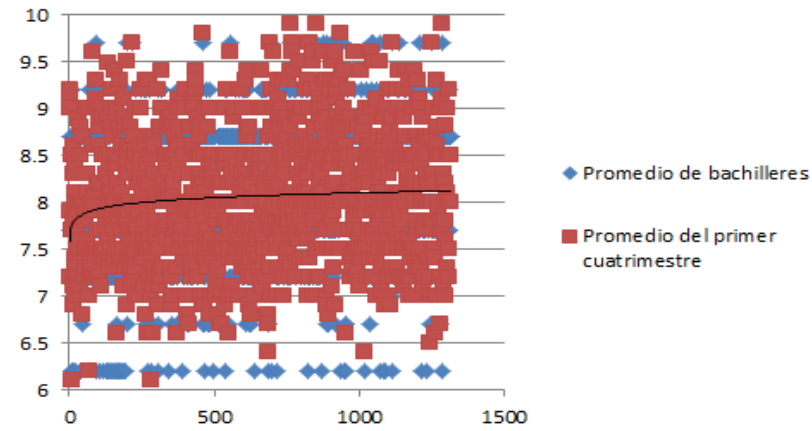


Figura 3 5: Ejemplo de ajuste lineal en el modelo de regresión logarítmica utilizando el conjunto de datos del promedio de los alumnos en el primer cuatrimestre con respecto a su promedio de bachilleres

- Exponencial $y=xe^x$

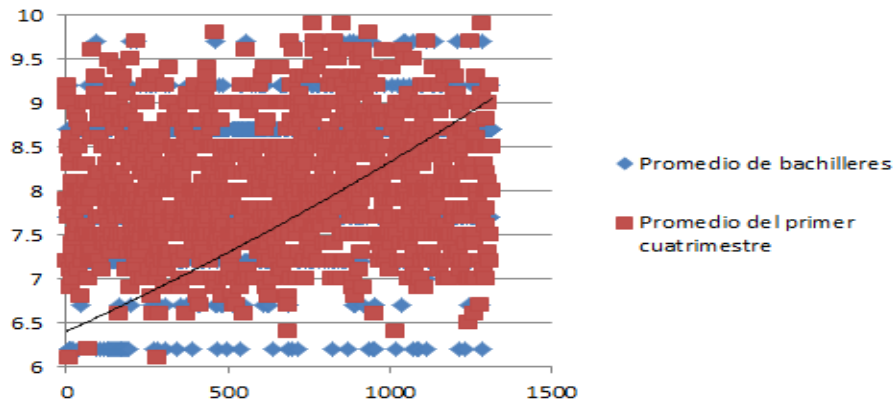


Figura 3 6: Ejemplo de ajuste lineal en el modelo de regresión exponencial utilizando el conjunto de datos del promedio de los alumnos en el primer cuatrimestre con respecto a su promedio de bachilleres

3.2.5 *Coeficiente de correlación de Pearson y Coeficiente de determinación*

Para medir la precisión en los modelos de regresión se necesitan utilizar los coeficientes de correlación y determinación, los cuales nos ayudan a buscar los mejores modelos.

- Coeficiente de correlación

El coeficiente de correlación (R) es lo que nos indica el ajuste lineal que se tiene en una gráfica de dispersión, entre más cercanos se encuentren los datos a la línea de regresión mejor va a ser el resultado de ajuste, y tenderá más a uno. Se utiliza para medir la calidad de los datos y su comportamiento.

- Coeficiente de determinación.

Es el cuadrado del coeficiente de correlación y es el que nos indica el comportamiento de los datos a futuro, la probabilidad de ajuste lineal. Entre más cercano sea su valor a uno mejor va a resultar el modelo de predicción.

3.3 Metodología KDD

Para realizar la metodología de este proyecto se va a basar en el método de extracción de conocimiento llamado *Knowledge Discovery in Databases (KDD)*, que se refiere al proceso de descubrir información útil a partir de grandes volúmenes de datos, es un proceso iterativo que extrae información de calidad para poder crear conclusiones dentro de los datos, (Kaufmann, 2001). Se pueden ver los pasos de esta metodología en la figura 3.7.

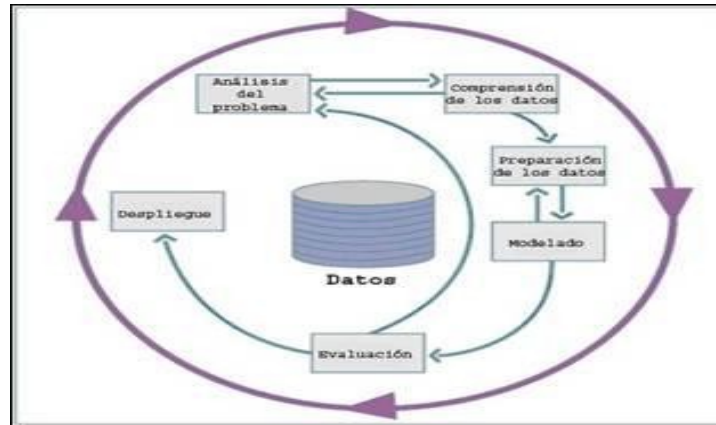


Figura 3 7: Metodología KDD

3.4 Reducción de la dimensionalidad

Se refiere a la reducción de atributos o datos que puede llegar a tener un conjunto, de tal manera de obtener solo aquellos atributos más relevantes para poder tener modelos más eficientes con los cuales se puedan trabajar.

La reducción puede hacerse ya sea buscándose los atributos más relevantes o solo el más relevante, esto sin afectar el entendimiento del modelo ni alterar los atributos.

Capítulo 4: Validación experimental

En este capítulo se presenta el desarrollo de la validación experimental basada en la metodología que anteriormente fue descrita.

4.1 Entendimiento del negocio

Se va a detectar a los alumnos que tengan posibilidades de tener bajo rendimiento escolar en la Universidad Tecnológica de Chihuahua usando los datos que se tienen como resultado en el examen EXANI, combinado con conjuntos de datos proporcionados por la institución, como lo son los promedios del primer cuatrimestre de los alumnos desde el año 2013 y el conjunto con datos de trayectorias el cual contiene más atributos, de esta manera poder ayudar a la institución a aumentar el número de alumnos que concluyen su carrera satisfactoriamente.

4.1.1 Hipótesis

Un sistema de predicción basado en algoritmos de regresión puede identificar a los estudiantes de nuevo ingreso de la Universidad Tecnológica de Chihuahua utilizando atributos como el promedio de bachilleres o la institución de procedencia con mayor probabilidad de fracasar en su carrera, basándose en datos históricos de los promedios del primer cuatrimestre.

4.2 Pre procesamiento y transformación de los datos

Se va a usar el programa de WEKA para poder observar el conjunto de datos de los resultados del examen EXANI.

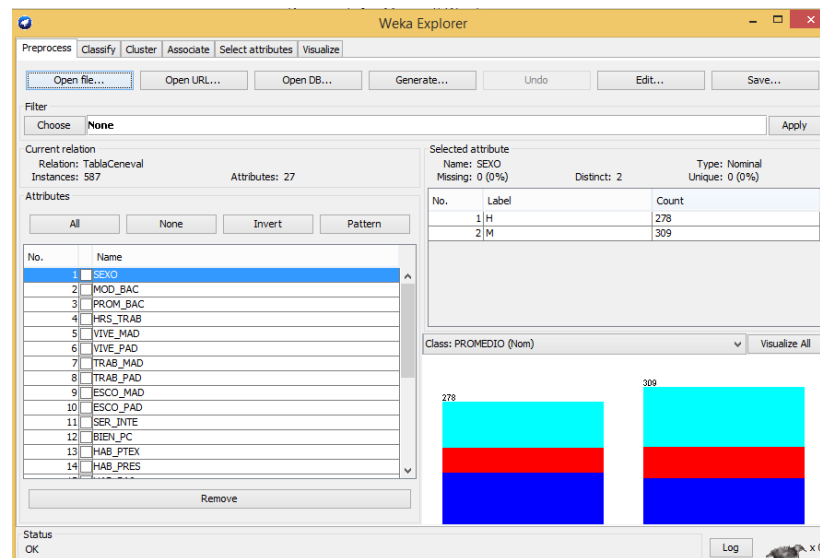


Figura 4 1: Vista de los atributos del conjunto de datos en WEKA

Aquí se puede visualizar cada una de las clases con sus respectivos atributos, y ver una gráfica para visualizar una mejor representación. Se da opción de “remove”, la cual ayuda a borrar atributos con menos importancia.

Usando el lenguaje de java en netbeans y cargando las respectivas librerías de Weka se puede realizar un programa que cumpla con las funciones que nos ofrece el Weka, así que se crea un JFrame Form donde se pueda abrir el conjunto de datos tal y como se muestra en la figura 4.2

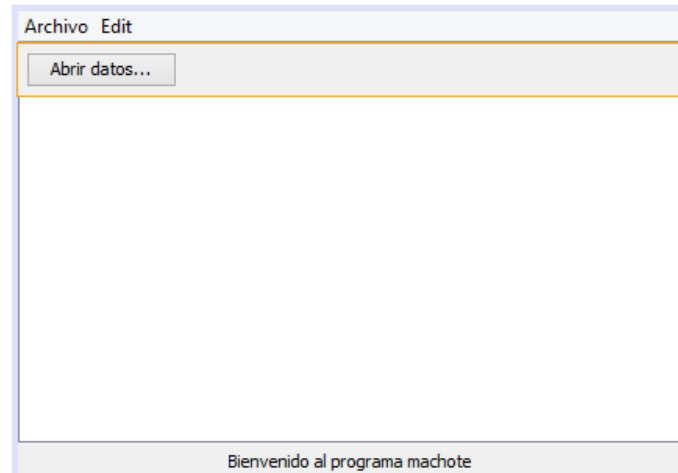


Figura 4 2: Forma para abrir el conjunto de datos del CENEVAL

De esta forma se puede abrir el conjunto de datos del examen EXANI para poder trabajar con él.

```

@attribute TRAB_PAD {padtiemcompleto,padtiemmedio,padnotrabaja,padotros,padnose}
@attribute ESCO_MAD {madnoestudio,madprimaria,madsecundaria,madbachillerato,madtecnic,madlicenciatura,madposgrado,madnose}
@attribute ESCO_PAD {padnoestudio,padprimaria,padsecundaria,padbachillerato,padtecnic,padlicenciatura,padposgrado,padnose}
@attribute SER_INTE {siinternet,nointernet}
@attribute BIEN_PC {nopc,unapc,dospc,trespc,cuatroomas}
@attribute HAB_PTEX {noeditardocumentospc,pcohabilitardocumentospc,hableditardocumentospc,muyhabilitardocumentospc}
@attribute HAB_PRES {nopresentacionespc,pcohabilitacionespc,habilpresentacionespc,muyhabilitacionespc}
@attribute HAB_BAJ {nobajarprogramaspc,pcohabilitarprogramaspc,habilbajarprogramaspc,muyhabilitarprogramaspc}
@attribute LI_INTER {noleerinfoinglesDeInternet,pcohabilleerinfoinglesDeInternet,habilleeinfoinglesDeInternet,muyhabilleerinfoinglesDeInternet}
@attribute LI_TACA {nolextosacademicosingles,pcohabilitextosacademicosingles,habilitextosacademicosingles,muyhabilitextosacademicosingles}
@attribute POS_SEL numeric
@attribute ICNE numeric
@attribute PERCEN numeric
@attribute PCNE numeric
@attribute PRLM numeric
@attribute PMAT numeric
@attribute PRV numeric
@attribute PESP numeric
@attribute PTIC numeric
@attribute PROMEDIO numeric

@data
H,general,8.2,menosCinco,sivimadm,novivpad,madotros,padtiemcompleto,madtecnic,padnose,nointernet,unapc,noeditardocumentospc,muyhabilitar
M,general,9.2,notrabaja,sivimadm,novivpad,madtiemcompleto,padnose,madsecundaria,padsecundaria,nointernet,unapc,noeditardocumentospc,muy
H,tecnologico,7.75,notrabaja,sivimadm,novivpad,madtiemmedio,padtiemcompleto,madlicenciatura,padbachillerato,siinternet,cuatroomas,pcohabilitar
M,general,8.2,notrabaja,sivimadm,novivpad,madnotrabaja,padtiemcompleto,madprimaria,padsecundaria,siinternet,unapc,pcohabilitardocumentos
M,general,8.2,notrabaja,novivpad,madtiemcompleto,padtiemcompleto,madtecnic,padnose,siinternet,trespc,pcohabilitardocumentospc
M,tecnologico,8.2,notrabaja,sivimadm,novivpad,madnotrabaja,padtiemcompleto,madprimaria,padsecundaria,siinternet,cuatroomas,pcohabilitardito
H,tecnico,7.75,menosCinco,sivimadm,novivpad,madnotrabaja,padtiemcompleto,madtecnic,padprimaria,siinternet,cuatroomas,pcohabilitardocum

```

Figura 4 3: Conjunto de datos del EXANI

4.3 Actividades de minería de datos

A continuación se muestran las actividades para cumplir cada objetivo planteado para resolver el objetivo principal.

4.3.1 Actividades del primer objetivo

El primer objetivo consiste en la búsqueda del mejor algoritmo para crear modelos de regresión de alta calidad para estimar el desempeño de cada estudiante de nuevo ingreso.

Para encontrar el mejor algoritmo en los modelos de regresión se necesita llevar a cabo un análisis de los datos durante un pre-procesamiento con reducción de la dimensionalidad utilizando herramientas como WEKA, Excel y programación en Java, y así obtener los coeficientes de determinación (R^2) de los algoritmos.

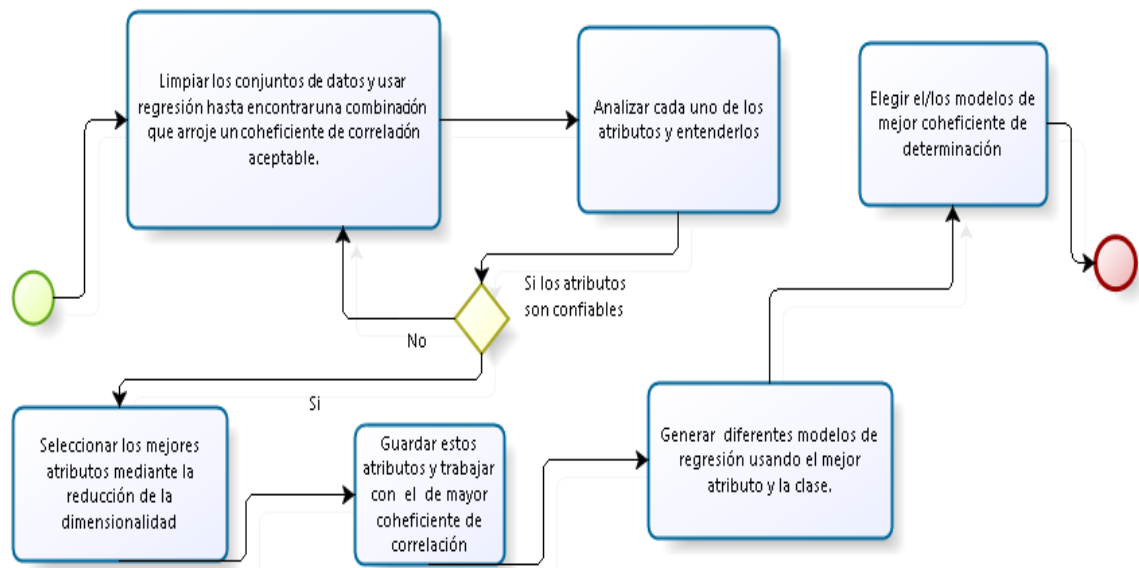


Figura 4 4: Actividades del objetivo 1

4.3.2 Actividades del segundo objetivo

El segundo objetivo consiste en validar la utilidad y novedad de los modelos aprendidos utilizando expertos, buscando la adaptación y mejora.

Mediante la reunión con expertos dentro de la institución se analizará el sistema para validar los modelos obtenidos, incluyendo el mejor, se discutirá sobre los atributos utilizados así como los algoritmos ejercidos, se busca con esto hacer mejoras o cambios para mejorar el sistema conforme a lo que la institución pueda requerir. Tal vez este proceso se pueda repetir más de una vez, conforme a los cambios que se pidan.

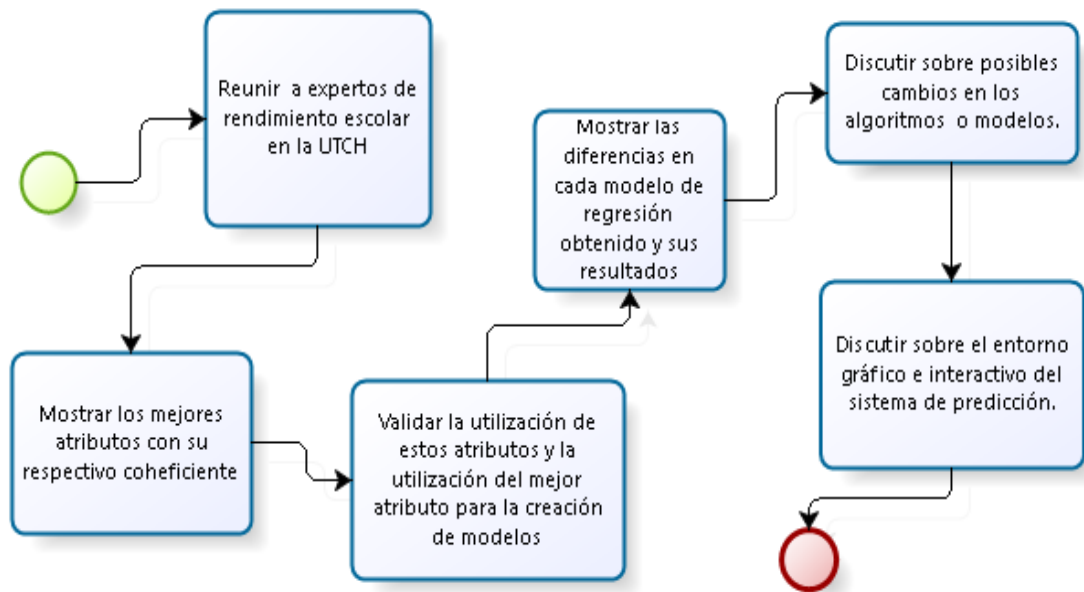


Figura 4 5: Actividades del objetivo 2

4.3.4 Actividades del tercer objetivo

El tercer objetivo consiste en crear una metodología propia para el desarrollo de este tipo de proyectos. Se basa en el método KDD.

En este objetivo a realizar se va a verificar cada actividad para completar este proyecto, ya que originalmente para realizarlo se utilizó la metodología KDD se necesita analizar los cambios o pasos añadidos para la creación de una nueva metodología.

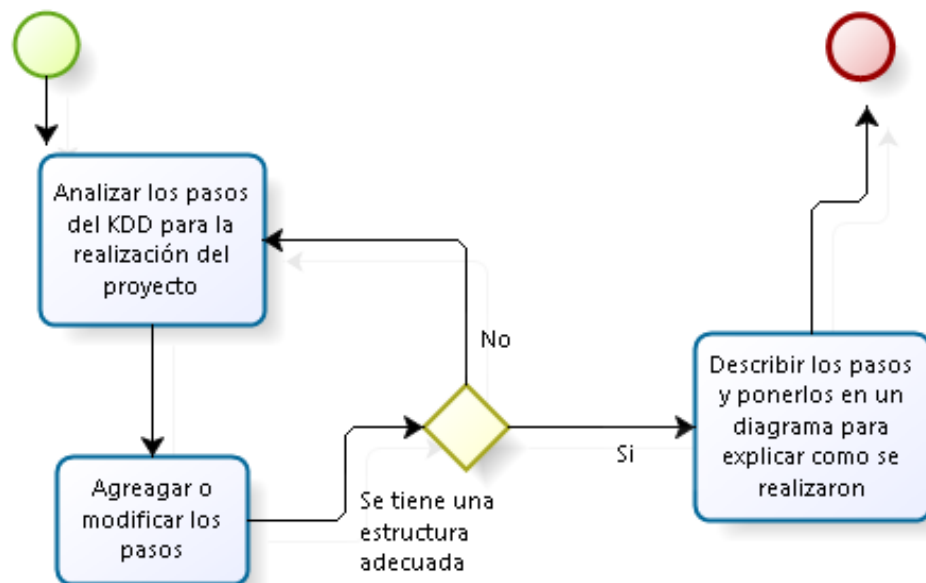


Figura 4 6: Actividades del objetivo 3

4.4 Resultados

A continuación se describirán los resultados obtenidos mediante la realización de los objetivos mencionados anteriormente.

4.4.1 Resultados de actividades del objetivo 1

“Obtener los conjuntos de datos por parte de la Universidad tecnológica y realizar la limpieza de datos. Para poder encontrar el mejor algoritmo se necesitan trabajar con diversos datos obtenidos por parte de la institución, los cuales fueron entregados por medio de archivos XLS ó .DBF, estos archivos contenían información de los EXANIS y de los promedios de los alumnos en primer cuatrimestre de los últimos 4 años, también se cuenta con el conjunto de datos de trayectorias. Estos archivos se cargaron en Weka como se aprecia en la Figura 4.7 para poder realizar un entendimiento de los datos y así comenzar el pre-procesamiento.

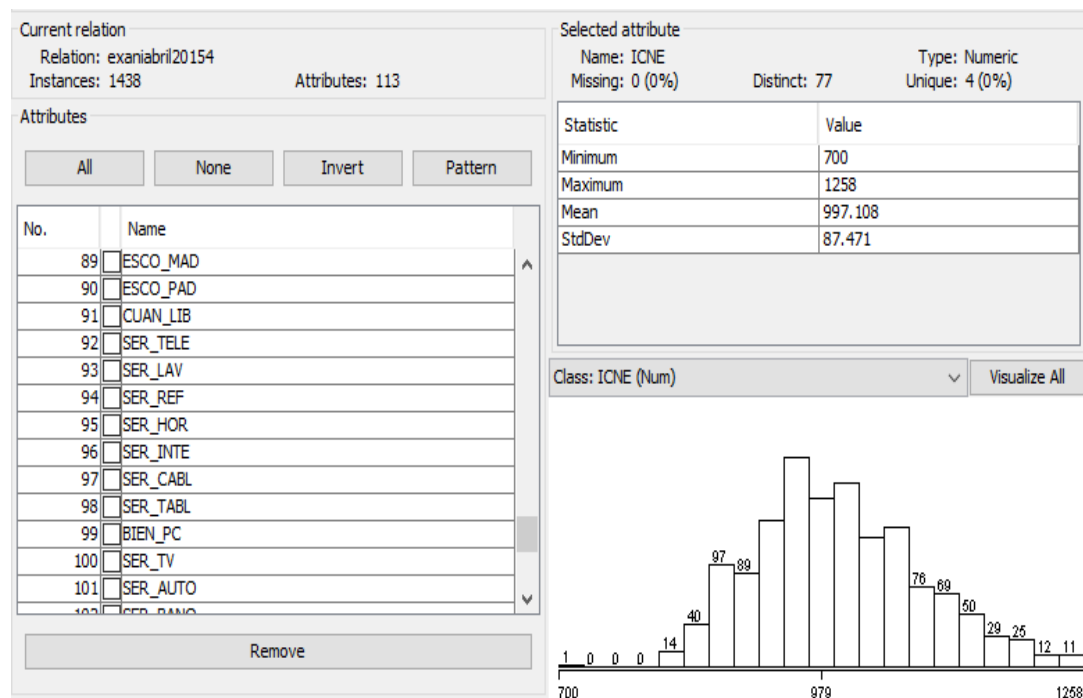


Figura 4 7: Ejemplo de los datos utilizando datos del EXANI de Abril del 2015 en Weka

En estos datos se borraron atributos innecesarios, repetidos ó no válidos, de igual manera también se necesitan aplicar diferentes filtros para clasificarlos como numéricos o nominales, el resultado final es un solo archivo del tipo ARFF el cuál se encuentra con datos limpios y listos para ser trabajados como en la Figura 4.8 y Figura 4.9, donde se excluye la información personal de los alumnos, mostrando solo los atributos de los 2 conjuntos de datos trabajados en este proyecto.

```
@relation Version 4 procesado-weka.filters.unsupervised.attribute.Remove-R5,10,17-weka.filters.unsupervised.attribute.NumericToNominal-R23-23-weka.filters.unsupervised.attribute.NumericToNominal-R29-31-weka.filters.unsupervised.attribute
```

```
@attribute Sexo (Hombre,Mujer)
@attribute Matricula numeric
@attribute Universidad {UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA,'UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA: UNIDAD ACADÉMICA BIS','UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA: UNIDAD ACADÉMICA CUAUHTÉMOC'}
@attribute Carrera {Técnico Superior Universitario-TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN,ÁREA: REDES Y TELECOMUNICACIONES,Técnico Superior Universitario-TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN}
@attribute Tipo plan {Intensivo,Despresurizado}
@attribute Turno {MATUTINO,DESPRESURIZADO,VESPERTINO}
@attribute APLI numeric
@attribute FECHA_APLI numeric
@attribute IDENTIFICA numeric
@attribute FOLIO numeric
@attribute MATRICULA numeric
@attribute DIA_NAC numeric
@attribute MES_NAC numeric
@attribute AÑO_NAC numeric
@attribute LI_MAD {1,2,3}
@attribute LI_PAD {1,2,3}
@attribute EDO_PROC numeric
@attribute NOM_PROC {PREPARATORIA JOHN F. KENNEDY,'PREPARATORIA ESTATAL POR COOPERACIÓN',COLEGIO DE BACHILLERES DEL ESTADO DE CHIHUAHUA,'COLEGIO NACIONAL DE EDUCACIÓN PROFESIONAL TÉCNICA','PRI}
@attribute CVE_PROC {1024711.0,153807.0,229688.0,117752.0,200220.0,223214.0,147807.0,103302.0,168586.0,160770.0,130087.0,193969.0,228360.0,239702.0,243793.0,245574.0,206736.0,208697.0,174071.0,126966.0,174834.0,102906.0}
@attribute AÑO_BAC {1,2}
@attribute REG_PROC {1,2,3}
@attribute MOD_BAC {1,2,3,5,6,7,8}
@attribute PROM_BAC {1,2,3,4,5,6,7,8}
@attribute BEC_DAC {1,2}
@attribute BEC_NEC {1,2}
@attribute BEC_HDA {1,2}
@attribute HRS_TRAB {1,2,3,4,5,6}
@attribute EST_ALCA {1,2,3}
@attribute NO_UNI {1,2,3,4,5,6,7}
@attribute SI_UNI {1,2,3,4,5,6,7}
@attribute SI_POSG {1,2,3,4,5,6,7}
@attribute ESCO_MAD {1,2,3,4,5,6,7,8}
@attribute ESCO_PAD {1,2,3,4,5,6,7,8}
@attribute CUAN_LIB {1,2,3,4,5,6,7,8}
@attribute SER_LAV {1,2}
```

Figura 4 8: Archivo completo en ARFF de promedios con EXANI.

```

@attribute SER_BANO {1,2,3,4,5}
@attribute VAC_RM {1,2,3,4,5}
@attribute CURP {LETJ951020HCHCRN03,NISO911101HCHTGM08,PACM931026MCHLNR06,GURN951007MCHLZN06,AEMD970419MCHLRN05,PERC900308HCHRSR06,JIMR860707HCHIMRF01,CAMJ970919HCHSRH00,RAPE870326HCH}
@attribute 39.31 {1,2,3}
@attribute Modalidad {1,2}
@attribute Turno {1,2}
@attribute familia {101,102,103,105,107,109}
@attribute area {1,2,3,5,9}
@attribute NombrePE {DESARROLLO DE NEGOCIOS □ REA MERCADOTECNIA; TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN □ REA SISTEMAS INFORMÁTICOS; MECATRÓNICA □ REA AUTOMATIZACIÓN; MANTENIMIENTO □ REA I}
@attribute Tipo de plan de estudios {1,2}
@attribute 5 {1,2}
@attribute 6 {1,2}
@attribute 7 {1,2}
@attribute 8 {1,2}
@attribute 9.1 {1,2,3,4,5,6,7,8,9}
@attribute 9.2 {1,2,3,4,5,6,7,8,9}
@attribute 10.1 {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18}
@attribute 10.2 {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18}
@attribute 11 {1,2,3,4,5,6,7,8,9,10}
@attribute 12 {1,2,3}
@attribute 13 {1,2}
@attribute 14 {1,2}
@attribute 15 {1,2,3,4,5}
@attribute 16 {1,2,3,4,5}
@attribute 17 {1,2}
@attribute 18.1 {1,2}
@attribute 18.2 {1,2}
@attribute 18.3 {1,2}
@attribute 18.4 {1,2}
@attribute 18.5 {1,2}
@attribute 19 {1,2,3,4,5}
@attribute 20 {1,2,3}
@attribute 21 {1,2,3,4,5}
@attribute 22 {1,2,3,4,5}
@attribute 23 {1,2}
@attribute 24 {1 2 3 4 5 6 7}

```

Figura 4 9: Archivo ARFF de trayectorias con EXANI, mostrando algunos de los campos.

En la selección de los mejores atributos en el conjunto de datos es cuando se comienza a trabajar con los datos ya procesados se puede utilizar regresión lineal utilizando librerías de Weka en un programa en java para comparar cada atributo con respecto a la clase que se quiere predecir, en este caso el promedio en el primer cuatrimestre. Esto da como resultados diferentes coeficientes de correlación en cada atributo, haciéndose observable la importancia de cada uno de ellos y el mejor como se ve en la Figura 4.10 y 4.11 con los respectivos conjuntos de datos.

24	BEC_DAC	0.3025
25	BEC_NEC	0.1796
26	BEC_HDA	0.1067
27	HRS_TRAB	0.0702
28	EST_ALCA	0.0034
29	NO_UNI	-0.0163
30	SI_UNI	0.0197
31	SI_POSG	-0.0380
32	ESCO_MAD	0.0513
33	ESCO_PAD	-0.0330
34	CUAN_LIB	-0.0286
35	SER_LAV	-0.0537
36	SER_REF	0.0599
37	SER_HOR	-0.0463
38	SER_INTE	-0.0040
39	SER_CABL	-0.0703
40	BIEN_PC	0.0277
41	SER_TV	0.0695
42	SER_AUTO	0.0166
43	SER_BANO	-0.0698
44	VAC_RM	0.0377

El mejor atributo es PROM_BAC con un R de 0.8407

Figura 4 10: Lista de atributos con su respectivo coeficiente de correlación y la selección del mejor del conjunto con promedio y EXANI.

139	35.1	-0.0520
140	35.11	0.1104
141	35.12	0.0243
142	36.11	-0.0820
143	36.12	-0.0820
144	36.13	-0.0820
145	36.14	-0.0778
146	36.15	-0.0820
147	36.16	-0.0820
148	36.17	-0.0820
149	36.21	-0.0820
150	36.22	-0.0820
151	36.23	-0.0820
152	36.24	-0.0896
153	36.25	-0.0820
154	36.26	0.0191
155	36.27	-0.0705
156	36.28	0.0141
157	37.1	0.0514
158	37.2	-0.0820
159	38.01	0.0066
160	38.02	-0.0820
161	38.03	0.0083
162	38.04	-0.0025
163	38.05	-0.0180
164	38.06	-0.0820
165	38.07	-0.0820
166	38.08	0.0427
167	38.09	-0.0820
168	38.1	-0.0820
169	38.11	0.0572
170	38.12	-0.0820
171	39.1	-0.0820
172	39.2	-0.0710
173	39.3	-0.0820
174	39.11	0.0021
175	39.21	-0.1155
El mejor atributo es PROM_BAC con un R de 0.6116		

Figura 4 11: Lista de atributos con su respectivo coeficiente de correlación y la selección del mejor del conjunto con trayectorias y EXANI

Es importante observar que en ambos conjunto de datos el atributo con mayor importancia según su coeficiente de correlación es PROM_BAC, el cual es el promedio de bachilleres de los alumnos, con este atributo se va a poder trabajar para crear mejores modelos utilizando el concepto de reducción de la dimensionalidad.

Cuando ya se tiene seleccionado el mejor atributo se pueden aplicar conceptos de reducción de la dimensionalidad, con esto se realizan diferentes transformaciones utilizando el mejor atributo con respecto a la clase, la cual es promedios, cuando ya se tienen estas transformaciones podemos hacer una comparación de sus coeficientes de determinación para poder elegir el que mejor nos convenga, en el caso del conjunto de datos de los promedios del primer cuatrimestre y resultados del EXANI los mejores modelos son el Lineal y cuadrático con un coeficiente de determinación mayor, como se puede observar en la figura 4.12 .

No.	Tipo	R2
1	Lineal	0.8544
2	Cuadrática	0.8544
3	Exponencial	0.8467
4	Logaritmico	0.7557
5	Potencial	0.7658

Figura 4 12: Modelos de regresión con su coeficiente de determinación promedios con EXANI

En el conjunto de datos de las trayectorias con los datos del EXANI es interesante que a pesar de que el promedio de bachilleres tiene un coeficiente de correlación menor de .61 a diferencia de los promedios con .84 se logra un modelo con un coeficiente de determinación mayor de .89,

No.	Tipo	R2
1	Lineal	0.3740
2	Cuadrática	0.3822
3	Exponencial	0.8939
4	Logaritmico	0.3836
5	Potencial	0.8539

Figura 4 13: Modelos de regresión con su coeficiente de determinación trayectorias con EXANI

4.4.2 Resultados de actividades del objetivo 2

“Presentación de resultados a directivos de la UACH”.

En una reunión programada por la Universidad Tecnológica de Chihuahua se muestran los resultados del trabajo obtenido mostrando los atributos ordenados base a su coeficiente de correlación, para que ellos puedan determinar si el atributo seleccionado como el mejor es de su agrado, primero se muestran los atributos del conjunto de datos de los promedios y EXANI esto se muestra en la figura 4.14.

No.	Atributo	R(Valor absoluto)
23	PROM_BAC	0.8407
19	CVE_PROC	0.3336
24	BEC_DAC	0.3025
18	NOM_PROC	0.2662
21	REG_PROC	0.265
14	ANO_NAC	0.1921
25	BEC_NEC	0.1796
4	Carrera	0.1653
5	Tipo plan	0.1532
6	Turno	0.1532
20	ANO_BAC	0.1424
1	Sexo	0.1158
26	BEC_HDA	0.1067
12	DIA_NAC	0.0832
17	EDO_PROC	0.0752
32	ESCO_MAD	0.0752
34	CUAN_LIB	0.0752
40	BIEN_PC	0.0752
43	SER_BANO	0.0752
44	VAC_RM	0.0733
9	IDENTIFICA	0.0726
39	SER_CABL	0.0703
27	HRS_TRAB	0.0695
41	SER_TV	0.0662
13	MES_NAC	0.0629
42	SER_AUTO	0.062
15	LI_MAD	0.0603
36	SER_REF	0.0599
35	SER_LAV	0.0537
11	MATRICULA	0.0507
37	SER_HOR	0.0463
28	EST_ALCA	0.0456
22	MOD_BAC	0.041
31	SI_POSG	0.038
10	FOLIO	0.029
30	SI_UNI	0.0197
29	NO_UNI	0.0163
16	LI_PAD	0.0158
7	APLI	0.0151
2	Matricula	0.0145
8	FECHA_APLI	0.0086
33	ESCO_PAD	0.0065
38	SER_INTE	0.004
3	Universidad	0.0007

Figura 4 14: Atributos ordenados con base a su coeficiente de correlación

También se muestran los atributos ordenados con del conjunto de datos de trayectorias con los resultados del EXANI.

No.	Atributo	R
21	PROM_BAC	0.6116
10	ANO_NAC	0.2258
20	MOD_BAC	0.2069
18	ANO_BAC	0.1777
59	11	0.1636
22	BEC_DAC	0.1548
17	CVE_PROC	0.1508
60	12	0.1246
19	REG_PROC	0.1232
122	32.16	0.1141
140	35.11	0.1104
11	SEXO	0.107
16	CIU_PROC	0.0997
117	32.11	0.0991
15	NOM_PROC	0.0915

Figura 4 15: Atributos ordenados con base a su coeficiente de correlación

Ellos analizan estos atributos y pueden llegar a proponer repetir el proceso con otro atributo de elevada importancia. Después de esto se muestra un análisis del atributo seleccionado y de la clase para dar un mejor entendimiento a las personas de la reunión, como se muestra en la figura 4.6.



Figura 4.16: Explicación del mejor atributo y la clase

De igual manera también se muestran los resultados de cada algoritmo, esto para mostrar por que se utilizó el concepto de reducción de la dimensionalidad y no la utilización de todos los datos, como se muestra en la figura 4.7.

Tabla 1: Tabla comparativa de cada modelo de regresión

Modelo tipo	R ²	Conjuntos	Datos utilizados	Fórmula
Lineal	.8544	EXANI y control escolar	El mejor atributo	$Y = a + bX + e$
Cuadrático	.8544	EXANI y control escolar	El mejor atributo	$Y = aX^2 + bX + c$
Exponencial	.8467	EXANI y control escolar	El mejor atributo	$Y = e^x$
lineal	.8460	EXANI y control escolar	Los 3 mejores atributos	$Y = a + bX + e$
Exponencia l	.8939	EXANI y trayectorias	El mejor atributo	$Y = e^x$
Potencial	.8539	EXANI y trayectorias	El mejor atributo	$Y = ax^b$

El atributo marcado en la tabla es el modelo Exponencial el cual maneja el mejor coeficiente de determinación manejando el mejor atributo, el mejor coeficiente de determinación encontrado en el trabajo en estos modelos es de .8939.

Después de mostrar estos resultados al comité encargado de evaluar este proyecto se le entregará un cuestionario el cual contendrá las siguientes preguntas de acuerdo al método DELPHI:

1.- Los atributos elegidos en el modelo con el coeficiente más alto ¿Son verdaderamente funcionales para todo tipo de alumnos?

2.- ¿Los atributos con mayor importancia son adecuados para realizar un sistema que mida el desempeño académico de un aspirante?

3.- ¿Es satisfactorio el valor del coeficiente de determinación en los modelos seleccionados para poder dar pie al siguiente paso?

Esto se hará con la finalidad de que se evalúen las respuestas en cada iteración o corrección del sistema, cuando las respuestas sean verdaderamente satisfactorias se continuará al siguiente paso.

Cuando los expertos de la institución autorizan el modelo obtenido se lleva a cabo la propuesta para la interfaz del sistema de predicción que van a utilizar, así van visualizando la utilización del sistema, tal como se observa en la figura 4.17.

Sistema de Predicción UTCH
para unidopromedioversion5

Nombre del postulante:
Alfredo De La Cruz

NOM_PROG:
CENTRO DE ESTUDIOS TECNOLÓGICOS INDUSTRIAL Y DE SERVICIOS

PROM_BAC:
8

Predecir... Ver rangos

Figura 4 17: Primera vista del sistema de predicción

Cuando se eligen los valores de los 2 atributos de nombre de procedencia y promedio de bachilleres seleccionados en este programa se puede ver la predicción al presionar el botón “predecir...” como se ve en la figura 4.18

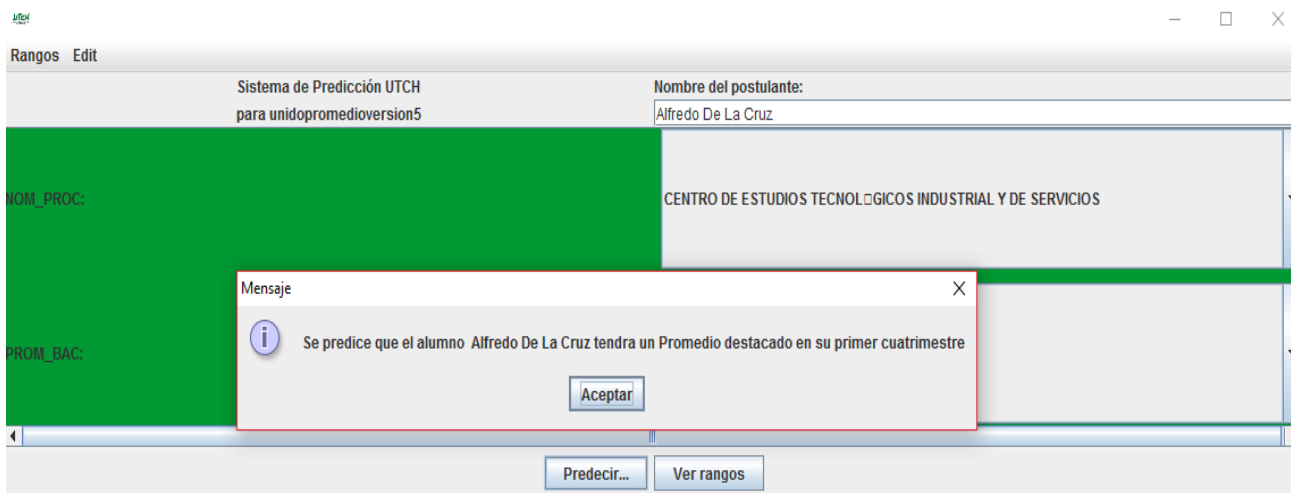


Figura 4 18: mensaje de predicción al presionar el botón “predecir..” según los valores indicados

Cuando se presiona el botón de “Ver rangos” aparece la información con respecto a los rangos pedidos por la institución. Esto se muestra en la figura 4.19.

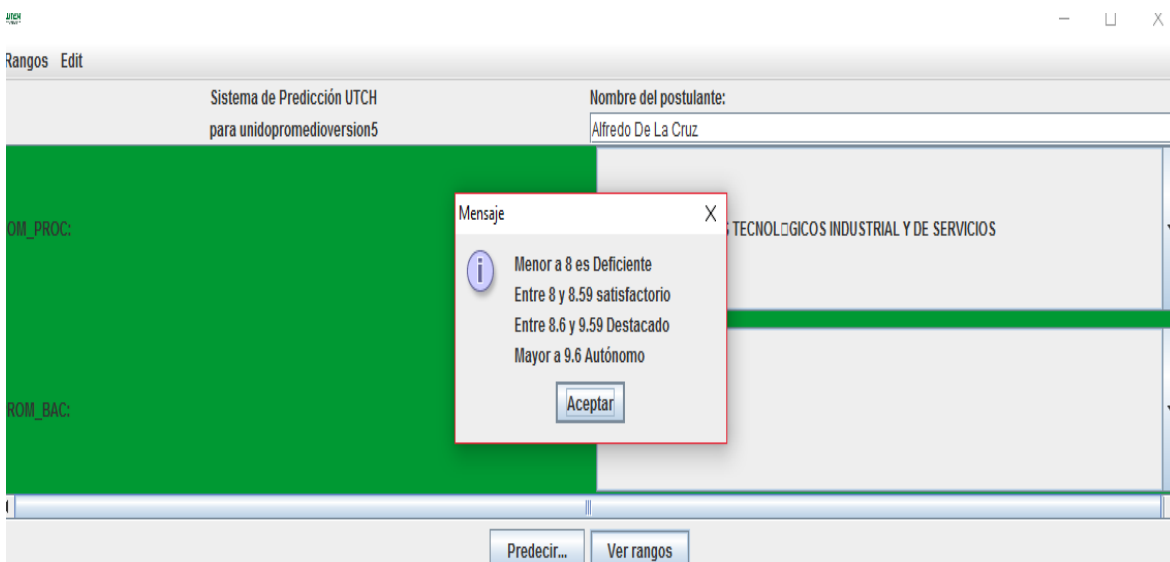


Figura 4 19: Información de rangos al presionar el botón “ver Rangos”

4.4.3 *Resultados de actividades del objetivo 3*

“Análisis de cada actividad para la realización de una nueva metodología”.

Durante la realización de este proyecto se fueron siguiendo una serie de pasos que a pesar de ser basados en la metodología de KDD fueron cambiando durante las actividades de este proyecto, naciendo así una nueva metodología como se muestra a continuación.

1. **Entendimiento del problema:** Se hablará con la Universidad para plantear el problema que hay sobre el bajo desempeño académico.
2. **Entendimiento de los datos:** Se trabajan los datos proporcionados por la UTCH y se analiza el significado de cada uno con ayuda de las personas expertas.
3. **Pre-procesamiento:** Se realizará tratamiento de los datos para generar variables a partir de las ya existentes con una estructura de datos apropiada.
4. **Reducción de la dimensionalidad:** Se buscaran los atributos con mayor importancia para generar diferentes modelos con minería de datos.
5. **Minería de datos:** Se realizará el modelado de los datos integrando los algoritmos de regresión, dependiendo de la cantidad que arroje el paso anterior.
6. **Interpretación y evaluación:** Se realizará la evaluación de los patrones que serán obtenidos con ayuda de expertos en la escuela.
7. **Despliegue de la aplicación:** Se dejará el sistema a la Universidad, para que puedan utilizarlo en futuras tomas de decisiones, según su conveniencia.

Se puede apreciar mejor esta nueva metodología en la figura 4.20.

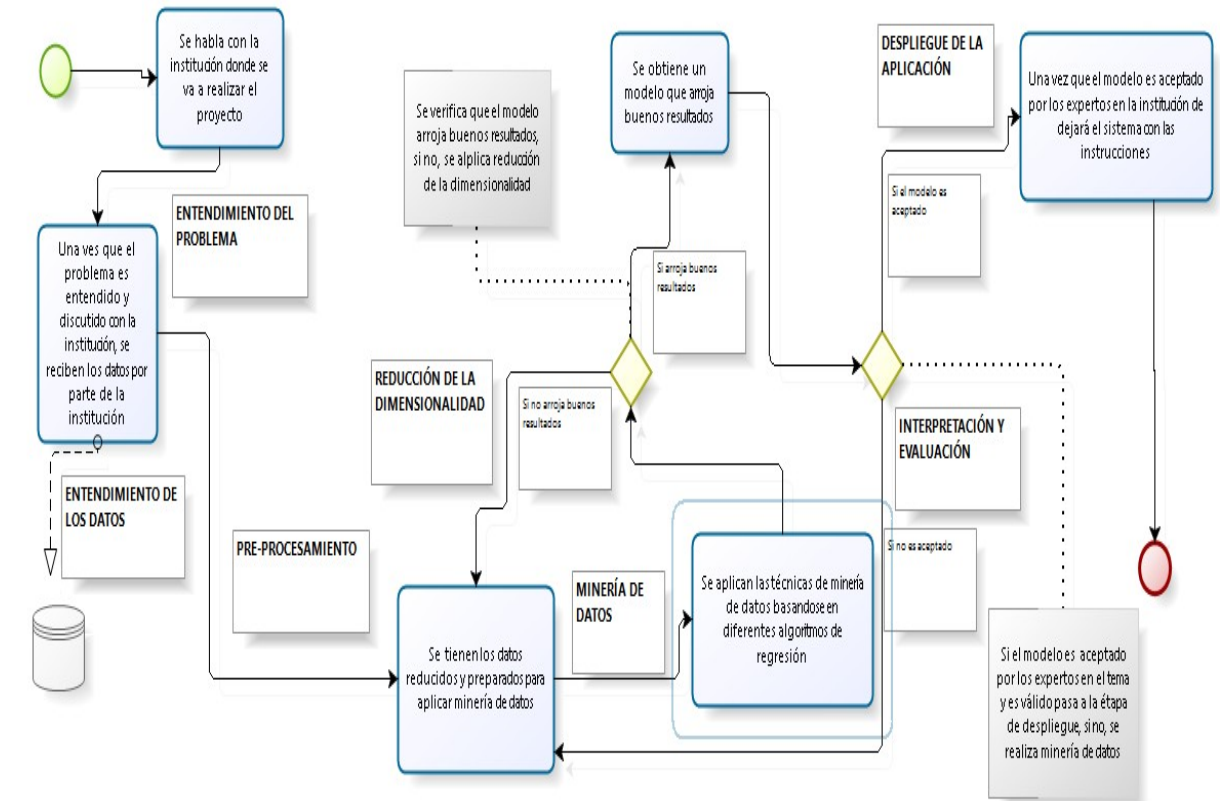


Figura 4.20: Metodología propia



Capítulo 5: Conclusiones

Basado en los resultados obtenidos en esta investigación se comprueba la hipótesis planteada para poder desarrollar un sistema de predicción que arroje buenos resultados para poder identificar a los alumnos que puedan llegar a tener bajo rendimiento escolar durante su primer cuatrimestre en la universidad.

5.1 Conclusiones del objetivo 1

“Búsqueda del mejor algoritmo para crear modelos de regresión de alta calidad para estimar el desempeño de cada estudiante de nuevo ingreso.”

Para buscar el mejor algoritmo de regresión no basta con conformarse en usar todas las variables que contiene un conjunto de datos, hay que buscar más allá, medir la importancia de los atributos y comparar cada modelo obtenido y trabajado, esto aumenta significativamente el resultado de los algoritmos dándonos al final un programa de predicción más confiable y con un buen resultado.

Es interesante observar como tomando solo el atributo más importante en un conjunto de datos se pueden obtener resultados favorables haciendo las transformaciones adecuadas a la línea de regresión, y ya se elige dependiendo de la distribución de los datos representados en la gráfica de dispersión.

Los mejores modelos obtenidos fueron el lineal y el cuadrático usando solo el mejor atributo en el conjunto de datos de promedios con EXANI y fue el logarítmico con el mejor atributo en el conjunto de datos con EXANI y trayectorias, esto comprueba que la reducción de la dimensionalidad verdaderamente funciona ya que el coeficiente de determinación es funcional para poder usarlo como un instrumento predictivo.

5.2 Conclusiones del objetivo 2:

“Validar la utilidad y novedad de los modelos aprendidos utilizando expertos, buscando la adaptación y mejora.”

Como ingenieros se pueden desarrollar diferentes proyectos buscando solamente resultados altos, sin tratar que sea algo entendible o verdaderamente funcional para los demás, es este tipo de situaciones que se busca la ayuda de expertos en el área del proyecto que se realiza, en este caso proyecto de evaluación del desempeño académico.

Con la aprobación de los expertos se puede dar pie a realizar la integración del sistema con mejoras y más entendimiento para todos. Es necesario realizar algunos cambios conforme a lo que los expertos sugieran.

En la siguiente figura se muestra el procedimiento del método Delphi que nos permitió dar pie a programar el sistema.

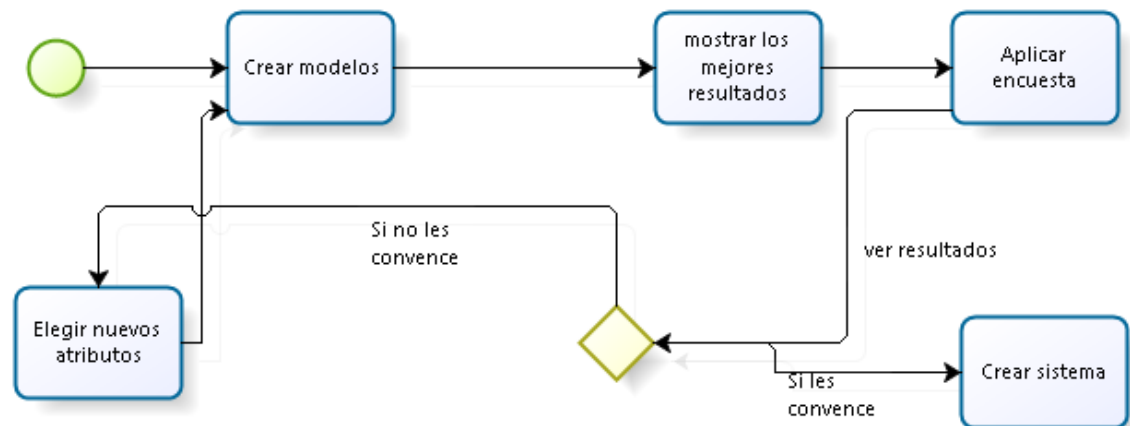


Figura 5 1: método Delphi

5.3 Conclusiones del objetivo 3

“Crear una metodología propia para el desarrollo de este tipo de proyectos. Se basa en el método KDD.”

Cuando se busca realizar un proyecto es fundamental ser ordenado y que cada paso sea trabajado con entendimiento, para esto hay diversas metodologías. Se ha probado con este proyecto que una metodología ya establecida no es suficiente para completar satisfactoriamente cada uno de los pasos requeridos para terminar el sistema, es por eso que la metodología propia es la adecuada y con los pasos que fueron necesarios para este trabajo.

Por lo tanto se concluye que la metodología propia basada en la metodología KDD es la adecuada para la realización de este tipo de proyectos, ya que se cumplió con los pasos establecidos de forma exitosa.

Bibliografía

- Arenas, A. G. (1996). Tecnologías de información y estrategias de negocios. *soluciones avanzadas*.
- Becerra, A. C., Murieta, B. I., Garza, H. D., & Vega, R. B. (2012). Perfiles que definen al alumno en su rendimiento escolar. *CIPITECH*.
- CENEVAL. (2014). <http://www.ceneval.edu.mx/>.
- Cruz Salazar, L. A., Muñoz Aldana, D. J., & Contreras Montes, J. A. (2008). Implementación de redes neuronales y lógica . *IEEE*.
- Devore, J. L. (2009). Probabilidad y Estadística para Ingeniería y Ciencias. *International Thomson Editores*.
- Fiegehen, L. E. (2005). Repitencia y deserción universitaria en América Latina. *IESALC*.
- Kaufmann, M. (2001). *Data Mining: Concepts ans Techniques*.
- Luna, F. V. (2007). Modelo lineal general jer'arquico. *Laboratorio de Investigaci'on y Asesor'ia Estad'istica, LINA E*.
- Miñano Pérez, P., Cantero Vicente, M. P., & Castejón Costa, J. L. (2008). PREDICCIÓN DEL RENDIMIENTO ESCOLAR DE LOS ALUMNOS A PARTIR DE LAS.
- Moxons, B. (1996). Defining data mining.
- Orallo, J. H. (2007). *Introducción a la minería de datos*. Bogotá: Pearson.
- OSMANBEGOVIĆ, 1., AGIĆ, H., & SULJIĆ, M. (2014). PREDICTION OF STUDENTS' SUCCESS BY APPLYING DATA MINING ALGORITHMS. *ISSN*:, 11.
- Osmanbegović, E., & Suljić, M. (2012). DATA MINING APPROACH FOR PREDICTING STUDENT PERFORMANCE. *Economic Review – Journal of Economics and Business, Vol. X, 10*.
- Porcel, E. A., Dapozo, G. N., & Victoria, M. (2009). Prediction of the academic performance of freshmen at the FACENA (UNNE) based on their socio–educational characteristics. *Facultad de Ciencias Exactas y Naturales y Agrimensura Universidad Nacional del Nordeste*.



- Pratiwi, O. N., & Informatika, T. (2010). Predicting Student Placement Class using Data. *Cikutra 204A*.
- R, J. J., H., A. G., & A., J. A. (2014). Predicción del rendimiento en una asignatura empleando la regresión logística ordinal. *SciELO*.
- Rokach, O. M. (2010). *Data Mining and Knowledge Discovery Handbook*. New York.
- Silberschatz, A., Korth, H. F., & Sudarshan, S. (2002). *FUNDAMENTOS DE BASES DE DATOS*. Madrid: Mc Graw Hill.
- Valquíria R. C. Martinho, C. N. (2013). An Intelligent System for Prediction of School Dropout Risk Group in Higher . *IEEE*.
- Viñuela, P. I. (2003). *REDES NEURONALES ARTIFICIALES Y*. Pearson Prentice Hall.
- Witten, I. H., & Eibe Frank, L. T. (1999). «*Weka: Practical Machine Learning Tools and Techniques with Java Implementations*».
- Yi-sheng, H., Chao, M., & Zhi-bin, L. (2010). System Prediction and Control Model Based on . *IEEE*.
- Yunta, L. R. (2001). Bases de datos documentales: estructura y uso. *CINDOC*.



JOSÉ ALFREDO DE LA CRUZ NAVA



26 AÑOS, CHIHUAHUA, SOLTERO. LICENCIA DE CONDUCIR 95434
FECHA DE NACIMIENTO: 27. 02. 1990
LAURO GOMEZ #13
MISAE L NUÑEZ, CHIHUAHUA, CHIHUAHUA
T E L . (6 1 4) 2 2 2 6 8 3 1
ALFREDODELAC@HOTMAIL.COM
OFUS27@GMAIL.COM
ALFREDODELACN@GMAIL.COM
TITULADO
CEDULA PROFESIONAL: 9410295
VISA Y PASAPORTE VIGENTES

OBJETIVO

Ingeniero en Sistemas

EDUCACIÓN

2008 – 2013 [Instituto Tecnológico De Chihuahua II] [Chihuahua, Chihuahua]

2005 – 2008 [Centro De Bachillerato Industrial y De Servicios 122] [Chihuahua, Chihuahua] (Titulado)

EXPERIENCIA LABORAL

2012-2013 [Safran (Labial-planta 3)] [Chihuahua, Chihuahua]

{6 meses de practicante y 2 meses de contratado como ingeniero en telecomunicaciones, supliendo a personal en el extranjero}

{Soporte a computadoras del área de diseño, monitoreo de servidores, manejo de servidores, implementación de soluciones de hardware, solución de problemas básicos de Catia, Ennovia, telefónica IP entre otros programas y Soporte a la red}

2013-2014 (Y PARTE DEL 2015) [Técnico en sistemas de Interceramic (contrato externo por medio de SYS)] [Chihuahua, Chihuahua]

{Soporte técnico a equipos de cómputo, programas, impresoras, site (switches, servidores, telefonía), sistema basado en tickets}

HABILIDADES

Lenguajes:

JAVA, C++, VISUAL BASIC, DELPHI, PHP.

Bases de datos:

ORACLE, SQL, MYSQL.