UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE CIENCIAS DE LA INGENIERÍA

CARRERA INGENIERÍA EN SISTEMAS

Tema de la Tesis

"Análisis inteligente de datos para el contraste de información académica con el fin de obtener patrones de comportamiento de los estudiantes de la UTEQ"

Previo a la obtención del título de:

Ingeniero en Sistemas

Autor:

Pablo Rafael Vélez Franco

Director de Tesis

PhD. Amilkar Puris Cáceres

Quevedo – Ecuador.

Año

2015

CÓDIGO DUBLÍN

1.	Título	M	ANÁLISIS INTELIGENTE DE DATOS PARA EL CONTRASTE DE INFORMACIÓN ACADÉMICA CON EL FIN DE OBTENER PATRONES DE COMPORTAMIENTO DE LOS ESTUDIANTES DE LA UTEQ
2.	Creador	M	Pablo Rafael Vélez Franco
3.	Materia	M	Facultad de Ciencias de la Ingeniería; Análisis inteligente de datos: Minería de datos
4.	Descripción	M	La presente tiene como objetivo mediante la aplicación de análisis inteligente de datos determinar los factores socioeconómicos que intervienen en el rendimiento académico de los estudiantes de la Universidad Técnica Estatal de Quevedo.
5.	Editor	M	Facultad de Ciencias de la Ingeniería: Carrera de Ingeniería en sistemas.
6.	Colaborador	0	PhD. Amilkar Puris Cáceres Ing. Jorge Guanín Fajardo
7.	Fecha	M	13-07-2015
8.	Tipo	M	Proyecto de Tesis
9,	Formato	M	Ms Word, Pdf
10	Identificador	M	Ninguna
11	Fuente	M	Minería de datos
12	Lenguaje	M	Español
13	Relación	0	Ninguno
14	Cobertura	0	Ninguno
15	Derechos	M	Ninguno
16	Audiencia	0	Proyecto de Investigación / Research Projec

UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD CIENCIAS DE LA INGENIERÍA

CARRERA INGENIERÍA EN SISTEMAS

Presentado al Consejo Directivo como requisito previo a la obtención del título de Ingeniero en Sistemas

Aprobado:

Ing. Carlos Márquez de la Plata

PRESIDENTE DEL TRIBUNAL DE TESIS

Ing. Washington Chiriboga

MIEMBRO DEL TRIBUNAL DE TESIS

Ing. José Luis Tubay

MIEMBRO DEL TRIBUNAL DE TESIS

QUEVEDO – LOS RIOS – ECUADOR

Año 2015

DEDICATORIA

Este trabajo está dedicado principalmente a Dios por haberme permitido llegar hasta este momento tan importante de mi vida.

A mis padres Rafael Vélez y Esperanza Franco, por brindarme todo el apoyo necesario, e inculcarme sus valores y principios. Ellos han formado lo que soy como persona, mi carácter, mi empeño, mi fortaleza y mi perseverancia para lograr mis objetivos.

A mis hermanos, José, María, Javier y en especial a mi ñañita Dexy, por el incondicional ánimo que me supieron brindar en cada etapa de mi vida.

A mis grandes amigos Johanna e Iván por haberme soportado y brindado su amistad, sin ellos no hay duda que este largo camino no hubiese sido el mismo.

Aquellas personas que siempre me impulsaron en seguir adelante.

AGRADECIMIENTO

Primero agradezco a Dios por haberme acompañado a lo largo de mi carrera y brindarme la paciencia y ánimo necesario para culminar este objetivo propuesto hace muchos años.

Un agradecimiento a todas las personas que de alguna forma fueron parte importante para la culminación de este proyecto.

A la prestigiosa Universidad Técnica Estatal de Quevedo y sus distinguidos docentes, por brindarme el conocimiento necesario para afrontar esta y cada una de las etapas siguientes en mi vida.

Al Ing. Jorge Guanín y el Dr. Amilkar Puris, por todo el conocimiento impartido y tiempo dedicado durante el desarrollo de esta investigación.

A quienes a pesar de mis defectos, han formado parte de mi vida y siempre han estado ahí ayudando a convertirme en la persona que soy ahora.

Para todos ellos: Muchas gracias y que Dios los bendiga.

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

Yo, Pablo Rafael Vélez Franco, declaro que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo, puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

Pablo Rafael Vélez Franco

CERTIFICACIÓN DEL DIRECTOR DE TESIS

El suscrito, PhD. Amilkar Puris Cáceres, Docente de la Universidad Técnica Estatal de Quevedo, certifica que el Egresado Pablo Rafael Vélez Franco, realizó la tesis de grado previo a la obtención del título de Ingeniero en Sistemas, titulada “Análisis inteligente de datos para el contraste de información académica con el fin de obtener patrones de comportamiento de los estudiantes de la UTEQ”, bajo mi dirección, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

_____

PhD. Amilkar Puris Cáceres

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

RESUMEN EJECUTIVO

En este trabajo se presenta el estudio de los diferentes factores socioeconómicos y su influencia sobre el rendimiento académico de los estudiantes de la Universidad Técnica Estatal de Quevedo.

La información para la presente investigación fue obtenida de los sistemas de información con los que cuenta la universidad, y se seleccionaron los registros de los estudiantes que ingresaron entre los años 2002 y 2008.

Se realizó la unificación de las diferentes fuentes de datos, identificación de información relevante, y limpieza de datos, procesos necesarios para tener un conjunto de datos óptimo para aplicar procesos de minería de datos.

Se realizaron diversos experimentos con diferentes conjuntos de datos y software de minería de datos los cuales, posterior a su análisis, se obtuvo resultados que permitieron determinar las características socioeconómicas que definen en gran medida el rendimiento de un estudiante.

El análisis de los resultados obtenidos, podría ayudar en el proceso de toma de decisiones académicas, y así desarrollar medidas de apoyo a los actuales y futuros estudiantes.

ABSTRACT

In this work the study of different socioeconomic factors and their influence on the academic performance of students in the State Technical University of Quevedo.

The information for this research was obtained from information systems are there in college, and records of students who entered between 2002 and 2008 were selected.

The unification of the different data sources, identifying relevant information and data cleansing processes necessary to have a set of optimal data to apply data mining processes was performed.

Various experiments with different sets of data and data mining software's, after analysis, which results allowed to determine the socioeconomic characteristics that largely define the performance of a student obtained was performed.

The analysis of the results, could help in the process of making academic decisions, and develop measures to support current and future students.

CAPÍTULO I: MARCO CONTEXTUAL DE LA INVESTIGACIÓN

1.1. INTRODUCCIÓN

Se ha estimado que la cantidad de información en el mundo se duplica cada pocos meses. La automatización de las actividades de negocio produce una corriente cada vez mayor de datos, ya que incluso las transacciones simples, tales como una llamada telefónica, el uso de una tarjeta de crédito, o una prueba médica, típicamente se registran en un ordenador (Frawley & Matheus, 1992).

Los datos están siendo recogidos y acumulados a un ritmo dramático. Hay una necesidad urgente de una nueva generación de teorías y herramientas computacionales para ayudar a los seres humanos en la extracción de información útil de los crecientes volúmenes de datos digitales. Estas teorías y herramientas son el tema del emergente campo de descubrimiento de conocimiento en bases de datos (KDD) (Fayyad & Smyth, 1996).

La minería de datos, también conocida como Descubrimiento de Conocimiento en Bases de datos, es el campo que nos permite descubrir información nueva y potencialmente útil de grandes cantidades de datos.

Recientemente, se ha incrementado el interés en utilizar la minería de datos en el estudio educacional, centrándose en el desarrollo de métodos de descubrimiento que utilicen los datos de plataformas educacionales y en el uso de esos métodos para comprender mejor a los estudiantes y el entorno en el que aprenden (Jiménez & Álvarez, 2010).

Es de gran importancia para mejorar la educación superior realizar un análisis al rendimiento académico de los estudiantes, aunque el rendimiento no permite tener un entendimiento exacto nos ayuda a tener una acercamiento de la realidad (Diaz, 2002).

Uno de los problemas relacionado con la educación superior que mayor preocupación genera a diversos países, es el abandono de los estudios por parte de los estudiantes y el extenso tiempo para obtener el título universitario (Tejedor & García, 2007).

Se dice que el rendimiento académico no es el producto de una única capacidad, sino el resultado sintético de una serie de factores que actúan en, y desde, la persona que aprende. Puede afirmarse, en términos educativos, que el rendimiento académico es un resultado del aprendizaje suscitado por la actividad educativa del profesor y producido en el alumno, aunque es claro que no todo aprendizaje es producto de la acción docente (Gómez & Oviedo, 2011).

El rendimiento académico puede estar condicionado por diversos factores de toda índole, por lo cual determinar las variables más relevantes que influyen en el ámbito educativo y su incidencia en el rendimiento es muy importante (Gargallo & Pérez, 2007).

Las estrategias y estilos de aprendizaje que tienen los estudiantes, junto con factores socioeconómicos y otros influyen en su desempeño académico, dado el gran número de estudiantes que reprueban o abandonan el año en las universidades del Ecuador, esto está siendo motivo de preocupación y estudio.

El fracaso de los estudiantes en la carrera universitaria, afectan muchos entornos de su vida, por lo cual es un problema que necesita la aplicación de medidas necesarias para evitarlo y de cierta forma corregir estos casos a futuro (Herrera & Nieto, 1999).

En el marco de las observaciones anteriores se han realizado diversos estudios para determinar si los factores socioeconómicos tiene relación con el rendimiento de los estudiantes univesitarios, como el realizado por (Ruiz Herrero, 2011) en el

cual determinó que la renta per cápita del distrito, tipo de centro educativo y la profesión de sus progenitores influyen en el rendimiento.

La finalidad del presente trabajo es realizar un análisis inteligente de datos a la información almacenada por la UTEQ referente a los estudiantes, para así obtener nuevos conocimientos sobre el comportamiento estudiantil.

Las universidades necesitan nuevos recursos para entender los factores externos que intervienen en el rendimiento del estudiante, tomar las medidas oportunas y tratar de disminuir la deserción estudiantil.

Como resultado del estudio se obtendrá modelos que muestren los patrones de comportamiento de los estudiantes, de acuerdo a diferentes características de sus entornos familiares y socioeconómicos.

Actualmente en la UTEQ no se ha realizado un adecuado estudio sobre el rendimiento de los alumnos y los factores que intervienen, estos patrones podrían en un futuro ayudar a la Unidad de Planeamiento Académico de la universidad en la toma de decisiones, evitando a tiempo las falencias y realizando los correctivos necesarios.

1.2. JUSTIFICACIÓN

El bajo rendimiento académico y la deserción de los estudiantes de educación superior es un aspecto que en la actualidad es muy preocupante y motivo de mucho estudio, no solo por las instituciones de educación, también por las entidades gubernamentales, por el gasto de recursos que este puede ocasionar.

La Universidad Técnica Estatal de Quevedo cuenta con un historial de datos académicos y socioeconómicos, los cuales con un correcto análisis podrían proporcionar información útil para la tomas de decisiones académicas, para este caso se utilizó datos de los estudiantes que ingresaron a la institución en el lapso del año 2002 y 2008.

En vista de lo anteriormente expuesto, con la realización de este trabajo se busca obtener nuevos conocimientos en el ámbito educativo, que ayuden a mejorar el rendimiento de los actuales y futuros estudiantes.

El conocimiento obtenido podría servir como referencia a la Unidad de Planeamiento Académico para desarrollar medidas de apoyo a los alumnos y realizar los correctivos necesarios para mejorar el rendimiento de los estudiantes, además de servir como punto de partida para futuras investigaciones.

1.3. SITUACIÓN ACTUAL DE LA PROBLEMÁTICA

1.3.1. ANÁLISIS DEL PROBLEMA

Las Instituciones de Educación Superior tienen como función principal formar profesionales capacitados que puedan desenvolverse en la sociedad, no obstante un gran número de estudiantes desertan o pierden el año por diversos motivos, lo cual genera gran preocupación.

La deserción de los estudiantes y su rendimiento puede estar relacionado con diferentes factores, gran parte de las investigaciones se han centrado en el entorno económico sin tomar en consideración otros aspectos.

Actualmente la UTEQ no cuenta con patrones de comportamiento de los estudiantes, los cuales son reglas o características del alumno que permiten interpretar el rendimiento, tomando en cuenta los aspectos personales, sociales e institucionales que intervienen.

La gran cantidad de información almacenada por la UTEQ referente a los estudiantes, dificulta su manipulación mediante técnicas tradicionales, por lo cual realizar un análisis inteligente de datos mediante las técnicas adecuadas, ayudaría a obtener información útil sobre los patrones de comportamiento de los estudiantes.

1.3.2. FORMULACIÓN

¿Cómo utilizar técnicas de análisis inteligente de datos para la obtención de patrones de comportamiento socioeconómico de los estudiantes de la UTEQ?

1.3.3. SISTEMATIZACIÓN

- ¿Existe fiabilidad en los datos almacenados sobre las características socioeconómicas de cada estudiante en la UTEQ?

- ¿Cuáles de las técnicas difusas para el análisis inteligente de datos pueden ser utilizadas para estudiar la influencia socioeconómica sobre el rendimiento estudiantil?

- ¿Cuáles son las características socioeconómicas que más influyen en rendimiento académico de los estudiantes de la UTEQ?

1.4. OBJETIVOS

1.4.1. Objetivo General

Obtener patrones socioeconómicos y su incidencia sobre el rendimiento académico de los estudiantes de la UTEQ utilizando análisis inteligente de datos.

1.4.2. Objetivos Específicos

- Analizar el estado de la información almacenada sobre los estudiantes.

- Aplicar las técnicas difusas de análisis inteligente de datos que se ajusten a la información con las características de la Universidad.

- Evaluar los resultados obtenidos de las técnicas difusas de análisis inteligente de datos para la obtención de patrones.

1.5. HIPÓTESIS

1.5.1. Planteamiento

La aplicación de técnicas difusas para el análisis de datos socioeconómicos de los estudiantes de la UTEQ permite identificar patrones con un alto grado de influencia sobre el rendimiento académico.

1.5.2. Matriz de Conceptualización

TIPOS DE VARIABLE	DEFINICIÓN CONCEPTUAL	DIMENSIÓN DE LA VARIABLE	INDICADOR
VARIABLE INDEPENDIENTE Técnicas difusas de análisis inteligente de datos	Permiten obtener información oculta en grandes cantidades de datos.	Modelación	• Número de iteraciones • Profundidad del árbol • Número de nodos
VARIABLE DEPENDIENTE Patrones de comportamiento	Conjunto de datos y estadísticas que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones.	Eficacia	• Tasa de acierto • Número de reglas • Desviación Estándar

CAPÍTULO II: MARCO TEÓRICO DE LA INVESTIGACIÓN

2.1. FUNDAMENTACIÓN CONCEPTUAL

2.1.1. KDD: Proceso de Extracción de conocimiento

La Extracción de conocimiento está principalmente relacionado con el proceso de descubrimiento conocido como *Knowledge Discovery in Databases* (KDD), que es el conjunto de pasos realizados para obtener nuevo conocimiento útil dentro de una gran cantidad de información (Han & Kamber, 2001).

La extracción del conocimiento está conformada por una secuencia iterativa de pasos:

- **Limpieza de datos** (para eliminar valores atípicos y los datos inconsistentes)
- **Integración de datos** (donde múltiples fuentes de datos pueden ser combinados)
- **Selección de datos** (donde los datos pertinentes a la tarea de análisis se recuperan de la base de datos)
- **Transformación de datos** (donde los datos se transforman o consolidan en formas apropiadas para la minería)
- **Minería de datos** (un proceso esencial donde se aplican métodos inteligentes para extraer patrones de datos)
- **Evaluación del patrón** (para identificar los patrones verdaderamente interesantes que representan el conocimiento sobre la base de algunas medidas de intereses)

- **Presentación Conocimiento** (donde se utilizan técnicas de visualización y de representación del conocimiento para presentar el conocimiento extraído para el usuario) (Han & Kamber, 2001)

2.1.1.1 Limpieza de datos

Datos del mundo real tienden a ser incompletos, ruidoso, e inconsistente. La limpieza de datos son rutinas para intentar rellenar los valores que faltan, suavizar el ruido (Han & Kamber, 2001).

2.1.1.1.1 Datos perdidos

Para el tratamiento de los datos perdidos se pueden usar diferentes métodos, para cada caso. Los siguientes son métodos para tratar los datos faltantes.

- **Ignorar la tupla:** Esto se hace generalmente cuando la etiqueta de clase falta (asumiendo la tarea de minería implica la clasificación). Este método no es muy eficaz, a menos que la tupla contiene varios atributos con valores que faltan. Es especialmente malo cuando el porcentaje de valores perdidos por atributo varía considerablemente.

- **Rellena el valor que falta manualmente:** En general, este enfoque requiere mucho tiempo y no puede ser factible dado un gran conjunto con muchos valores que faltan datos.

- **Utilizar una constante global para rellenar el valor que falta:** Vuelva a colocar todos los valores de atributos que faltan por la misma constante, tales como una etiqueta como "Desconocido". Si los valores perdidos son reemplazados por, digamos, "Desconocido", el programa de la minería puede pensar erróneamente que forman un concepto interesante, ya que

todos ellos tienen un valor en ese campo común de "Desconocido". Por lo tanto, aunque este método es simple, no es infalible.

- **Utilizar el valor promedio:** Utiliza el valor promedio para rellenar el atributo faltante.

- **Utilice el valor más probable para completar el valor faltante:** Esto se puede determinar con la regresión, herramientas basadas en la inferencia bayesiana. (Han & Kamber, 2001).

Completar el valor faltante con el valor probable es una estrategia muy popular. En comparación con los otros métodos, se utiliza la mayoría de la información de los datos actuales para predecir los valores que faltan. Al tener en cuenta los valores de los otros atributos en su estimación del valor perdido, hay una mayor probabilidad de que se conservan las relaciones entre atributos (Han & Kamber, 2001).

2.1.1.1.2 Datos ruidosos

El ruido puede ser causado por valores aleatorios, las siguientes son métodos para el tratamiento de datos ruidosos:

- **Binning**: Métodos de intervalos para suavizar un valor de datos ordenados mediante la consulta de su entorno. Dado que los métodos de agrupación consultan los valores vecinos, realizan suavizado local.

- **Regresión**: Los datos pueden ser suavizadas por el ajuste de los datos a una función, como por ejemplo con la regresión. La regresión lineal consiste en encontrar la "mejor" línea colocar dos atributos (o las variables), de modo que un atributo puede ser usado para predecir la otra.

Regresión lineal múltiple es una extensión de la regresión lineal, en donde están implicados más de dos atributos y los datos se ajustan a una superficie multidimensional.

- **Clustering**: Los valores atípicos pueden ser detectados por la agrupación, donde los valores similares se organizan en grupos o "clusters". Intuitivamente, los valores que se encuentran fuera del conjunto de los grupos pueden ser considerados valores atípicos (Han & Kamber, 2001).

Muchos métodos de suavizado de los datos también son métodos para la reducción de datos que implica discretización. Por ejemplo, las técnicas de agrupación descritos anteriormente reducen el número de valores distintos por atributo. Esto actúa como una forma de reducción de datos para los métodos de minería de datos basados en la lógica, como la inducción de árbol de decisión, que en repetidas ocasiones hacen comparaciones de valor en los datos ordenados. (Han & Kamber, 2001)

2.1.1.1.3 Limpieza de datos como proceso

Los valores perdidos, el ruido y las inconsistencias contribuyen a los datos inexactos. El primer paso en los datos como un proceso de limpieza es la detección de discrepancia. Las discrepancias pueden ser causadas por varios factores, incluyendo las formas mal diseñados de entrada de datos que tienen muchos campos opcionales, errores humanos en la entrada de datos, errores deliberados (por ejemplo, los encuestados no quieren divulgar información sobre ellos mismos), y la decadencia de datos (por ejemplo, las direcciones obsoletas) (Han & Kamber, 2001).

Las discrepancias también pueden surgir de las representaciones de datos inconsistentes y el uso inconsistente de códigos. Los errores en los dispositivos de instrumentación que los datos de registro y errores del sistema, u otra fuente de

discrepancias. Los errores también pueden ocurrir cuando los datos son (inadecuadamente) utilizados para fines distintos a los previstos originalmente.

Como punto de partida, utilizar cualquier conocimiento que ya tenga sobre las propiedades de los datos. Tal conocimiento se conoce como metadatos.

El segundo proceso es la transformación de datos, el cual sirve para corregir discrepancias. Este proceso, sin embargo, es propenso a errores y consume mucho tiempo. Algunas transformaciones pueden introducir más discrepancias.

Algunas discrepancias anidadas sólo pueden ser detectadas después de que otros han sido corregidos. Por ejemplo, un error como "20004" en un campo de año sólo puede surgir una vez que todos los valores de fecha se han convertido a un formato uniforme. Las transformaciones se hacen a menudo como un proceso por lotes mientras el usuario espera sin realimentación. Sólo después de la transformación es completa puede que el usuario vuelva atrás y compruebe que no hay nuevas anomalías que han sido creadas por error. Típicamente, se requieren numerosas iteraciones antes de que el usuario está satisfecho.

Un enfoque para aumento de la interactividad en la limpieza de datos es el desarrollo de lenguajes declarativos para la especificación de los operadores de transformación de datos. Este tipo de trabajo se centra en definir poderosas extensiones a SQL y algoritmos que permiten a los usuarios expresar especificaciones de limpieza de datos de manera eficiente.

A medida que descubrimos más acerca de los datos, es importante mantener la actualización de los metadatos para reflejar este conocimiento. Esto le ayudará a acelerar la limpieza de datos en futuras versiones del mismo almacén de datos. (Han & Kamber, 2001)

2.1.1.2 **Integración de datos**

La minería de datos a menudo requiere de la integración de datos, fusión de datos de múltiples almacenes. Es probable que la tarea de análisis de datos implicará la integración de datos, que combina datos de múltiples fuentes en un almacén de datos coherente.

Estas fuentes pueden incluir múltiples bases de datos, cubos de datos o archivos planos. Hay una serie de cuestiones a considerar durante la integración de datos.

La integración de esquemas y objetos puede ser complicada. ¿Cómo se pueden emparejar los equivalentes a entidades del mundo real a partir de múltiples fuentes de datos? Esto se conoce como el problema de identificación de entidad. (Han & Kamber, 2001)

La redundancia es otro tema importante. Un atributo (como ingresos anuales, para ejemplo) puede ser redundante si se puede "deriva" de otro atributo o conjunto de atributos.

Algunas redundancias pueden ser detectadas por análisis de correlación. Dados dos atributos, dicho análisis puede medir la fuerza con un atributo implica la otra, sobre la base de los datos disponibles.

Se debe tener en cuenta que la correlación no implica causalidad. Es decir, si A y B están correlacionados, esto no implica necesariamente que A causa B o que B causa A.

Por ejemplo, en el análisis de una base de datos demográficos, podemos encontrar que los atributos que representan el número de hospitales y el número de robos de autos en una región están correlacionados. (Han & Kamber, 2001)

2.1.1.3 Transformación de datos

En la transformación de datos, los datos se transforman o consolidan en formas apropiadas para la minería. Transformación de datos puede incluir lo siguiente:

- Suavizado de datos: que trabaja para eliminar el ruido de los datos. Tales técnicas incluyen categorización, la regresión, y la agrupación.

- Agregación: donde se aplican operaciones de resumen o de agregación de los datos. Por ejemplo, los datos de ventas de diarios pueden ser agregados a fin de calcular las cantidades totales mensuales y anuales. Este paso se usa típicamente en la construcción de un cubo de datos para el análisis de los datos en múltiples granularidades.

- La generalización de los datos: donde de bajo nivel o "primitivo" de datos (en bruto) se sustituyen por conceptos de alto nivel a través del uso de jerarquías de conceptos. Por ejemplo, los atributos categóricos, como la calle, se pueden generalizar a conceptos de alto nivel, al igual que la ciudad o el país. Del mismo modo, los valores para los atributos numéricos, como la edad, pueden correlacionarse con conceptos de alto nivel, al igual que los jóvenes, de mediana edad y de alto nivel.

- Normalización: donde los datos de atributos se escalan para caer dentro de un pequeño rango especificado.

- Construcción de atributos: donde los nuevos atributos se construyen y se añaden desde el conjunto dado de atributos para ayudar al proceso de minería.

2.1.1.4 Minería de datos

En pocas palabras, la minería de datos se refiere a la extracción o "minería" de conocimiento de grandes cantidades de datos. Muchas personas tratan de minería de datos como sinónimo de otro término utilizado popularmente, el descubrimiento del conocimiento de datos, o KDD. Alternativamente, otros ven la minería de datos simplemente como un paso esencial en el proceso de descubrimiento de conocimiento.

El paso de la minería de datos puede interactuar con el usuario o una base de conocimientos. Los patrones interesantes se presentan al usuario y se pueden almacenar los nuevos conocimientos en la base de conocimientos.

La arquitectura de un sistema de extracción de datos típica puede tener los siguientes componentes principales:

- Base de datos: Esta es uno o un conjunto de bases de datos, almacenes de datos, hojas de cálculo u otros tipos de depósitos de información. La limpieza de datos y técnicas de integración de datos se pueden realizar sobre los datos.

- Base de conocimiento: Este es el conocimiento que se utiliza para guiar la búsqueda o evaluar el grado de interés de los patrones resultantes. Tal conocimiento puede incluir jerarquías de conceptos, utilizados para organizar los atributos o valores de atributos en diferentes niveles de abstracción.

- Motor de la minería de datos: Esto es esencial para el sistema de minería de datos y lo ideal consiste en un conjunto de módulos funcionales para tareas como la caracterización, la asociación y el análisis de correlación,

clasificación, predicción, análisis de conglomerados, análisis de las demás, y el análisis de la evolución.

- Módulo de evaluación del patrón: Este componente típicamente emplea medidas Intereses e interactúa con los módulos de minería de datos con el fin de enfocar la buscar hacia patrones interesantes. Alternativamente, el módulo de evaluación de modelo puede estar integrado con el módulo de minería, dependiendo de la implementación del método de minería de datos utilizado.

- Interfaz de usuario: Este módulo se comunica entre los usuarios y el sistema de extracción de datos, lo que permite al usuario interactuar con el sistema mediante la especificación de una consulta de la minería de datos o tarea, proporcionando información para ayudar a enfocar la búsqueda, y la realización de la minería de datos exploratorio basado en la minería de datos intermedia resultados. Además, este componente permite al usuario navegar por la base de datos y almacenamiento de datos esquemas o estructuras de datos, evaluar los patrones minados, y visualizar los patrones en diferentes formas.

Desde una perspectiva de almacenamiento de datos, minería de datos puede ser visto como una etapa avanzada de procesamiento analítico en línea (OLAP). Sin embargo, la minería de datos va mucho más allá del limitado alcance de procesamiento analítico de estilo de resumen de los sistemas de almacenamiento de datos mediante la incorporación de las técnicas más avanzadas para el análisis de datos.

La minería de datos implica una integración de técnicas de múltiples disciplinas como tecnología de base de datos y almacenamiento de datos, estadísticas, aprendizaje automático, computación de alto rendimiento, reconocimiento de

patrones, redes neuronales, visualización de datos, recuperación de información, procesamiento de imágenes y señales, datos y análisis espacial o temporal.

Mediante la realización de la minería de datos, conocimientos interesantes, regularidades, o información de alto nivel se puede extraer de las bases de datos. El conocimiento descubierto puede ser aplicado a la toma de decisiones, control de procesos, gestión de información y procesamiento de consultas.

Por lo tanto, la minería de datos es considerada una de las fronteras más importantes de los sistemas de bases de datos y de la información y uno de los desarrollos interdisciplinarios más prometedores de la tecnología de la información.

En principio, la minería de datos debería ser aplicable a cualquier tipo de depósito de datos, así como a los datos transitorios, tales como flujos de datos. (Han & Kamber, 2001)

2.1.1.4.1 Base de datos relacionales

Un sistema de base de datos, también llamado un sistema de gestión de base de datos (DBMS), consiste en una colección de datos relacionados entre sí, conocidas como una base de datos, y un conjunto de programas de software para administrar y acceder a los datos. Los programas de software implican mecanismos para la definición de las estructuras de base de datos; para el almacenamiento de datos; para compartir, o acceso a datos concurrente, distribuida; y para garantizar la coherencia y la seguridad de la información almacenada, a pesar de los fallos del sistema o intentos de acceso no autorizado.

Una base de datos relacional es una colección de tablas, cada una de las cuales se le asigna un nombre único. Cada tabla consta de un conjunto de atributos (columnas o campos) y por lo general almacena un gran conjunto de tuplas

(registros o filas). Cada tupla en una tabla relacional representa un objeto identificado por una clave única y descrita por un conjunto de valores de atributo.

Un modelo de datos semántico, tal como una entidad-relación (ER) modelo de datos, a menudo se construye para bases de datos relacionales. Un modelo de datos ER representa la base de datos como un conjunto de entidades y sus relaciones. (Han & Kamber, 2001)

Cuando se aplica la minería de datos a bases de datos relacionales, podemos ir más allá mediante la búsqueda de tendencias o patrones de datos.

2.1.1.4.2 Base de datos transaccionales

En general, una base de datos transaccional consiste en un archivo en el que cada registro representa una transacción. Una transacción típicamente incluye un número único de identidad transacción y una lista de los elementos que componen la transacción (como los artículos comprados en una tienda) (Han & Kamber, 2001).

2.1.1.4.3 Base de datos temporales

Una base de datos temporal suele almacenar datos relacionales que incluyen atributos relacionados con el tiempo. Estos atributos pueden implicar varias marcas de tiempo, cada uno con diferente semántica. (Han & Kamber, 2001)

2.1.1.4.4 Base de datos de secuencias

Una base de datos de secuencia almacena secuencias de eventos ordenadas, con o sin una noción de tiempo concreto (Han & Kamber, 2001).

2.1.1.4.5 Base de datos de series de tiempo

Una base de datos de series de tiempo almacena secuencias de valores o eventos obtenidos en mediciones repetidas de tiempo. Algunos ejemplos incluyen los datos recogidos de la bolsa de valores, control de inventarios, y la observación de los fenómenos naturales (como la temperatura y el viento) (Han & Kamber, 2001).

2.1.2. Minería de datos

La minería de datos permite obtener conocimiento que se encuentra oculto en grandes cantidades de datos, determinando las relaciones que existan entre los datos, lo cual ayudará en la toma de decisiones (Chaudhuri & Dayal, 1997).

2.1.3. Minería de Datos Educacional

Es el uso de los beneficios de la minería de datos aplicada al entorno educacional, determinando tendencias existentes en la información almacenada, para una mejor toma de decisiones (Jiménez & Álvarez, 2010).

2.1.4. Vista Minable

El concepto vista minable es aplicado a base de datos relacionales, donde se obtiene un subconjunto de datos, los cuales están en condiciones de aplicar diversos procesos de minería de datos (Jiménez & Álvarez, 2010).

2.1.5. Arquitectura de un Sistema de minería de datos

La arquitectura de un sistema de minería de datos puede estar conformada por los siguientes componentes:

2.1.6. Data Warehouse

Es la integración de datos consolidados, almacenados en un dispositivo de memoria no volátil, proveniente de múltiples y posiblemente diferentes fuentes de datos. Con el propósito del análisis y a partir de este tomar decisiones en función de mejorar la gestión del negocio. Contiene un conjunto de cubos de datos que permiten a través de técnicas de OLAP consolidar, ver y resumir los datos acorde a diferentes dimensiones de estos (Chaudhuri & Dayal, 1997).

2.1.7. Knowledge base

Este es el conocimiento del dominio que se utiliza para guiar la búsqueda o evaluar el grado de interés de los patrones resultantes. Tal conocimiento puede incluir jerarquías de conceptos, que se utilizan para organizar los atributos o valores de atributos en diferentes niveles de abstracción (Han & Kamber, 2001).

2.1.8. Data mining engine

Esto es esencial para el sistema de minería de datos y lo ideal consiste en un conjunto de módulos funcionales para tareas tales como la caracterización, la asociación y el análisis de correlación, clasificación, predicción, análisis de conglomerados, análisis de las demás, y el análisis de la evolución (Han & Kamber, 2001).

2.1.9. Pattern evaluation module

Este componente suele emplear medidas de intereses e interactúa con los módulos de minería de datos con el fin de centrar la búsqueda hacia patrones interesantes (Han & Kamber, 2001).

2.1.10. User interface

Este módulo es el encargado de realizar la comunicación entre los usuarios y el sistema de extracción de datos, lo que permite al usuario interactuar con el sistema mediante la especificación de una consulta de minería de datos o una tarea, proporcionando información para ayudar a enfocar la búsqueda (Han & Kamber, 2001).

2.1.11. Preprocesamiento de datos

Las bases de datos del mundo real de hoy son altamente susceptibles a, datos desaparecidos e inconsistentes, causando ruido debido a su normalmente gran tamaño (a menudo varios gigabytes o más) y su probable origen de múltiples fuentes. Datos de baja calidad dará lugar a resultados de la minería de baja calidad (Han & Kamber, 2001).

2.1.12. Lógica Difusa

La lógica difusa es un método matemático para el procesamiento de datos inciertos. Mediante el uso de las propiedades básicas y operaciones definidas para conjuntos difusos, cualquier estructura de reglas compuestas puede ser descompuesta y se reduce a una serie de reglas canónicas simples (Ross, 2004).

En la lógica difusa la suma de los niveles de pertenencia de un elemento debe sumar 1, es decir si un color tiene un porcentaje de iluminación de 40% debería tener un 60% de no iluminación (Puente, 1996).

Lotfi A. Zadeh fue el creador de esta teoría (Zadeh, 1965).

2.1.13. Variables Difusas

Los valores contenidos por las variables difusas son representados por términos lingüísticos. Las variables difusas permiten trabajar con medidas de incertidumbre, las cuales permiten definir de mejor forma el mundo real, ya que las características de ciertos fenómenos no son verdadero o falso, sí o no, claro u oscuro, tienen estados intermedios (Soler, 2007).

2.1.14. Reglas de clasificación IF-THEN

Las reglas son una buena manera de representar la información o bits de conocimiento. Un clasificador basado en reglas utiliza un conjunto de reglas si-entonces para la clasificación. Una regla IF-THEN es una expresión de la forma:

IF condition THEN conclusión

La parte "IF" de una regla se conoce como el antecedente de la regla o condición. La parte “THEN” es el consecuente regla.

En el antecedente de la regla, la condición consiste en una o más pruebas de atributos que se procesan con lógica AND, y el consecuente de la regla contiene una predicción de la clase (Han & Kamber, 2001).

2.1.15. Técnicas de lógica difusa

2.1.15.1. Técnicas predictivas (Pittsburg)

En el estilo Pittsburg cada individuo codifica un conjunto de reglas completo que modela el sistema analizado (Martinez, 2009).

En el caso de las técnicas predictivas estas necesitan de un conocimiento previo para generar el modelo de los datos, dicho modelo necesita ser comprobado después de realizar el proceso de minera de datos (Pérez & Santín, 2008).

2.1.15.2. Técnicas descriptivas (Michigan)

En el estilo Michigan cada individuo codifica una sola regla difusa que modela una relación de variables de entrada/salida concreta (Martinez, 2009).

En el caso de las técnicas descriptivas no se asume la presencia de variables independientes o dependientes, estas son generadas mediante la identificación de patrones (Pérez & Santín, 2008).

2.1.16. Conceptualización del nivel socioeconómico

El nivel socioeconómico de una persona es determinado mediantes varios factores, por lo general este es compartido con el resto de integrantes de una familia. El nivel socioeconómico de una persona está relacionado con diferentes factores de la familia como su grado de adquisición y entorno, el cual dista de la clase social la cual está sujeta a otros elementos (Garbanzo, 2013).

El nivel o grado socioeconómico no puede ser conceptualizado de forma clara y simple, cada nación o cultura lo puede determinar de forma diferente, por ello un único concepto y grado no puede ser válido en diferentes regiones (Garbanzo, 2013).

2.1.17. Importancia del rendimiento

La valoración del rendimiento académico en la educación superior es imprecisa, debido a que se construye con base en las notas obtenidas. Por ello, se reconoce

la necesidad de diferenciar entre el rendimiento académico inmediato, que son las notas, y el mediato, referente a los logros profesionales y personales de los estudiantes (Rodríguez F. y., 2014).

2.1.18. Software de Minería de Datos

2.1.18.1. Knime

KNIME es una moderna plataforma de análisis de datos que le permite realizar estadísticas sofisticadas y la minería de datos en sus datos para analizar las tendencias y predecir posibles resultados. Su entorno de trabajo visual combina el acceso de datos, transformación de datos, la investigación inicial, potentes análisis predictivo y visualización, también proporciona la capacidad de desarrollar informes basados en su información (Knime, 2014).

2.1.18.2. Keel

Keel es un software de código libre realizado en Java, el cual permite realizar diversas tareas de minería de datos. Una de sus principales ventajas es que posee una interfaz simple basada en el flujo de datos para diseñar los experimentos con diferentes conjuntos de datos y algoritmos. Keel posee una gran variedad de algoritmos para la extracción de conocimiento, técnicas de procesamiento de datos, métodos de imputación y muchas características más.

2.1.18.3. Weka

Weka es un software de minería de datos el cual posee herramientas para realizar las diversas etapas en el proceso de extracción del conocimiento, además de integrar un gran número de algoritmos para utilizar. Es un sistema desarollado en

Java el cual permite ser usado de forma independiente o interactuar con otros sistemas (University of Waikato, 2014).

2.1.18.4. Rapidminer Studio

RapidMiner es sin duda la más poderosa interfaz de usuario gráfica e intuitiva para el diseño de procesos de análisis. Permite la carga de cientos de datos, transformación de datos, modelado de datos, y los métodos de visualización de datos con acceso a una lista de fuentes de datos completa (Rapidminer, 2014).

2.2. MARCO REFERENCIAL

Los siguientes artículos sobre estudios realizados para determinar los factores socioeconómicos que intervienen el rendimiento académico fueron revisados como parte de la presente investigación:

2.2.1. Factores asociados al rendimiento académico en estudiantes universitarios

En esta investigación se analizan los diferentes factores socioeconómicos asociados al rendimiento académico, para el cual realizaron diversos estudios enfocados a determinados entornos. Obteniendo información relevante sobre los estudiantes universitarios y su rendimiento académico (Vargas, 2006).

2.2.2. A Case Study on Using Data Mining for University Curricula:

Los autores desarrollaron un sistema de modelado de procesos de aprendizaje, que tiene como objetivo alcanzar éxito en el aprendizaje en términos de mejor promedio. Esto se realiza mediante la aplicación de la Minería de Datos

Educacional, aplicada a los planes de estudio de los anteriores estudiantes y sus niveles de éxito obtenido. (Sakurai & Takada, 2012)

2.2.3. Determinantes del rendimiento académico en estudiantes universitarios de primer año de Economía:

En esta investigación se realiza un estudio sobre los datos recopilados mediante encuestas a los estudiantes universitarios. Se utilizó la información proporcionada por quienes se encontraban en el primer año de economía. Obteniendo como resultados la relación entre el rendimiento y factores como la asistencia y los trabajos realizados en grupo (Bartual & Poblet, 2009).

2.2.4. Factores socioeconómicos que intervienen en el desempeño académico de los estudiantes

En esta investigación se pretende determinar cómo influyen ciertas variables socioeconómicas sobre el rendimiento de estudiantes universitarios. Se realizaron encuestas a estudiantes seleccionados determinando que las variables utilizadas no intervienen en el rendimiento. (Armenta & Pacheco, 2008).

CAPÍTULO III: METODOLOGÍA DE LA INVESTIGACIÓN

3.1. MATERIALES Y METODOS

3.1.1. EQUIPOS Y MATERIALES

Los recursos materiales utilizados en la investigación fueron los siguientes:

3.1.1.1. Hardware

Tabla 1. Materiales Hardware

CANTIDAD	MATERIAL	DESCRIPCIÓN
1	Computador	Utilizado durante el proceso de investigación. - HP Pavilion g4 **Características:** - Intel Core i5 - 4Gb RAM - 640Gb HD
1	Impresora	Epson EcoTank L355

3.1.1.2. Software

Tabla 2. Materiales Software

CANTIDAD	DESCRIPCIÓN	VALOR
1	RapidMinner	$ 0.00
1	Weka	$ 0.00
1	SQL SERVER 2008 Express	$ 0.00
1	Keel	$ 0.00
1	Knime	$ 0.00
1	LibreOffice	$ 0.00

3.1.1.3. Suministros

Tabla 3. Suministros

MATERIAL	**DESCRIPCIÓN**
Internet	Internet Banda Ancha
Varios	- Anillados - Empastados - Resma de hojas

3.1.1.4. Personal

Tabla 4. Personal

PERSONAL	**DESCRIPCIÓN**
AUTOR	Rafael Vélez Franco
DIRECTOR DE TESIS	PhD. Amilkar Y. Puris Cáceres
ASESOR DE TESIS	Ing. Jorge Guanín Fajardo

3.1.1.5. Presupuesto

Tabla 5. Presupuesto

DETALLE	VALOR
Internet	$ 200.00
Impresiones	$ 60.00
Transportación	$ 200.00
Suministros oficina	$ 30.00
Empastado	$ 50.00
Anillado	$ 10.00
Total	**$ 550.00**

3.2. MÉTODOS Y TÉCNICAS UTILIZADOS EN LA INVESTIGACIÓN

3.2.1 Método inductivo

El método inductivo es un proceso en el que, a partir del estudio de casos particulares, se obtienen conclusiones o leyes universales que explican o relacionan los fenómenos estudiados (Rodríguez, 2005).

3.2.2 Método deductivo

La deducción es un proceso mental o de razonamiento que va de lo universal o general a lo particular. Consiste en partir de una o varias premisas para llegar a una conclusión (Hurtado & Toro, 2007).

Mediante este método se pudo deducir los atributos de los estudiantes de la Universidad Técnica Estatal de Quevedo que mejor definen su rendimiento académico.

3.2.3 Método descriptivo

El método descriptivo tiene como característica principal permitir determinar características primordiales del objeto de estudio y detallar cada una de sus partes (Bernal, 2010).

Este método permitió describir de forma detallada los patrones de comportamiento académico de los estudiantes de acuerdo a los atributos que definen su rendimiento.

3.2.4 Método analítico

En este método se distinguen los elementos de un fenómeno y se procede a revisar ordenadamente cada uno de ellos por separado, a partir de la experimentación y el análisis de un gran número de casos, se establecen leyes universales (Rodríguez, 2005).

Este método nos permitió estudiar los atributos de los estudiantes de forma individual, y determinar la relación que tienen con otros atributos y con el rendimiento académico.

3.3. TIPO DE INVESTIGACIÓN

3.3.1 Diseño de investigación

La metodología cuasi experimental fue utilizada en esta investigación por el hecho que desarrollamos diferentes experimentos para encontrar modelos que permitan obtener patrones de comportamiento de los estudiantes.

Posteriormente realizamos un análisis de los algoritmos los cuales permitieron obtener un alto índice de precisión y patrones de comportamiento académico que permitan obtener los atributos de un estudiante que representan en mayor medida su rendimiento.

3.3.2 Cuasi experimentos

Los diseños cuasi experimentales permiten encontrar la relación que existe entre una variable independiente y varias variables dependientes, los cuales tienen un mejor grado de seguridad que los experimentos considerados verdaderos (Hernández, 2004).

3.3.3 Pasos del cuasi experimento

Para esta investigación se realizaron varios cuasi experimentos, cada uno de ellos con un diferente subconjunto de atributos los cuales fueron obtenidos mediante algoritmos de selección. Una vez concluidos se realiza un análisis de los resultados obtenidos para determinar el subconjunto de atributos que ayuda a obtener mejores reglas.

Paso 1: Definir los atributos que se incluirán en el cuasi experimento.

Paso 2: Elegir las técnicas de procesamiento aplicadas a los datos.

Paso 3: Seleccionar los algoritmos de minería de datos a ejecutar.

Paso 5: Realizar un análisis sobre los resultados obtenidos (Hernández, 2004).

3.4. TÉCNICAS DE INVESTIGACIÓN

3.4.1 Observación Directa

La observación se utilizar como herramienta de medida en diferentes entornos, la cual consiste en realizar un registro sistemático de diversas las conductas, el cual debe ser confiable y válido (Hernández, 2004).

La utilización de esta técnica nos permitió comparar los resultados obtenidos de los diferentes algoritmos de acuerdo a sus medidas de precisión y cobertura, además de las reglas obtenidas.

3.5. POBLACIÓN Y MUESTRA

Para la presente investigación, la población estuvo conformada por los estudiantes de la Universidad Técnica Estatal de Quevedo que cursaron en el lapso de los años 2002 – 2008, la cual se encuentra almacenada en los servidores de la institución, y se trabajó con la población en su totalidad y no una muestra.

CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

4.1. RESULTADOS

En este capítulo se describirá la forma en la cual se obtuvieron los resultados de la investigación que permitieron determinar los factores socioeconómicos de los estudiantes de la UTEQ que generan gran influencia sobre el rendimiento académico.

La información recopilada fue procesada a través de diferentes sistemas de minería de datos, mediante diferentes algoritmos para la obtención de las reglas.

4.1.1. Extracción de datos

La UTEQ cuenta con gran cantidad de información sobre los estudiantes, recopilada a través de los años mediantes encuestas realizadas a los mismos. La información comprendida va desde las calificaciones obtenidas en cada uno de los años, hasta información socioeconómica como la ubicación del domicilio, número de hijos, trabajo e ingresos.

Los datos antes mencionados se encuentran almacenados en diferentes medios y sistemas de la UTEQ, por lo cual fue necesario un proceso de unificación y homogenización de los datos, realizados a través de procedimientos almacenados en los diferentes DBMS donde estaba almacenada la información.

Gran parte de la información utilizada estaba almacenada en los servidores del Sistema Informático Universitario, el cual mediante las facilidades del administrador se puedo obtener la información más relevante sobre los factores socioeconómicos de los estudiantes.

4.1.2. Procesamiento

Uno de los pasos principales al realizar una investigación es el procesamiento de los datos, ya que de esto dependerán las siguientes etapas de minería de datos y los resultados finales.

Existen gran cantidad de técnicas de preprocesamiento de datos, entre los cuales está la limpieza de los datos ruidosos, los cuales pueden generar inconsistencia en los resultados obtenidos.

Mediante procedimientos almacenados se logró conformar un conjunto de datos, con la información referente a los datos socioeconómicos de los estudiantes, la cual se detalla a continuación.

Tabla 6. Atributos generales

ITEM	ATRIBUTO	DESCRIPCIÓN	TIPO	VALOR
1	Periodo	Periodo de estudio del estudiante	categórica	{..}
2	Facultad	Nombre de la facultad de estudio	categórica	{..}
3	Carrera	Nombre de la carrera de estudio	categórica	{..}
4	Sexo	Género del estudiante	categórica	{H,M}
5	edad_ingreso	Edad de ingreso a la Universidad	numérica	{15..99}
6	sostenimiento	Sostenimiento de la institución secundaria	categórica	{FIS,PAR }
7	localizacion	Localización geográfica de la institución secundaria	categórica	{L, F , O }
8	promedio_pre	Media de las notas de admisión	numérica	{1..10}

9	promedio_1	Media de las notas del primer curso	numérica	{1..10}
10	asis_1	media de asistencia a clases en el primer año de carrera	numérica	{1..100}
11	taprobacion_1	Tiempo (cursos académicos) que tarda en aprobar el 1er curso	numérica	{1..3}
12	IndAprob_1	Índice de número de materias aprobadas sin suspensión (materias aprobadas/total materias del semestre)	numérica	{1..5}
13	promedio_2	Media de las notas del segundo curso	numérica	{1..10}
14	asis_2	media de asistencia a clases en el segundo año de carrera	numérica	{1..100}
15	taprobacion_2	Tiempo (cursos académicos) que tarda en aprobar el 2do curso	numérica	{1..3}
16	IndAprob_2	Índice de número de materias aprobadas sin suspensión (materias aprobadas/total materias del semestre)	numérica	{1..5}
17	promedio_3	Media de las notas del tercer curso	numérica	{1..10}
18	asis_3	media de asistencia a clases en el tercer año de carrera	numérica	{1..100}
19	taprobacion_3	Tiempo (cursos académicos) que tarda en aprobar el 3er curso	numérica	{1..3}
20	IndAprob_3	Índice de número de materias aprobadas sin suspensión (materias aprobadas/total materias del semestre)	numérica	{1..5}

21	promedio_4	Media de las notas del cuarto curso	numérica	{1..10}
22	asis_4	media de asistencia a clases en el cuarto año de carrera	numérica	{1..100}
23	taprobacion_4	Tiempo (cursos académicos) que tarda en aprobar el 4to curso	numérica	{1..3}
24	IndAprob_4	Índice de número de materias aprobadas sin suspensión (materias aprobadas/total materias del semestre)	numérica	{1..5}
25	promedio_5	Media de las notas del quinto curso	numérica	{1..10}
26	asis_5	media de asistencia a clases en el quinto año de carrera	numérica	{1..100}
27	taprobacion_5	Tiempo (cursos académicos) que tarda en aprobar el 5to curso	numérica	{1..3}
28	IndAprob_5	Índice de número de materias aprobadas sin suspensión (materias aprobadas/total materias del semestre)	numérica	{1..5}
29	num_hijos	¿Cuántos hijos tienen?	categórica	{0…5}
30	tamaño_familia	¿Número de integrantes de su familia?	numérica	1 ,2 3, 4 5, 6 7. más de 6 integrantes}
31	nivel_educacion_familia	¿De la familia que vive con usted, cuál es el nivel educativo más alto que se alcanzó?	categórica	{N (Ninguna), PC (Primaria Completa), PI (Primaria

				incompleta), SC (Secundaria completa), SI (Secundaria incompleta), SUC (Superior completa), SUI (Superior incompleta)} CN Estudios de 4to nivel (Posgrado)
32	media_edad_familia	¿Cuál es la edad promedio de su familia?	categórica	{1 (30 – 40), 2 (41 – 60), 3 (Más de 60}
33	ViveCon	¿Describa con quién vive usted?	categórica	{I (Soy independient), SM (Solo con Mamá), SP (Solo con Papá), AP (Ambos Padres), SPJ (Su pareja), OF (Otro familiar)}

34	jornada_trabajo	¿Cuál es su jornada de trabajo?	categórica	{NO (No trabaja), MT (Medio tiempo), TC (Tiempo Completo), EV (Eventual)}
35	total_matriculas	Total de matrículas en la Universidad	numérica	{1..15}
36	financiamiento	Sostenimiento económico del estudiante.	categórica	{AF (Ayuda familiar para 1 o 2 hijos estudiando), CP (Cuenta propia), AFM (Ayuda familiar para más de 3 hijos estudiando), CR (Préstamo o crédito vigente)}

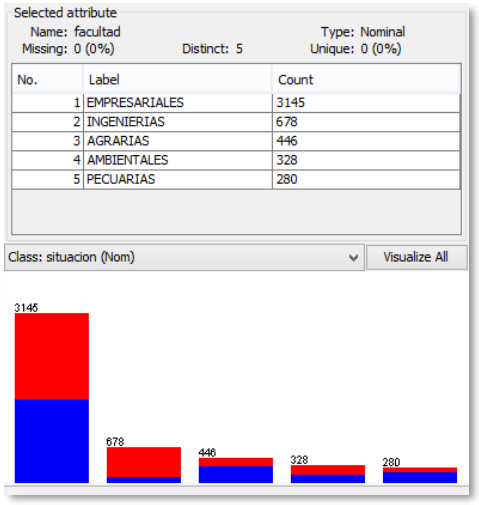
37	Zona	Situación geográfica donde reside el estudiante.	categórica	{UQ (Urbana en Quevedo), MQ (Marginal en Quevedo), FQ (Fuera de Quevedo con servicios básicos), CZR (Cualquier zona rural sin servicios básicos)}
38	Ingreso	Ingresos mensuales del estudiante o la familia con la quien vive.	categórica	{SCU (Sobre los $ USD 400), DDT (Desde $ USD 200 hasta $ USD 399), DCN (Desde $ USD 100 hasta $ USD 199), MN (Menos de $ USD 99)}

39	Vivienda	Tipo de vivienda en donde reside el estudiante.	categórica	{1 (Propia y de hormigón armado), 2 (Propia y Madera o mixta), 3 (Prestada 4. Hipotecada o Arrendada)}
40	horas_labo	Uso de las salas de internet para trabajos, investigaciones, etc relacionados al proceso de estudio	numérica	{0…320}
41	media_promedios	Media de todos los promedios		
42	media_IndAprobacion	Media de todos los índices de aprobación		
43	Situación	CLASE	categórica	{APROBADO, ABANDONA}

4.1.2.1 Análisis de los atributos

A continuación se realiza un breve análisis de los atributos presentes en el conjunto de datos, lo cual permitirá una mejor interpretación y conocimiento de la información a tratar.

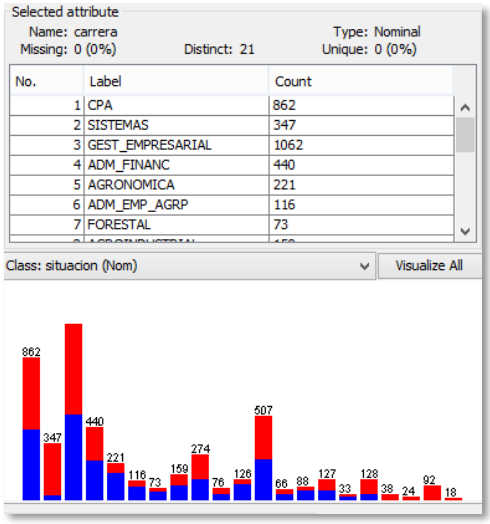
4.1.2.1.1 Atributo Facultad



En el grafico siguiente se puede observar que el atributo Facultad es de tipo nominal, posee 5 valores posibles de acuerdo a las Facultadas con las que cuenta la Universidad y que fueron objeto de estudio. No posee valores perdidos. Y podemos observar que la facultad con más registros es Empresariales y la que tiene menos Pecuarias.

Ilustración 1. Atributo Facultad

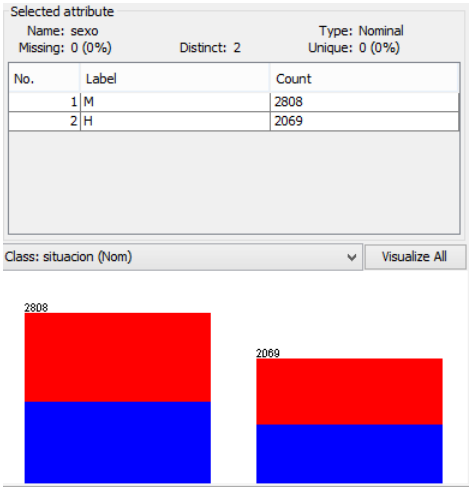
4.1.2.1.2 Atributo Carrera



El atributo carrera es de tipo nominal, cuenta con 21 valores posibles, no posee valores perdidos. La carrera con mayor cantidad de registros es Gestión Empresarial y la de menor número Industrial.

Ilustración 2. Atributo Carrera

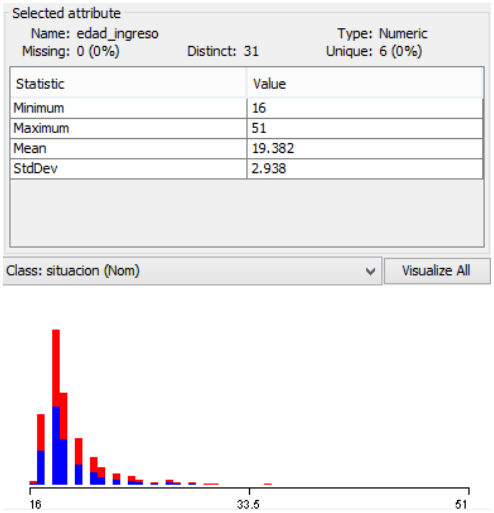
4.1.2.1.3 Atributo Sexo



El atributo sexo representa el género del estudiante, como se puede apreciar en la siguiente ilustración que no se poseen valores perdidos, y tiene 2 posibles valores M para las mujeres y H para los hombres, teniendo una mayor cantidad de registros con valor M.

Ilustración 3. Atributo Sexo

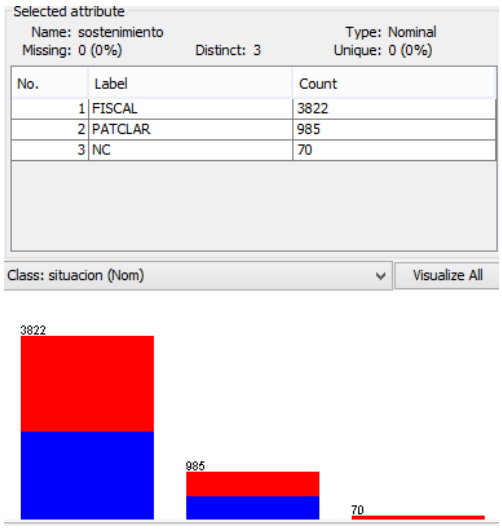
4.1.2.1.4 Atributo Edad_Ingreso



El atributo edad_ingreso es de tipo numérico y representa, la edad del estudiante al ingresar a la universidad, como se puede observar en la siguiente gráfica, los valores van desde un mínimo de 16 hasta 51 años, teniendo como promedio 19, además sin presencia de valores perdidos.

Ilustración 4. Atributo Edad_ingreso

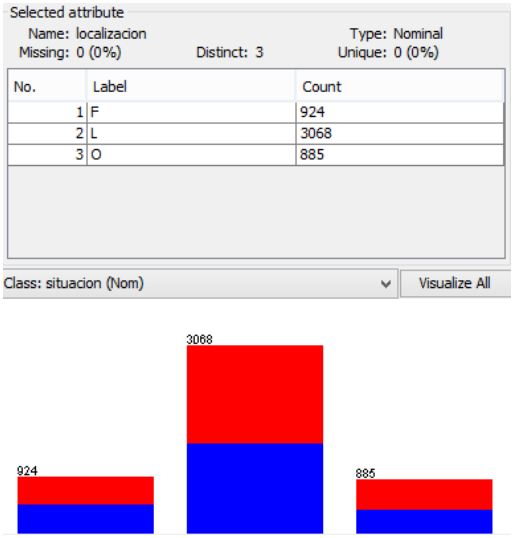
4.1.2.1.5 Atributo Sostenimiento



El atributo sostenimiento es de tipo nominal y representa el tipo de institución de la institución secundaria, y tiene 3 valores posibles Fiscal, Patclar y NC, indicando si es tipo Fiscal, Particular o no fue ingresada la respuesta.

Ilustración 5. Atributo sostenimiento

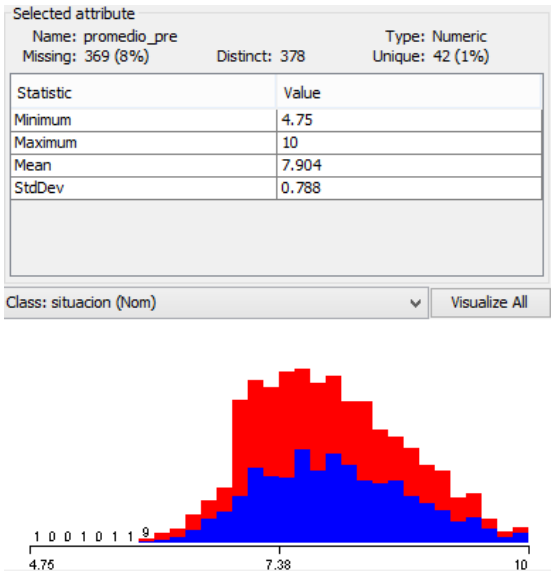
4.1.2.1.6 Atributo Localización



El atributo localización es de tipo nominal e indica la localidad del estudiantes, como se puede ver en la siguiente gráfica, tiene 3 posibles valores L, F, O, para indicar si vive en Quevedo, fuera de Quevedo y en otro lugar respectivamente.

Ilustración 6. Atributo Localización

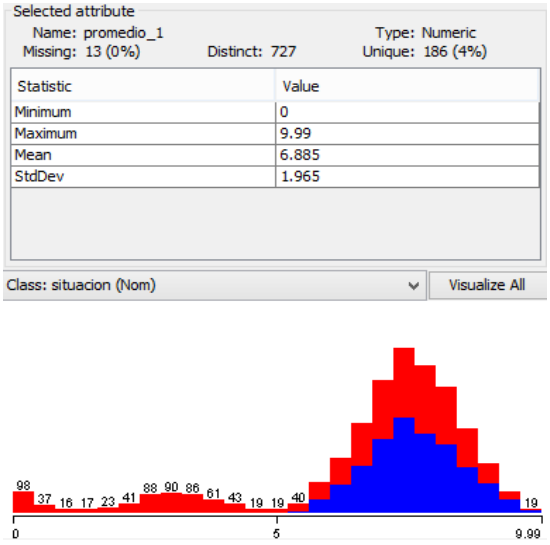
4.1.2.1.7 Atributo Promedio_pre



Este atributo representa la media de las notas obtenidas durante el proceso de admisión, siendo de tipo numérico, presentando un 8% de valores perdidos, En promedio 7.9 con una desviación estándar de 0.78, presentando valores desde un minino de 4.70 y un máximo de 10.

Ilustración 7. Atributo Promedio_pre

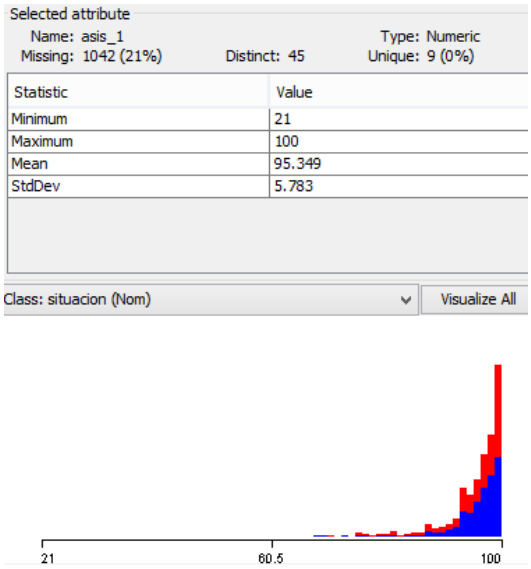
4.1.2.1.8 Atributo Promedio_1



Este atributo es de tipo numérico y represente la media de las calificaciones obtenidas en el primer año, teniendo valores entre un mínimo de 0 y un máximo de 9.99, promedio de 6.88 y una desviación estándar de 1.96 y conteniendo 13 registros con valores perdidos.

Ilustración 8. Atributo Promedio_1

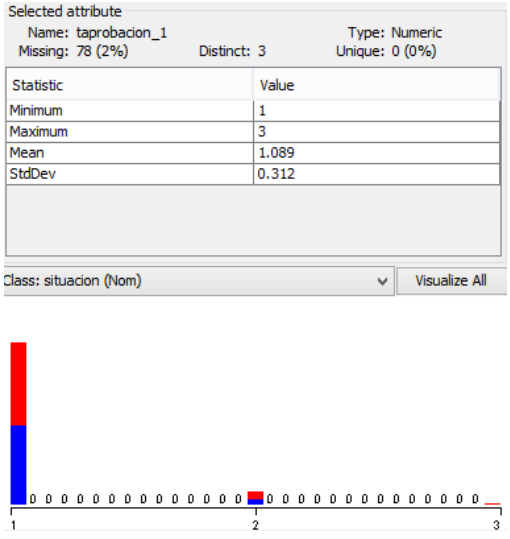
4.1.2.1.9 Atributo Asis_1



Este atributo es de tipo numérico y representa la media de asistencias a clases durante el primer año de la carrera, teniendo valores comprendidos entre un mínimo de 21 y un máximo de 100, promedio de 95, desviación estándar de 5.78, presentando un 21% de valores perdidos.

Ilustración 9. Atributo Asis_1

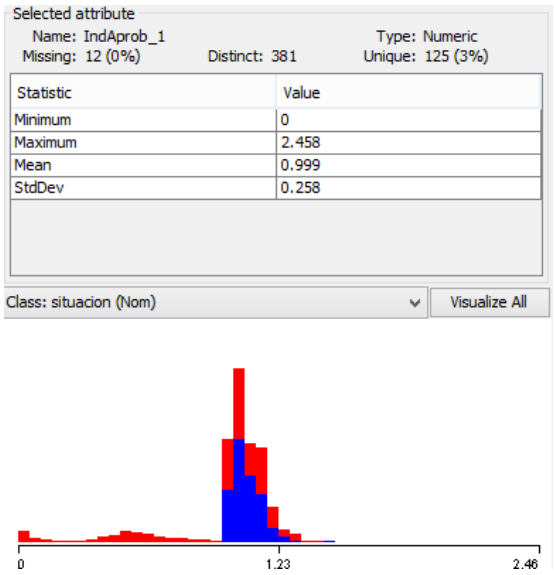
4.1.2.1.10. Atributo Taprobacion_1



Este atributo es de tipo numérico y representa el tiempo medido en cursos académicos que tarda el alumno en aprobar el primer año de la carrera, teniendo valores comprendidos entre un mínimo de 1 y un máximo de 3, promedio de 1.08, desviación estándar de 0.312, presentando un 2% de valores perdidos.

Ilustración 10. Atributo Taprobacion_1

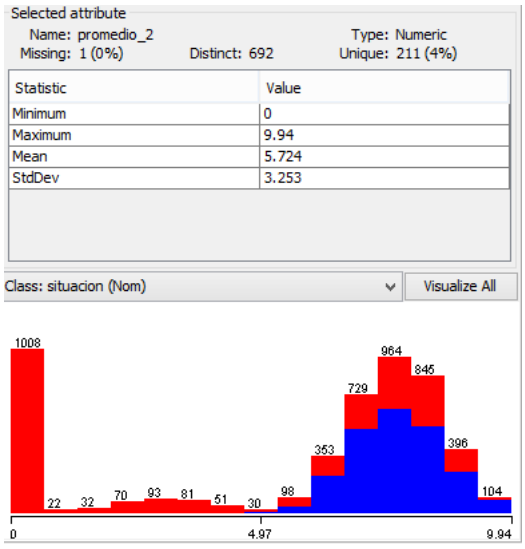
4.1.2.1.11. Atributo IndAprob_1



Este atributo es de tipo numérico y representa el índice de número de materias aprobadas sin suspensión, teniendo valores entre un mínimo de 0 y un máximo de 2.458, promedio de 0.99, presentando 12 registros con valores perdidos.

Ilustración 11. Atributo IndAprobacion_1

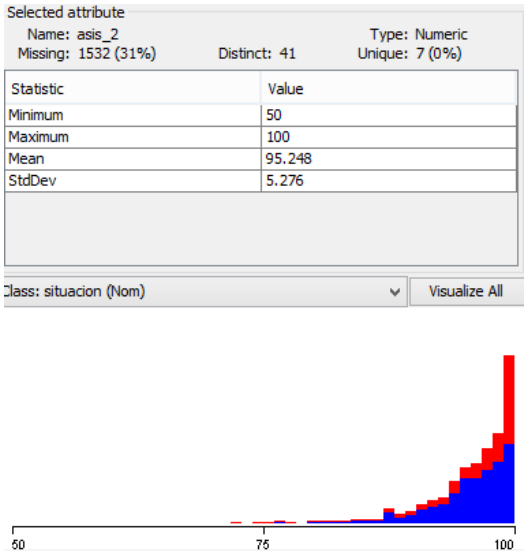
4.1.2.1.12. Atributo Promedio_2



Este atributo es de tipo numérico y represente la media de las calificaciones obtenidas en el segundo año, teniendo valores entre un mínimo de 0 y un máximo de 9.94, promedio de 5.724 y una desviación estándar de 3.25.

Ilustración 12. Atributo Promedio_2

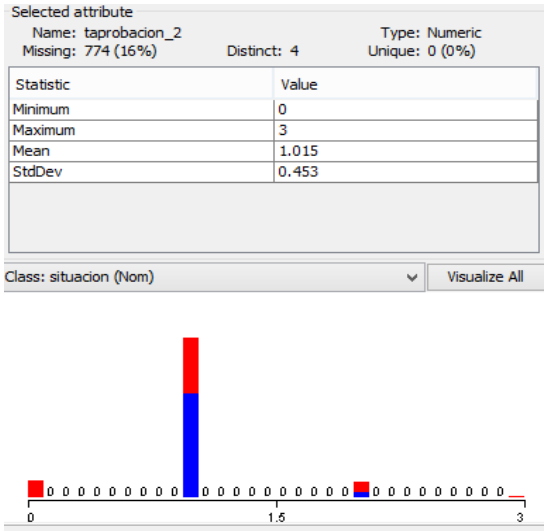
4.1.2.1.13. Atributo Asis_2



Este atributo es de tipo numérico y representa la media de asistencias a clases durante el segundo año de la carrera, teniendo valores comprendidos entre un mínimo de 50 y un máximo de 100, promedio de 95.24, desviación estándar de 5.27, presentando un 31% de valores perdidos.

Ilustración 13. Atributo Asis_2

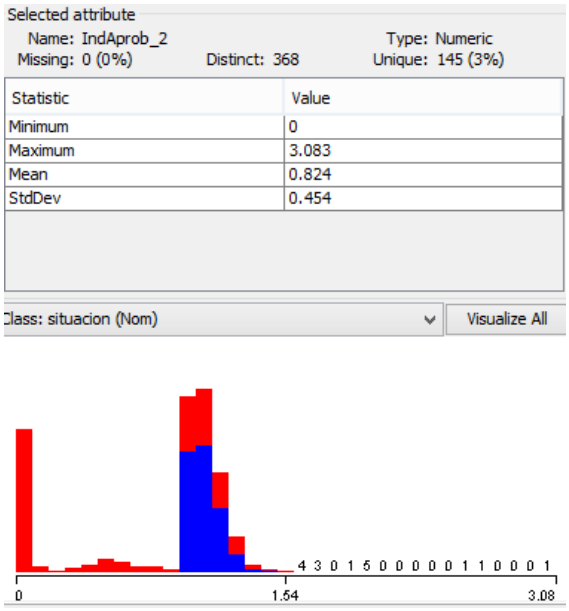
4.1.2.1.14. Atributo Taprobacion_2



Este atributo es de tipo numérico y representa el tiempo medido en cursos académicos que tarda el alumno en aprobar el segundo año de la carrera, teniendo valores comprendidos entre un mínimo de 0 y un máximo de 3, presentando un 16% de valores perdidos.

Ilustración 14. Atributo Taprobacion_2

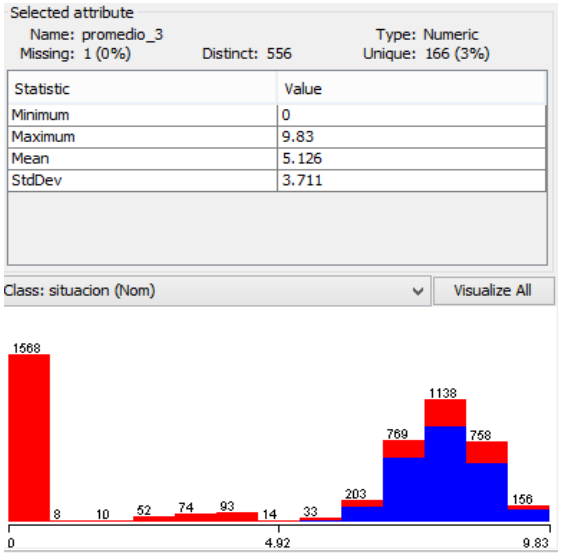
4.1.2.1.15. Atributo IndAprob_2



Este atributo es de tipo numérico y representa el índice de número de materias aprobadas sin suspensión en el segundo año de la carrera, teniendo promedio de 0.824 y desviación estándar de 0.454, presentando 0 registros con valores perdidos.

Ilustración 15. Atributo IndAprobacion_2

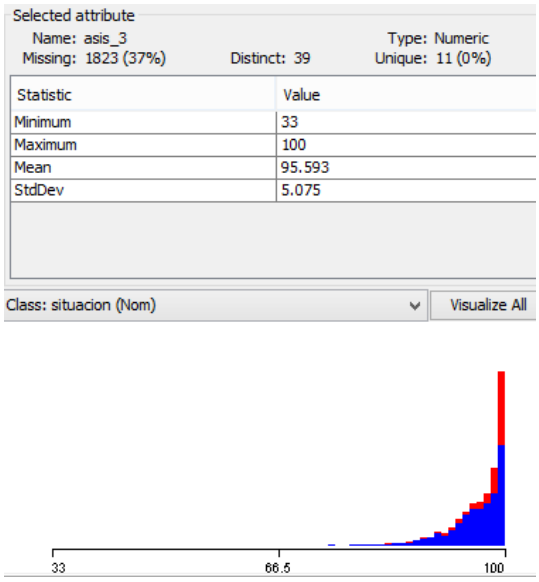
4.1.2.1.16. Atributo Promedio_3



Este atributo es de tipo numérico y represente la media de las calificaciones obtenidas en el tercer año, teniendo valores entre un mínimo de 0 y un máximo de 9.83, promedio de 5.12 y una desviación estándar de 3.

Ilustración 16. Atributo Promedio_3

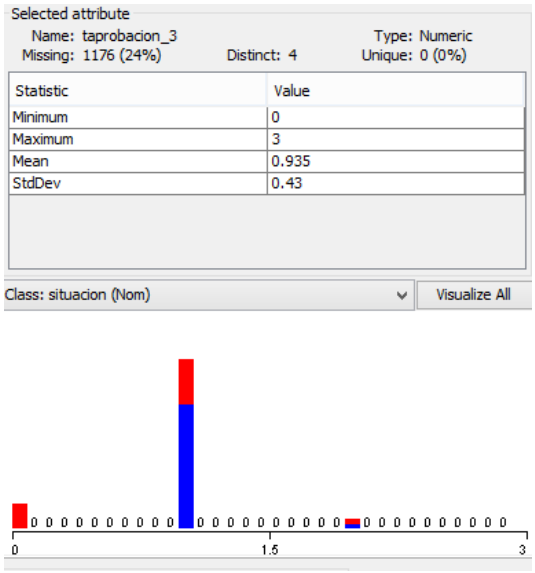
4.1.2.1.17. Atributo Asis_3



Este atributo es de tipo numérico y representa la media de asistencias a clases durante el tercer año de la carrera, teniendo valores comprendidos entre un mínimo de 33 y un máximo de 100, promedio de 95.59, desviación estándar de 5.07, presentando un 37% de valores perdidos.

Ilustración 17. Atributo Asis_3

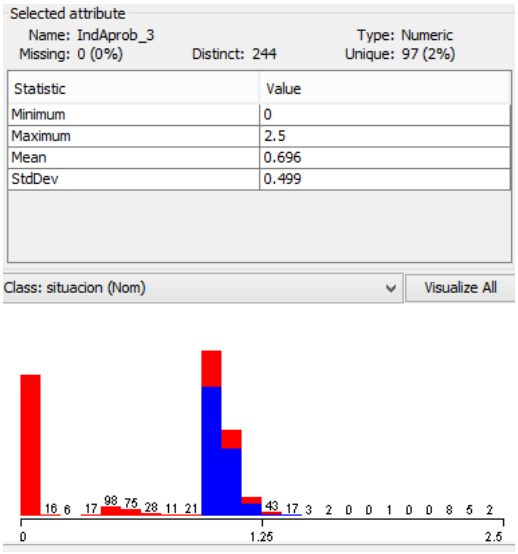
4.1.2.1.18. Atributo Taprobacion_3



Este atributo es de tipo numérico y representa el tiempo medido en cursos académicos que tarda el alumno en aprobar el tercer año de la carrera, teniendo valores comprendidos entre un mínimo de 0 y un máximo de 3, promedio de 0.935, desviación estándar de 0.43, presentando un 24% de valores perdidos.

Ilustración 18. Atributo Taprobacion_3

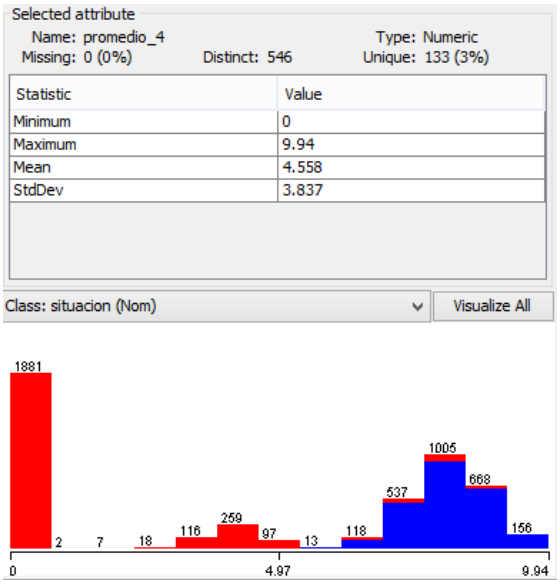
4.1.2.1.19. Atributo IndAprob_3



Este atributo es de tipo numérico y representa el índice de número de materias aprobadas sin suspensión en el tercer año de la carrera, teniendo valores entre un mínimo de 0 y un máximo de 2.5, promedio de 0.696 y desviación estándar de 0.499, presentando 0 registros con valores perdidos.

Ilustración 19. Atributo IndAprob_3

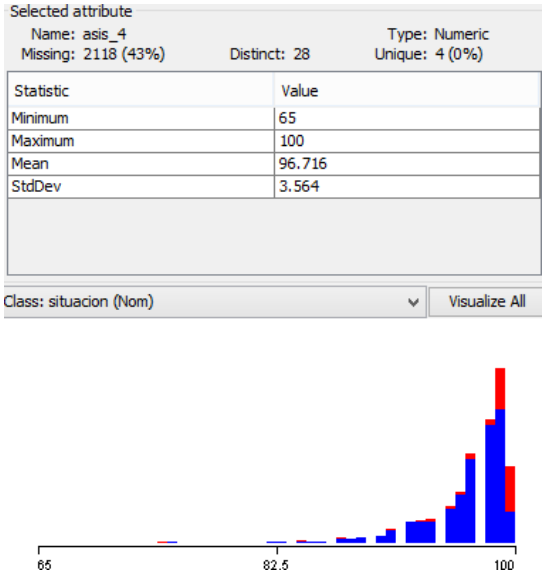
4.1.2.1.20. Atributo Promedio_4



Este atributo es de tipo numérico y represente la media de las calificaciones obtenidas en el cuarto año, teniendo valores entre un mínimo de 0 y un máximo de 9.94, promedio de 4.558 y una desviación estándar de 3.837 y conteniendo 0 registro con valores perdidos.

Ilustración 20. Atributo Promedio_4

4.1.2.1.21. Atributo Asis_4



Este atributo es de tipo numérico y represente la media de las calificaciones obtenidas en el cuarto año, teniendo valores entre un mínimo de 65 y un máximo de 100, promedio de 96.71 y una desviación estándar de 3.564 y conteniendo 43% de valores perdidos.

Ilustración 21. Atributo Asis_4

4.1.2.1.22. Atributo Taprobacion_4

Este atributo es de tipo numérico y representa el tiempo medido en cursos académicos que tarda el alumno en aprobar el cuarto año de la carrera, teniendo valores comprendidos entre un mínimo de 0 y un máximo de 3, promedio de 0.869, desviación estándar de 0.391, presentando un 31% de valores perdidos

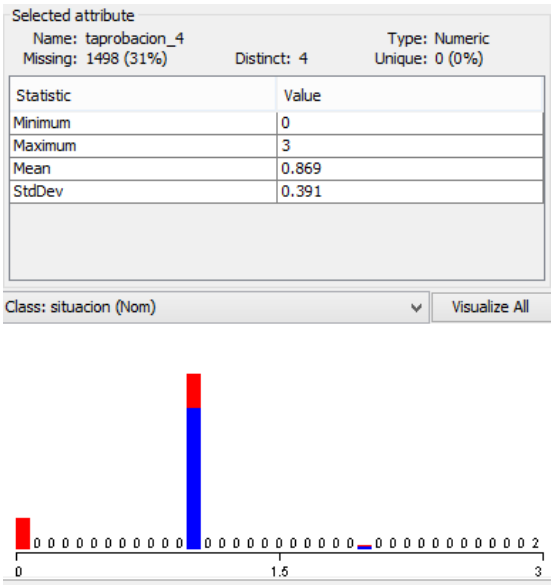
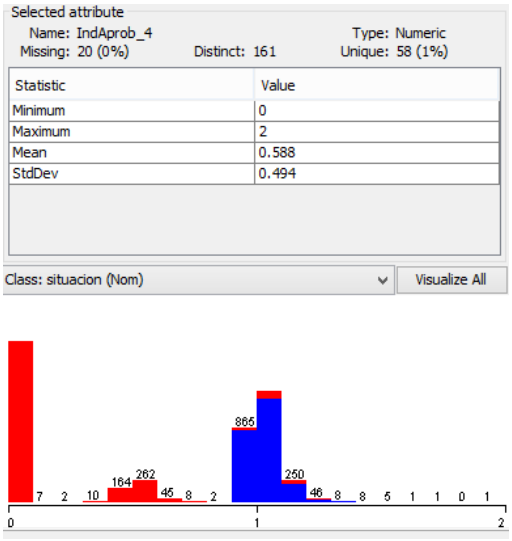


Ilustración 22. Atributo Taprobacion_4

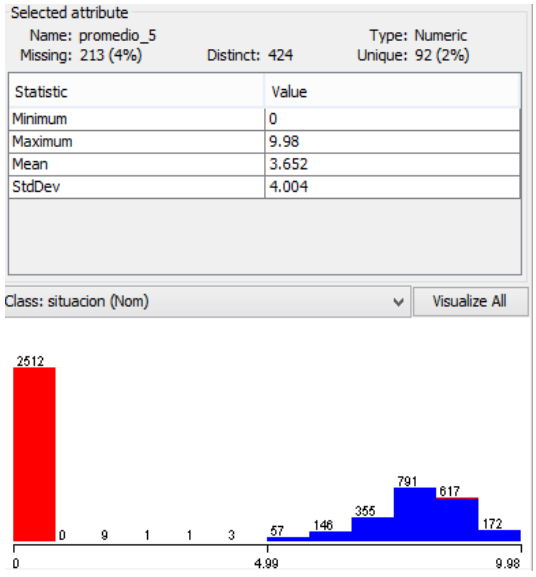
4.1.2.1.23. Atributo IndAprob_4



Este atributo es de tipo numérico y representa el índice de número de materias aprobadas sin suspensión en el cuarto año de la carrera, teniendo valores entre un mínimo de 0 y un máximo de 2, promedio de 0.588 y desviación estándar de 0.494, presentando 20 registros con valores perdidos.

Ilustración 23. Atributo IndAprob_4

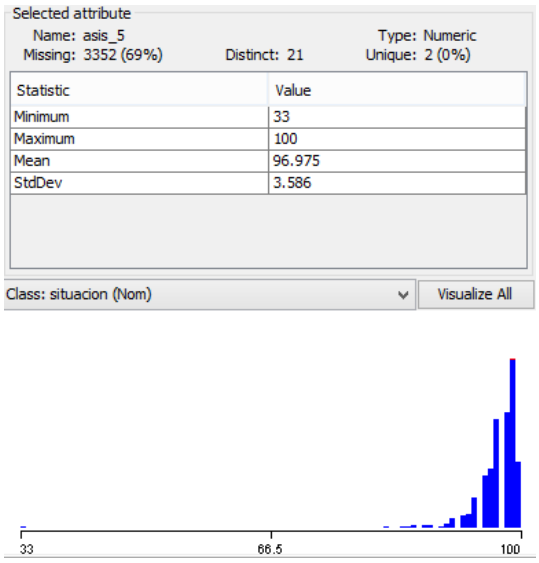
4.1.2.1.24. Atributo Promedio_5



Este atributo es de tipo numérico y represente la media de las calificaciones obtenidas en el quinto año, teniendo valores entre un mínimo de 0 y un máximo de 9.98, promedio de 3.652 y una desviación estándar de 4.004 y conteniendo 4% de valores perdidos.

Ilustración 24. Atributo Promedio_5}

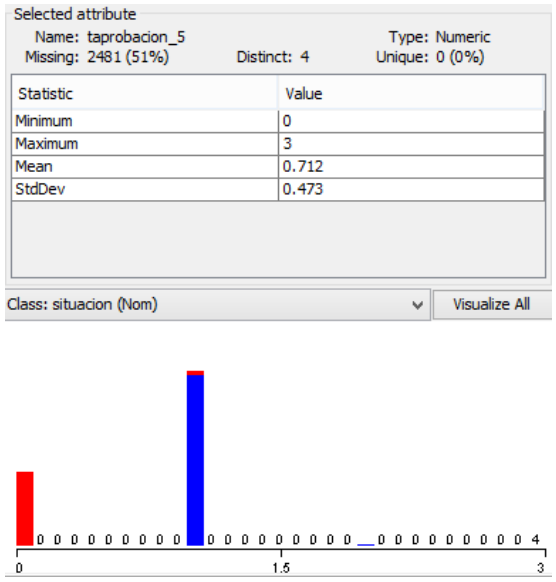
4.1.2.1.25. Atributo Asis_5



Este atributo es de tipo numérico y represente la media de las calificaciones obtenidas en el quinto año, teniendo valores entre un mínimo de 33 y un máximo de 100, promedio de 96.975 y una desviación estándar de 3.568 y conteniendo 69% de valores perdidos.

Ilustración 25. Atributo asis_5

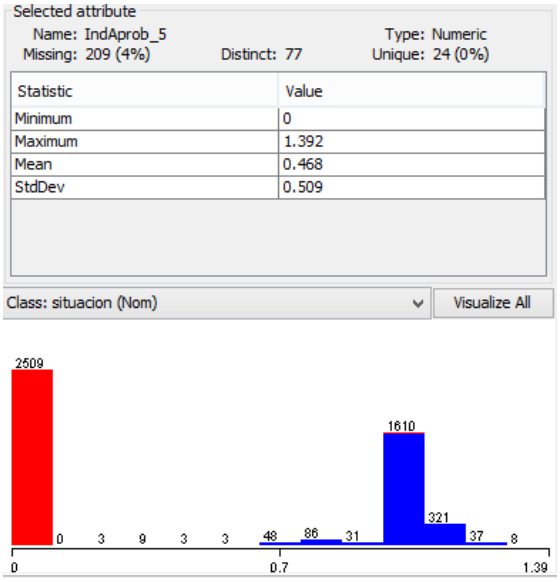
4.1.2.1.26. Atributo Taprobacion_5



Este atributo es de tipo numérico y representa el tiempo medido en cursos académicos que tarda el alumno en aprobar el quinto año de la carrera, teniendo valores comprendidos entre un mínimo de 0 y un máximo de 3, promedio de 0.712, desviación estándar de 0.473, presentando un 51% de valores perdidos.

Ilustración 26. Atributo Taprobacion_5

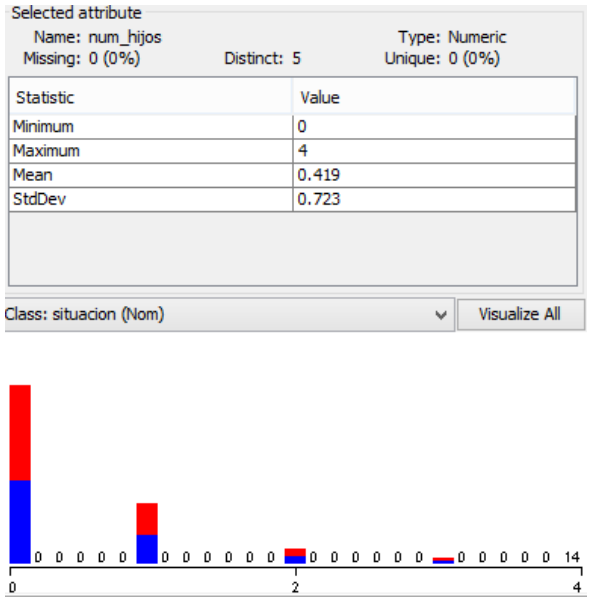
4.1.2.1.27. Atributo IndAprob_5



Este atributo es de tipo numérico y representa el índice de número de materias aprobadas sin suspensión en el quinto año de la carrera, teniendo valores entre un mínimo de 0 y un máximo de 1.392, promedio de 0.468 y desviación estándar de 0.509, presentando 4% de valores perdidos.

Ilustración 27. Atributo IndAprob_5

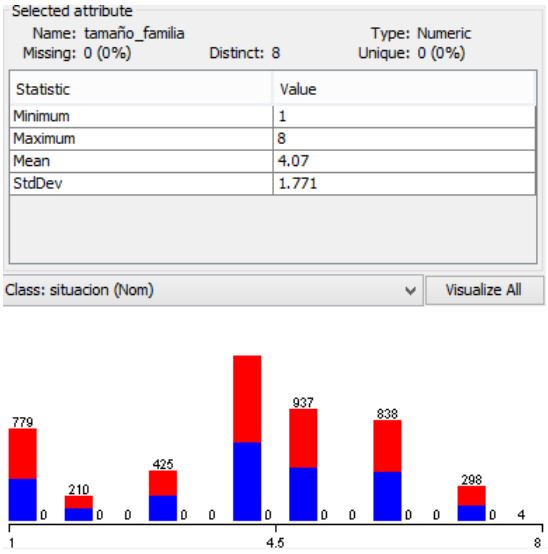
4.1.2.1.28. Atributo Num_hijos



Este atributo es numérico y representa el número de hijos que tiene el estudiante, tiene valores entre un mínimo de 0 y un máximo de 4 hijos, promedio de 0.419, una desviación estándar de 0.723 y no posee valores perdidos.

Ilustración 28. Atributo num_hijos

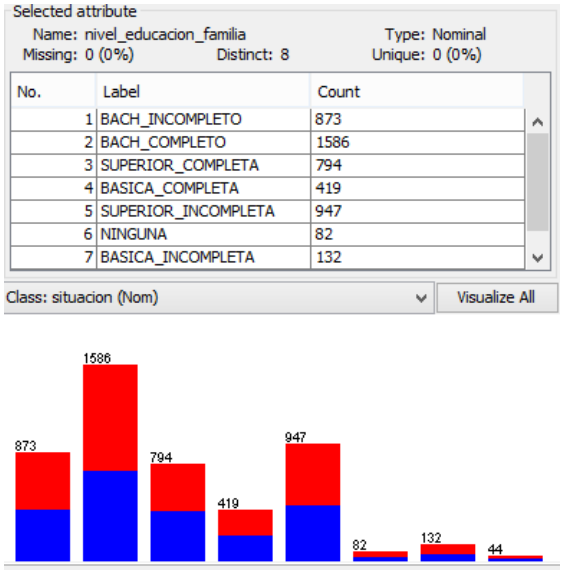
4.1.2.1.29. Atributo Tamaño_familia



En el siguiente grafico se puede observar el atributo Tamaño_familia, que es de tipo numérico y representa el número de integrantes de la familia del estudiante, tiene valores entre un mínimo de 1 y un máximo de 8 integrantes, un promedio de 4.07 integrantes y una desviación estándar de 1.77, no posee valores perdidos.

Ilustración 29. Atributo Tamaño_familia

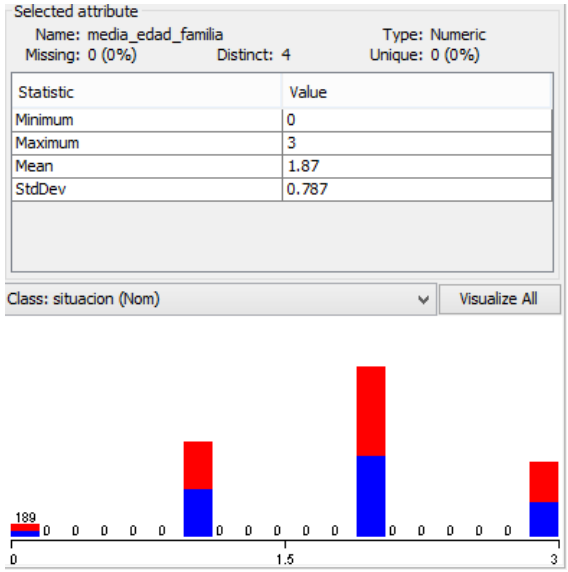
4.1.2.1.30. Atributo Nivel_educacion_familia



El atributo es de tipo nominal, puede contener 8 valores posibles, indicando el máximo grado de educación alcanzado por algún integrante de su familia, no posee valores perdidos, el valor con mayor cantidad de registros es BACHILLERATO COMPLETO y el menor CUARTO NIVEL.

Ilustración 30. Atributo Nivel_educacion_familia

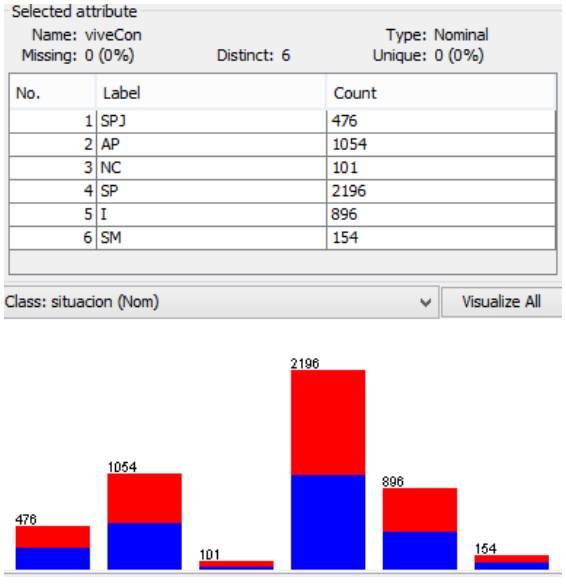
4.1.2.1.31. Atributo Media_edad_familia



Este atributo es de tipo numérico, y corresponde al promedio de edad de la familia de los estudiantes, como podemos apreciar en la siguiente gráfica, sus valores posibles oscilan entre 0 y 3, tiene una desviación estándar de 0.78, promedio de 1.87, y no presenta valores perdidos.

Ilustración 31. Atributo Edad_media_familia

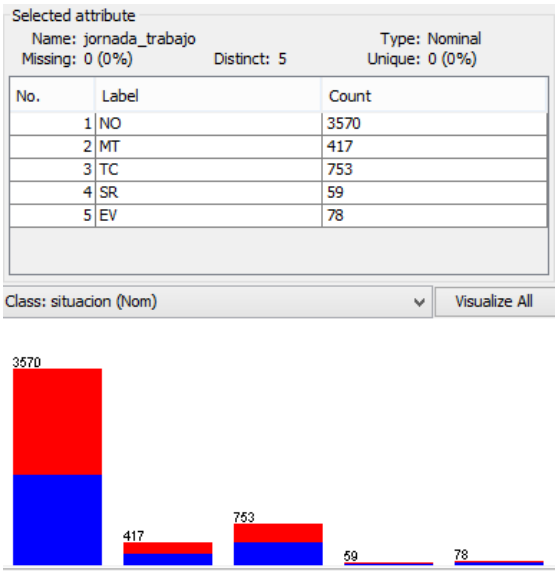
4.1.2.1.32. Atributo Vive_con



En la siguiente grafica se puede observar el atributo vive_con, el cual es de tipo nominal, puede llegar a tener entre 6 valores posibles, indicando con quien vive el estudiantes, entre las posibilidades se encuentra, independiente, ambos padres, pareja. No contiene valores perdidos.

Ilustración 32. Atributo vive_con

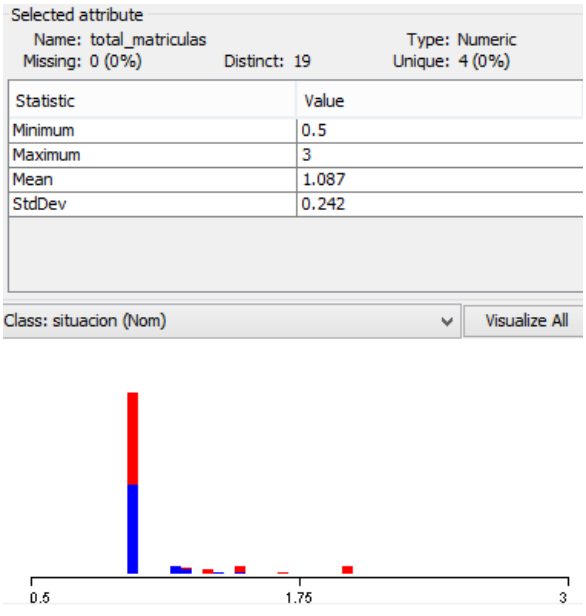
4.1.2.1.33. Atributo Jornada_trabajo



Este atributo es de tipo nominal, indica si el estudiante estudia y que jornada de trabajo tiene, los valores posibles son 6, entre los cuales encontramos, no trabaja, medio tiempo y tiempo completo. No posee valores perdidos.

Ilustración 33. Atributo Jornada_trabajo

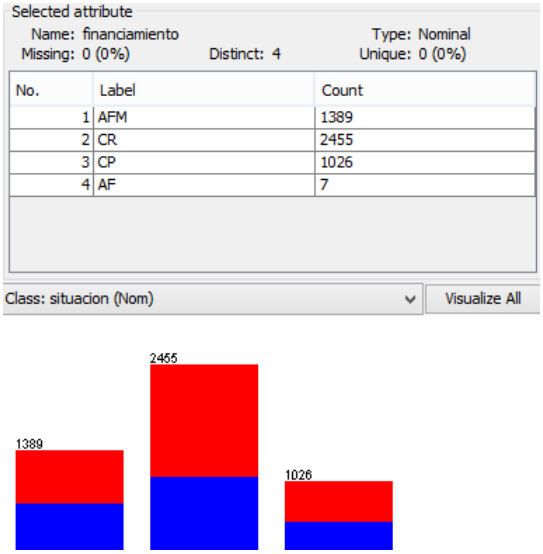
4.1.2.1.34. Atributo Total_matriculas



Este atributo es de tipo numérico y representa el número de matrículas realizadas en la Universidad, sus valores se encuentran en un mínimo de 0 y máximo de 5, una media de 1.08 y desviación estándar de 0.242, no posee valores perdidos.

Ilustración 34. Atributo Total_matriculas

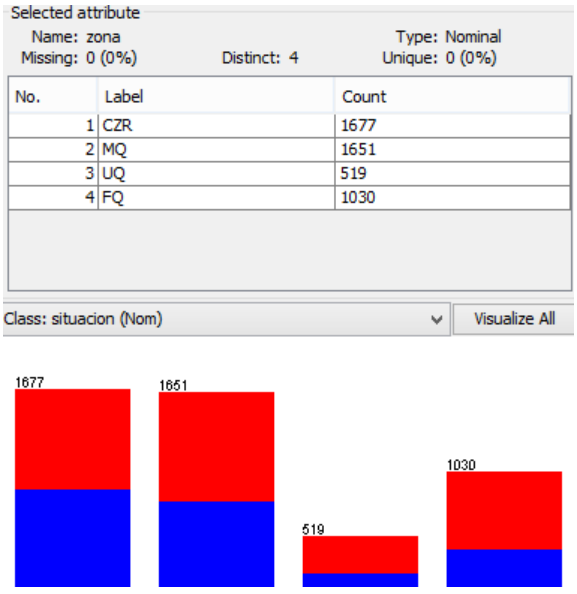
4.1.2.1.35. Atributo Financiamiento



Este atributo es de tipo nominal, tiene 4 posibles valores, indican cual es el sostenimiento económico del estudiante, entre los que encontramos ayuda familiar, cuenta propia y préstamo vigente, este atributo no contiene valores faltantes.

Ilustración 35. Atributo financiamiento

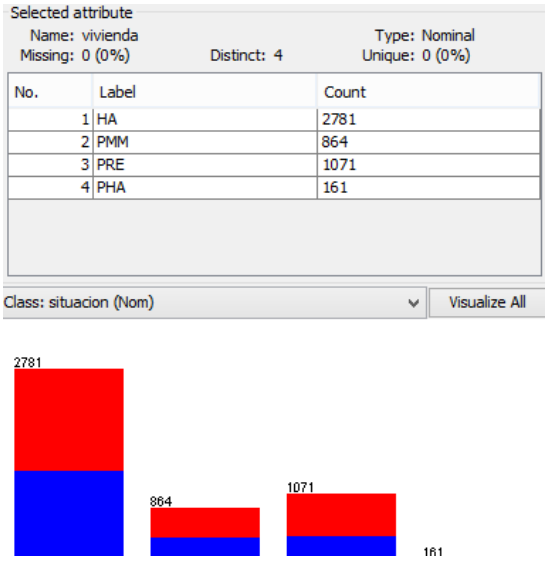
4.1.2.1.36. Atributo Zona



Este atributo es de tipo nominal, representa la ubicación geográfica donde reside el estudiante, tiene 4 valores posibles entre los que encontramos, urbana Quevedo, marginal Quevedo. No posee valores perdidos.

Ilustración 36. Atributo zona

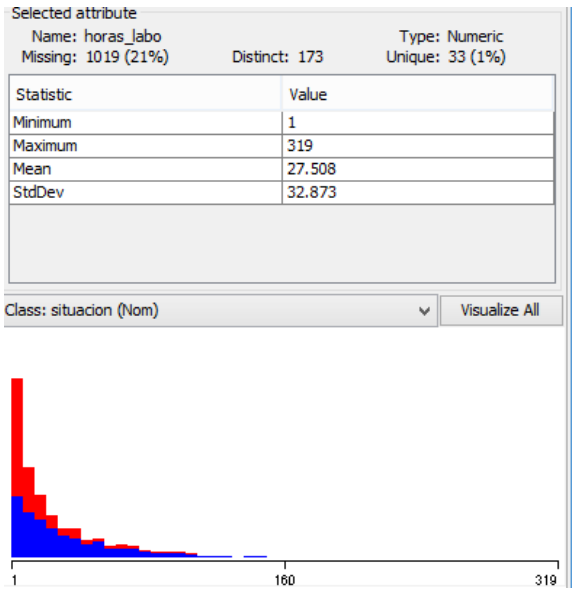
4.1.2.1.37. Atributo Vivienda



Este atributo es de tipo nominal y representa el tipo de vivienda en la que reside el estudiante, tiene 4 valores posibles, entre los que encontramos, propia de hormigón, propia de madera, prestada, hipotecada. No posee valores perdidos.

Ilustración 37. Atributo vivienda

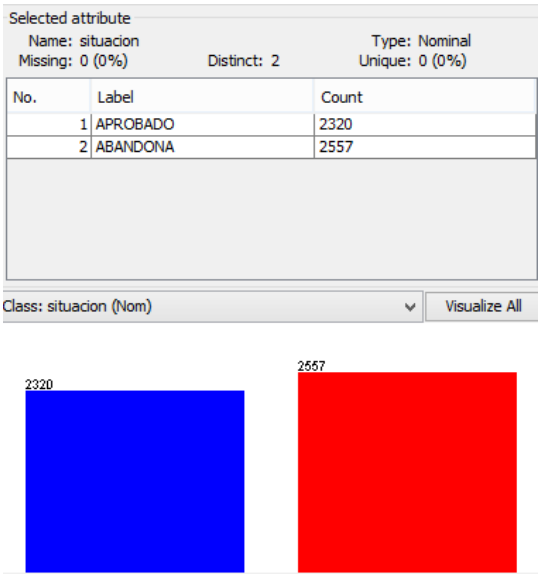
4.1.2.1.38. Atributo Horas_lab



Este atributo es numérico y representa el número de horas de salas de internet para investigaciones y trabajos. Sus valores se encuentran en un mínimo de 1 y máximo de 219, con un promedio de 27.50, desviación estándar de 32.87, tiene un 21% de valores perdidos.

Ilustración 38. Atributo Horas_lab

4.1.2.1.39. Atributo Situación



Este atributo es la clase, es de tipo nominal y nos permite conocer si el estudiante ha culminado los estudios en la universidad, tiene 2 posibles valores Aprobado y Abandona, como podemos observar el número de registros que tienen el valor de aprobado es de 2320 y el de abandona 2557.

Ilustración 39. Atributo Situación

4.1.2.2 Selección de Atributos

Posterior a la integración y combinación de las diferentes fuentes de datos, se debe realizar una selección de los atributos más relevantes, los cuales permitan obtener resultados más puntuales.

Para la realización de la investigación se logró recolectar 43 atributos relacionados a los estudiantes de la UTEQ, a los cuales se le aplicará diferentes técnicas y métodos para determinar los más significativos para el evento a estudiar.

Para agilizar este proceso se utilizó el programa Weka el cual permite mediante gráficos analizar los diferentes atributos, con la cantidad de instancias y valores nulos. Los atributos que tienen gran porcentaje de valores perdidos ocasionan ruido por lo cual debe analizar si es recomendable excluir del conjunto de datos.

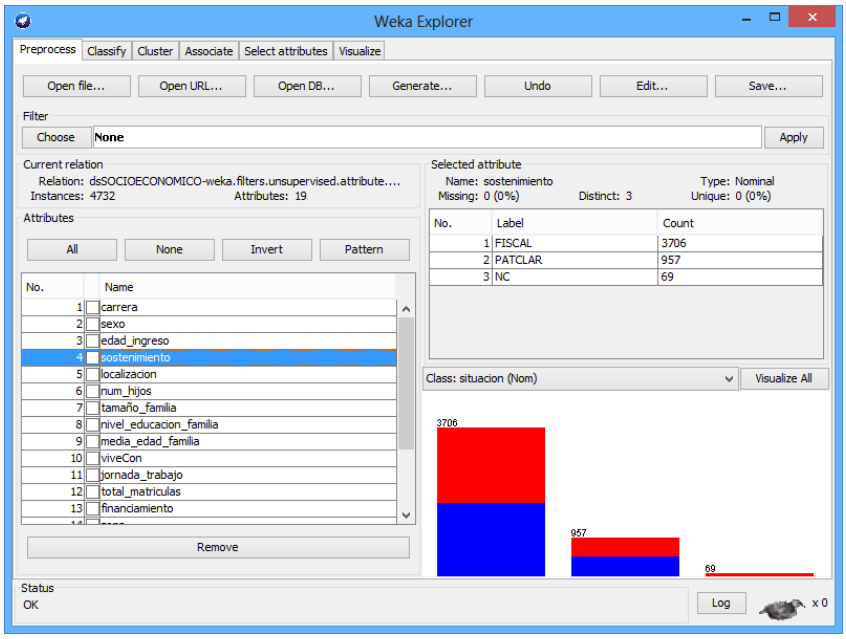


Ilustración 40 Procesamiento datos en Weka

En la imagen anterior podemos ver el análisis de los datos en el sistema Weka, el cual nos permite analizar cada uno de los atributos, y conocer sus valores, porcentaje de datos faltantes.

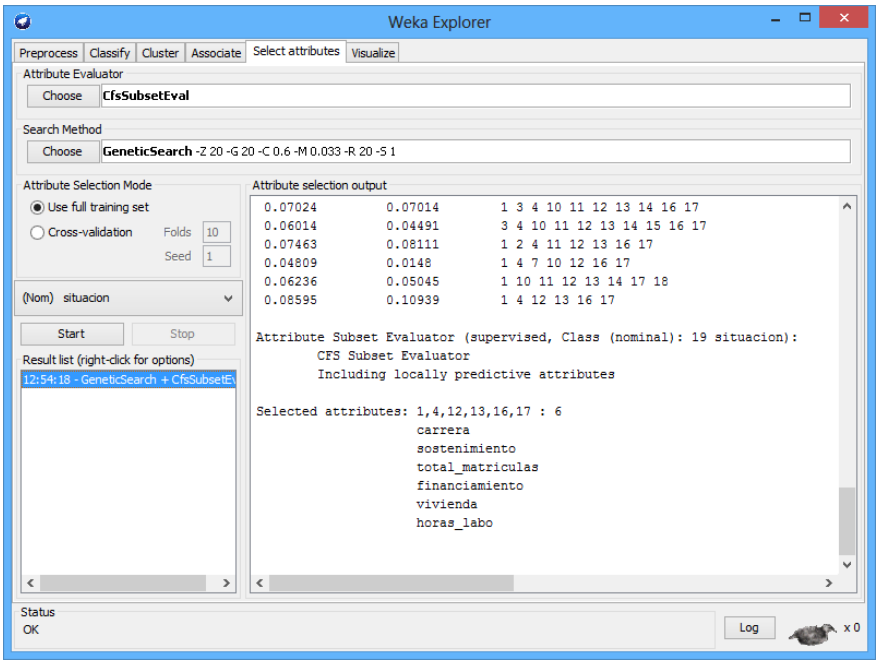


Ilustración 41 Selección atributos Weka

En la imagen anterior se puede observar, la interfaz de selección de los atributos que provee el software Weka, dando como resultados los atributos que tienen una mayor relación con la clase, en este caso la clase de la investigación es Aprobado y Abandono.

4.1.2.3 Transformación de Atributos

La transformación de datos es una etapa del KDD que permite obtener información apropiada para la minería. Entre las cuales encontramos la eliminación de datos faltantes, la generalización donde se reemplaza por concentos de alto nivel, como el caso tener un atributo llamado Calle reemplazar por Ciudad.

Otro punto importante es la normalización y creación de atributos, donde los atributos se escalan para que sus valores entren en un pequeño rango especificado, o permiten la creación de nuevos atributos que ayuden al proceso de minería de datos.

El autor de esta investigación realizó las transformaciones que permitieron un mejor análisis, el atributo “Tamaño de familia” que contiene el número de personas que integran la familia del estudiantes, por un atributo que las categoriza en pequeñas, medianas y grandes. Además de la creación de atributos como la media de los promedios y los índices de aprobación que permiten tener una mejor comprensión que los promedios de cada año.

4.1.2.4 Conjunto de datos

Luego de realizar diferentes procesos de selección de atributos, se obtuvo los siguientes conjuntos de datos, a los cuales se aplicó las técnicas de minería de datos para comprobar el grado de efectividad de los resultados obtenidos.

Tabla 7. Vista minable 1

NOMBRE	ATRIBUTOS
Vista minable 1	edad_ingreso
	sostenimiento
	localizacion
	promedio_pre
	promedio_1
	num_hijos
	tamaño_familia
	nivel_educacion_familia
	media_edad_familia
	viveCon
	jornada_trabajo
	total_matriculas
	Financiamiento
	Zona
	Ingreso
	vivienda
	horas_labo

Tabla 8. Vista minable 2

NOMBRE	ATRIBUTOS
Vista minable 2	sostenimiento
	localizacion
	promedio_pre
	promedio_1
	taprobacion_1
	IndAprob_1
	num_hijos
	tamaño_familia

Vista minable 2	nivel_educacion_familia
	media_edad_familia
	viveCon
	financiamiento

Tabla 9. Vista minable 3

NOMBRE	ATRIBUTOS
Vista minable 3	edad_ingreso
	localizacion
	num_hijos
	tamaño_familia
	nivel_educacion_familia
	media_edad_familia
	viveCon
	total_matriculas
	financiamiento
	zona
	vivienda
	horas_labo

Tabla 10. Vista minable 4

NOMBRE	ATRIBUTOS
Vista minable 4	edad_ingreso
	sostenimiento
	localizacion
	promedio_pre
	promedio_1
	asis_1
	taprobacion_1

	IndAprob_1
	num_hijos
	tamaño_familia
	nivel_educacion_familia
	media_edad_familia
	viveCon
	jornada_trabajo
	financiamiento
	Zona
	ingreso
	vivienda
	horas_labo

Tabla 11. Vista minable 5

NOMBRE	ATRIBUTOS
Vista minable 5	edad_ingreso
	sostenimiento
	localizacion
	promedio_pre
	promedio_1
	asis_1
	taprobacion_1
	IndAprob_1
	num_hijos
	tamaño_familia
	nivel_educacion_familia
	media_edad_familia
	viveCon
	jornada_trabajo
	financiamiento

	Zona
	ingreso
	vivienda
	horas_labo

Tabla 12. Vista minable 6

NOMBRE	ATRIBUTOS
Vista minable 6	facultad
	Carrera
	edad_ingreso
	Sostenimiento
	Localización
	promedio_pre
	promedio_1
	taprobacion_1
	IndAprob_1
	PIndAcademico
	promedio_2
	taprobacion_2
	IndAprob_2
	SIndAcademico
	num_hijos
	nivel_educacion_familia
	media_edad_familia
	viveCon
	jornada_trabajo
	total_matriculas
	financiamiento
	Zona
	Ingreso

Vista minable 6	Vivienda
	horas_labo
	media_promedios
	Situación

4.1.2.5 Datos faltantes

Los datos faltantes pueden ocasionar que los resultados obtenidos mediante la minería de datos no sean precisos, por lo cual es de importancia tratarlos de forma correcta para que se puede realizar la minera de datos, entre los procesos para tratarlos se encuentran, ignorar los registros con muchos valores perdidos, rellenar manualmente, utilizar una constante para el valor faltante, promedios, o el valor más probable en referencia a sus vecinos más cercanos.

Cada atributo requiere un tratamiento diferente, por lo cual el autor de esta investigación realizó diversos métodos de imputación, ignorando registros que contengan un gran porcentaje de atributos con valores desconocidos, crear constantes para valores faltantes en atributos como “jornada_trabajo”, “sostenimiento”, “zona”.

Posterior se realizó imputación mediante algoritmos basándose en el vecino más cercano, mediante el software Keel, como se puede observar en la siguiente imagen, permitiendo así completar los datos faltantes de acuerdo a los valores probables en base a registros con características similares.

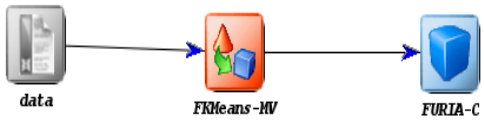


Ilustración 42. Imputación datos Keel

4.1.2.6 Discretización de datos

Se utilizó el sistema Keel para realizar una discretización de los datos, permitiendo obtener reglas de mejor calidad, al reducir el número de valores posibles de los atributos. En la siguiente imagen se puede observar el proceso de discretización de datos mediante Keel.

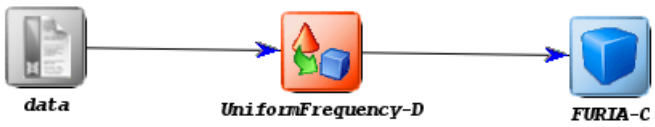


Ilustración 43. Discretización datos Keel

4.1.3 Extracción del conocimiento

Después de haber culminado con la preparación de los datos, estos se encuentran en las condiciones necesarias para aplicar los procesos de minería de datos. El autor de la investigación utilizó el software Keel para esta etapa.

El software Keel posee un conjunto de algoritmos de lógica difusa que fueron utilizaron en cada una de las vistas minables, para poder comparar los resultados obtenidos.

Para mejorar los resultados se aplicaron la técnica de sobremuestreo Oversampling, la cual nos ayuda a que los casos positivos “Aprobado” y los negativos “Abandona” tengan igual número de instancias, con lo cual el modelo no despreciará ninguno por tener pocos registros.

4.1.3.1 Resultados

Después de analizar los resultados obtenidos de aplicar los diferentes algoritmos a las vistas minables, se seleccionó la vista minable 6 por tener la mejor precisión y cobertura en las reglas obtenidas.

Los siguientes fueron los algoritmos utilizados en la investigación:

Tabla 13. Algoritmos utilizados en la investigación

Item	Algoritmo	Tipo	Enfoque	Referencia
1	FURIA (FURIA-C)	Clasificación	Fuzzy Rule Learning	(Huhn & E., 2009)
2	AdaBoost algorithm (ADABOST-C)	Clasificación	Evolutionary Fuzzy Rule Learning	(Hoffmann & Junco, 2004)
3	LogitBoost algorithm (LOGITBOOST-C)	Clasificación	Evolutionary Fuzzy Rule Learning	(Otero & Sánchez, 2006)
4	MaxLogitboost algorithm (MAXLOGITBOOST)	Clasificación	Evolutionary Fuzzy Rule Learning	(Otero & Sánchez, 2006)
5	Steady-State Genetic Algorithm (SGERD-C)	Clasificación	Evolutionary Fuzzy Rule Learning	(Mansoori & Zolghadri, 2008)
6	GP algorithm (GP-C)	Clasificación	Evolutionary Fuzzy Rule Learning	(Sánchez & Couso, 2001)
7	Hybrid Fuzzy GBML (GBML-C)	Clasificación	Evolutionary Fuzzy Rule Learning	(Yamamoto & Nakashima, 2005)
8	GAP algorithm (GAP-C)	Clasificación	Evolutionary Fuzzy Rule Learning	(Yamamoto & Nakashima, 2005)

Los resultados obtenidos de aplicar los algoritmos antes mencionados a la vista minable seleccionada fueron los siguientes:

Tabla 14. Resultados obtenidos minería de datos

ITEM	ALGORITMO	N° REGLAS	% CLASIFICACIÓN EN PRUEBAS	% CLASIFICACIÓN EN ENTRENAMIENTO
1	FURIA-C	18	97,48 ± 0,0111	99,21 ± 0,0017
2	ADABOST-C	14	70,85 ± 0.0218	73,29 ± 0,0192
3	LOGITBOOST-C	27	85,69 ± 0,0368	89,37 ± 0,0204
4	MAXLOGITBOOST-C	33	71,63 ± 0,0215	72,65 ± 0,0194
5	SGERD-C	18	71,61 ± 0,0514	72,67 ± 0,0462
6	GP-C	38	88,45 ± 0,0571	89,72 ± 0,0473
7	GBML-C	45	71,79 ± 0,0281	74,67 ± 0,0246
8	GAP-C	35	83,53 ± 0.0364	83,93 ± 0.0238

Posterior a la obtención de los resultados, se realizó el respectivo análisis de los valores obtenidos para identificar el algoritmo que nos entregará mejores reglas.

4.1.3.1.1 Análisis de resultados obtenidos

En primer lugar se analizó el algoritmo que genera un menor número de reglas, lo cual podría permitir una mejor interpretación de las reglas.

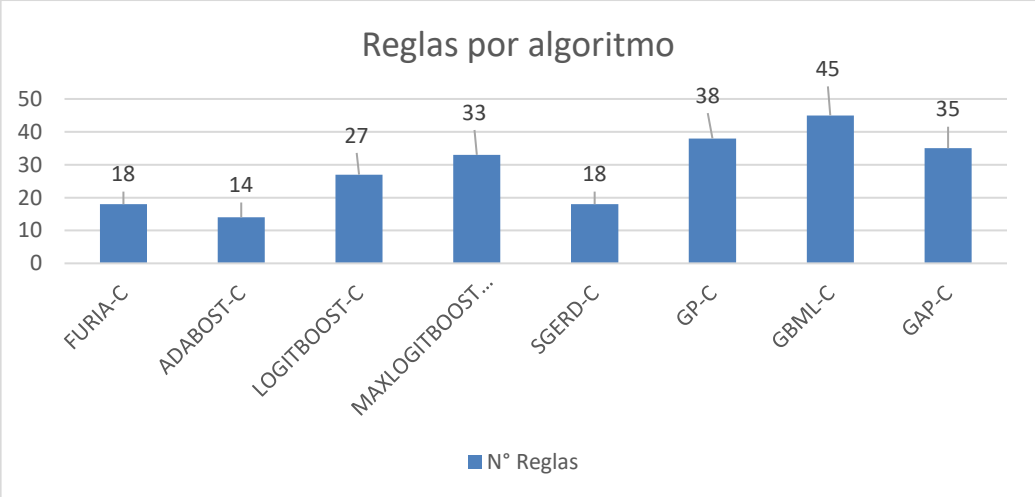


Ilustración 44: Reglas por algoritmo

Como se puede observar en el grafico 1, el algoritmo que devolvió un menor número de reglas es el ADABOST con una cantidad de 14 reglas, siguiendo los algoritmos FURIA y LOGITBOOST ambos con 18 reglas, a diferencia de los algoritmos GBML y GP que devolvieron un mayor número de reglas 45 y 38 respectivamente.

Posterior se realizó un análisis sobre la precisión de cada uno de los algoritmos obtenidos luego de aplicar a la vista minable.

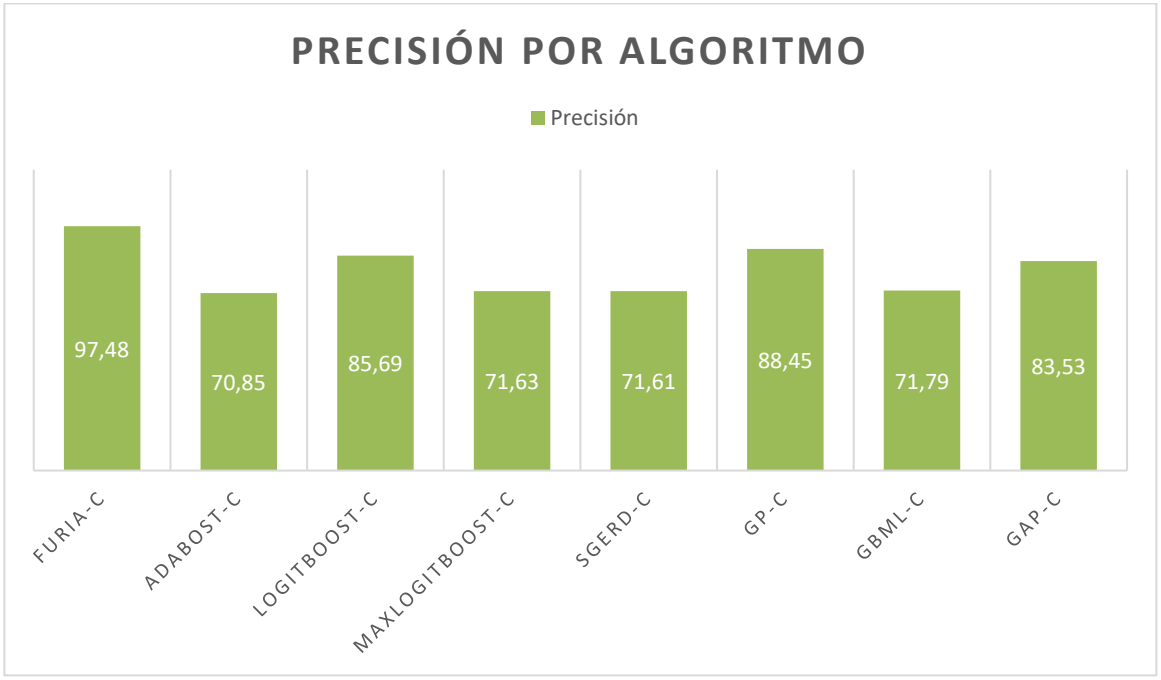


Ilustración 45: Precisión por algoritmo

Como se puede observar en el grafico 2, el algoritmo FURIA tiene el mayor grado de precisión llegando a 97.48% frente a GP y LOGITBOOST con 88.45% y 85.69% respectivamente, el algoritmo con menor precisión fue le ADABOST con un 70.85%.

A continuación se realizó el análisis de la desviación estándar de los algoritmos.

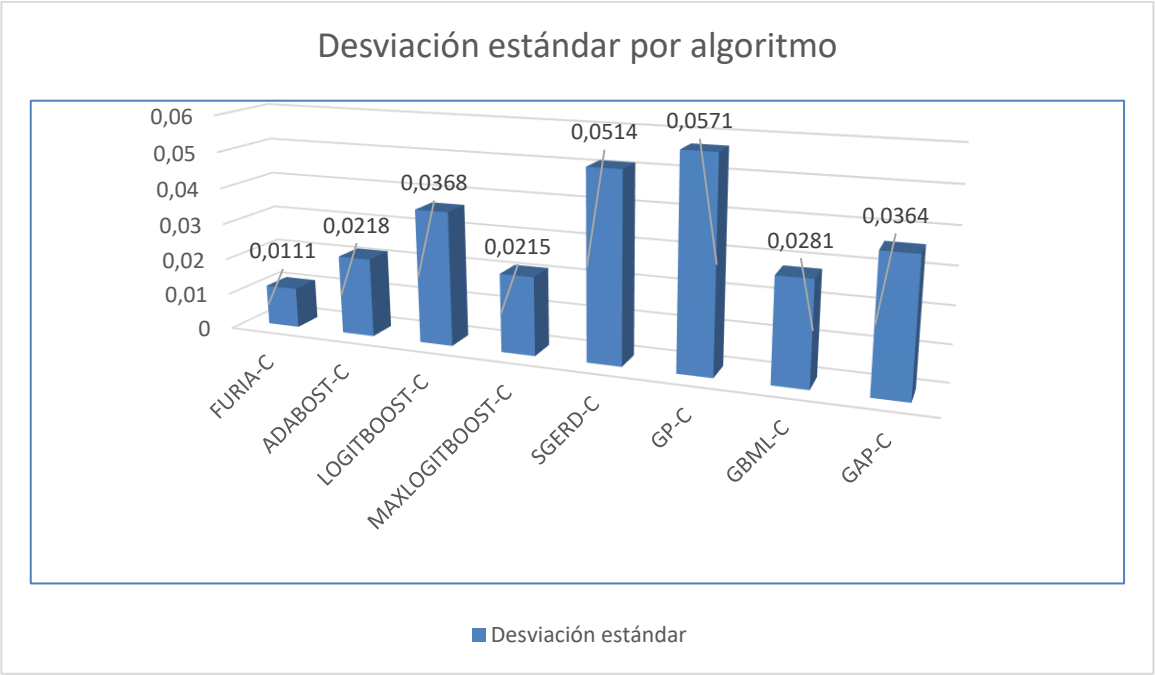


Ilustración 46: Desviación estándar por algoritmo

Como se puede observar en grafico 3, los algoritmos con menor desviación estándar son FURIA y MAXLOGITBOOST con 0.0111 y 0.0215 respectivamente.

Luego de analizar los gráficos anteriores, pertenecientes al número de reglas, precisión y desviación estándar, se puede determinar que el algoritmo ADABOST a pesar de proveer la menor cantidad de reglas sin embargo tiene el más bajo grado de precisión de los algoritmos utilizados. El algoritmo SGERD es otro de los algoritmos con menor número de reglas, pero con un alto grado de desviación estándar, por último el algoritmo FURIA tiene al cantidad de 18 reglas, la mayor precisión obtenida en la ejecución de los algoritmos y una baja desviación estándar.

Por lo antes mencionado para el autor de la investigación, el algoritmo FURIA es quien mejor resultados presenta al analiza la información socioeconómica de los estudiantes de la UTEQ.

4.1.3.2 Base de reglas

Luego de haber determinado el algoritmo que mejores resultados otorga, se analizó la base de reglas, los cual permitió obtener una mejor visión de las mismas y su relevancia, y permitiendo determinar los factores socioeconómicos que tienen mayor incidencia en el rendimiento académico.

A continuación se detallan las reglas obtenidas junto al gradó de precisión de las mismas.

Tabla 15. Reglas obtenidas mediante algoritmo FURIA

N°	DESCRIPCIÓN REGLA	PRECISÓN
R1	Si la media de todos sus promedios se encuentra entre 6.90 y 7.60 y la facultad es PECUARIA y la localización es dentro de la ciudad **entonces** Aprueba	99%
R2	Si la media de su promedios está entre 7.58 y 7.86 y vive dentro de Quevedo y el nivel de educación de su familia es BACHILLETATO COMPLETO **entonces** Aprueba	97%
R3	Si no ha realizado doble matricula y la edad media de su familia es entre 30 y 40 años y la facultad en la que estudia es EMPRESARIALES y viene de un colegio FISCAL **entonces** Aprueba	99%
R4	Si la edad de ingreso a la universidad es mayor de 20 años y estudia la carrera de ECONOMIA y tiene una media de promedios entre 6.90 y 7.60 y hace uso de los laboratorios entre 12 y 20 horas **entonces** Aprueba.	85%

R5	Si el promedio de primer año se encuentra entre 7.10 y 7.58 y vive con su pareja y la media de todos sus promedios está entre 6.90 y 7.60 **entonces** Aprueba	94%
R6	Si la media de sus promedios se encuentra entre 6.90 y 7.60 y su promedio en el pre estuvo entre 7.80 y 8.20 y el tamaño de su familia está entre 5 y 6 personas y la edad media de los integrantes de su familia entre 41 y 60 años **entonces** Aprueba	100%
R7	Si estudia la carrera CPA y la media de sus promedios se encuentra entre 6.90 y 7.60 y la edad de ingreso a la universidad es mayor a 22 años **entonces** Aprueba	96%
R8	Si el promedio de primer año se encuentra entre 7.30 y 7.62 y la media de sus promedios se encuentra entre 6.90 y 7.60 y vive en una zona marginal de Quevedo **entonces** Aprueba	97%
R9	Si la media de sus promedios se encuentra entre 6.90 y 7.60 y el promedio de su primer años está entre 6.78 y 7.11 y vive en una zona rural sin servicios básicos **entonces** Aprueba	98%
R10	Si estudia la carrera de ECONOMIA y vive de forma independiente **entonces** Aprueba	98%

R11	Si tiene una media de promedios entre 6.90 y 7.60 y el nivel de educación de su familia es EDUCACIÓN BÁSICA y el promedio de primer año fue mayor que 8 **entonces** Aprueba	97%
R12	Si trabaja a tiempo completo y no posee vivienda propia y el tamaño de su familia está entre 5 y 6 personas **entonces** Aprueba	97%
R13	Si el promedio del primer año se encuentra entre 7.10 y 7.58 y la edad media de su familia está entre 41 y 60 años **entonces** Aprueba	99%
R14	Si el promedio de segundo años está entre 6.77 y 7.19 y viene de un colegio PARTICULAR **entonces** Aprueba	98%
R15	Si la edad de ingreso a la universidad es mayor de 20 años y la carrera es CPA hace uno de los laboratorios entre 12 y 20 horas **entonces** Aprueba	85%
R16	Si el promedio de segundo año está entre 7.58 y 7.86 y el promedio del pre y vive fuera de la ciudad de Quevedo **entonces** Reprueba	76%
R17	Si el promedio de primer año se encuentra entre 7.10 y 7.58 y estudia la carrera de Marketing **entonces** Reprueba	81%
R18	Si la media de sus promedios es menor a 7.86 y estudia la carrera de SISTEMAS y vive en casa propia de hormigón **entonces** Reprueba	80%

4.2. DISCUSIÓN

Después de analizar los resultados obtenidos de la ejecución de los algoritmos aplicados en esta investigación (ver Tabla 14) se pudo determinar que el algoritmo FURIA es el que mejor se adata a los datos universitarios, obteniendo además de un alto grado de precisión un conjunto de reglas de menor tamaño.

Como se puede observar en la Tabla 15, se obtuvieron 18 reglas con el algoritmo FURIA, las cuales permiten determinar la influencia de los factores socioeconómicos de los estudiantes sobre su rendimiento académico.

Realizando un análisis a la regla denominada R1 podemos observar que si un estudiante de la facultad de Ciencias Pecuarias tiene un promedio entre 6.90 y 7.60 y vive dentro de la ciudad tiene una gran probabilidad de aprobar.

Revisando la regla R2, podemos observar que si la media de los promedios de un estudiante está entre 7.58 y 7.86 y vive dentro de Quevedo, mientras que el nivel de educación alcanzado por su familia es de BACHILLETATO COMPLETO entonces tiene un 97% de probabilidades de que Apruebe.

La regla R3 indica que si el alumno no ha realizado doble matricula y la edad media de su familia es entre 30 y 40 años, además de estudiar en la facultad de EMPRESARIALES y haber estudiado en un colegio FISCAL entonces Aprueba.

Es de resaltar en base a las reglas obtenidas no se muestra mucha relevancia al hecho de tener un trabajo, aunque no podemos asegurar que no influye en el rendimiento académico de los estudiante, conclusiones similares a las obtenidas por (Armenta & Pacheco, 2008) en un estudios realizados para determinar la influencia de factores socioeconómicos y el rendimiento estudiantil.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

Una vez finalizada la investigación realizada sobre los factores socioeconómicos que infieren sobre el rendimiento académico en los estudiantes de la UTEQ, se obtuvo las siguientes conclusiones.

- Al analizar la información almacenada sobre los estudiantes se evidenció la falta de datos socioeconómicos; para resolver este inconveniente se utilizaron las diferentes técnicas de preprocesamiento que permitan el tratamiento de ausencia de datos, con la aplicación de métodos de imputación.

- La aplicación de técnicas difusas de análisis inteligente, permitió obtener información relevante sobre los patrones en el rendimiento académico de los estudiantes.

- Se analizaron los resultados obtenidos, seleccionando el que mejor precisión aportaba al modelo con un 97% de precisión en las pruebas realizadas y determinando los factores socioeconómicos que tienen mayor influencia en el rendimiento.

- Se obtuvo un conjunto de reglas que permiten observar los patrones de comportamiento académico y la relación con el rendimiento de los estudiantes de la UTEQ (ver Tabla 15).

5.2. RECOMENDACIONES

Luego de culminar el proceso de investigación se plantean las siguientes recomendaciones:

- Continuar recopilando información por parte de la Universidad Técnica Estatal de Quevedo sobre los estudiantes, la cual mediante un correcto análisis puede servir para apoyar a diversos procesos administrativos y educativos.

- Depurar y realizar la debida documentación de la información almacenada en los sistemas de información de la UTEQ, para un mejor análisis y entendimiento en futuras investigaciones.

- Continuar con investigaciones sobre la influencia de los factores socioeconómicos en los estudiantes de la UTEQ, mediante el uso de minería de datos, para tomar medidas que ayuden a mejorar el rendimiento académico.

CAPÍTULO VI: LITERATURA CITADA

6.1. BIBLIOGRAFÍA

Alcalá R; Alcala-Fdez J; Casillas J; Cordón O; Herrera F. (2006). Hybrid learning models to get the interpretabilityaccuracy trade-off in fuzzy modeling. *Soft Comput 10(9)*, 717–734.

Álvarez, M. (1994). *Fundamentos de Inteligencia Artificial.* Murcia: Universidad de Murcia.

Álvaro Sicilia Camacho; Miguel Ángel Delgado Noguera. (2002). Educación Física y Estilos de Enseñanza. *Publicaciones: INDE. Barcelona-España*.

Armenta, N., & Pacheco, C. (2008). Factores socioeconómicos que intervienen. *REVISTA IIPSI*.

Arnold, & Blessie. (2013). Learning Styles of Teacher Education Students: Basis inproving the Teaching. *Procedia - Social and Behavioral Sciences.* (págs. 3-11). Readcube.

Ayyanathan, N., Kannammal, A., & Rekha, B. (May de 2012). Students' Communicative Competence Prediction and Performance Analysis of Probabilistic Neural Network Model. *IJCSI International Journal of Computer Science Issues, Vol. 9*, 302-317.

Bartual, T., & Poblet, M. C. (2009). Determinantes del rendimiento académico en estudiantes. *Revista de Formación e Innovación Educativa Universitaria*.

Bernadó-Mansilla E; Ho TK. (2005). Domain of competence of XCS classifier system in complexity measurement space. *IEEE Trans Evol Comput 9(1)*, 82–104.

Bernal, C. A. (2010). *Metodología de la investigación.* Colombia: PEARSON EDUCACION.

Brassard, Gilles; Bratle, Paul. (1997). *Fundamentos de Algoritmia.* Madrid : Prentice Hall.

Cano JR; Herrera F; Lozano M. (2003). Using evolutionary algorithms as instance selection for data reduction in KDD: An experimental study. *IEEE Trans Evol Comput 7(6)*, 561–575.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*.

Chellatamilan, T., Ravichandran, M., Suresh, R. M., & Kulanthaivel, G. (2011). Effect of Mining educational Data improve Adaptation of learning in e-Learning System. *Chennai and Dr.MGR University Second International Conference on Sustainable Energy and Intelligent System (SEISCON 2011)* , (págs. 922-927). India.

Dais-ujat. (2007). *Avances en Informática y Sistemas Computacionales Tomo II.* Mexico: Universidad Autónoma de Tabasco.

Diaz, P. A. (2002).

Duran, R., & Costaguta, R. (2007). Mineria de datos para descubrir estilos de aprendizaje. *Iberoamericana*, 10.

Dykinson. (2008). I Seminario sobre Sistemas Inteligentes. (pág. 256). Librería-Editorial Dykinson.

Fayyad & Simoudis. (1997). Data mining and knowledge discovery. *Proceedings of international Conference on knowledge discovery and Data mining*, (págs. 3-16).

Fayyad, U., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*.

Francisco Ascolano, Miguel Cazorla, Isabel Alfonso, Otto Colomina, Miguel Lozano. (2003). *Inteligencia Artificial, modelos, técnicas y areas de aplicacion.* España: Thonsom.

Francisco J. Martínez-López, J. C. (2010). *Sistemas inteligentes de marketing para modelado causal.* Granada, España: DELTA.

Fraw-ley, P.-S. &. (1992). Knowledge Discovery in Databases: an overview. *Magazine*, 213-228.

Frawley, W., & Matheus, C. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*.

Garbanzo, G. (2013). Factores asociados al rendimiento académico en estudiantes universitarios desde el nivel socioeconómico. *Revista Electrónica Educare*.

García, J. L.; Sánchez, C.; Jiménez, M. A.; Gutiérrez, M. (2012). Estilos de aprendizaje y estrategias de aprendizaje:. *Revista de Estilos deAprendizaje*, 65-78.

Gargallo, B., & Pérez, C. (2007). Actitudes ante el aprendizaje y rendimiento académico en los estudiantes universitarios. *Revista Iberoamericana de Educación (ISSN: 1681-5653)*, 1.

Gilles Brassard;Paul Bratle. (1997). *Fundamentos de Algoritmos.* Madrid : Prentice Hall.

Gómez, D., & Oviedo, R. (2011). Factores que influyen en el rendimiento académico del estudiante universitario. *Tecnociencia Chihuahua*, 2.

González Peiteado, M. (2009). “Estilos de enseñanzas predominantes en los alumnos de los centros de formación del profesorado de la provincia de Pontevedra. *Revista Innovación Educativa*, 237-245.

González-Peiteado, M. (2013). Los estilos de enseñanza y aprendizaje como soporte de la actividad docente. *Revista Estilos de aprendizaje*, 1-20.

Goyal, M., & Vohra, R. (May de 2012). Applications of Data Mining in Higher Education. *IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1*, 115-120.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers.

Hernández, R. (2004). *Metodología de la investigación.* MCGRAW-HILL.

Herrera, M. E., & Nieto, S. (1999). FACTORES IMPLICADOS EN EL RENDIMIENTO ACADÉMICO DE LOS ALUMNOS DE LA UNIVERSIDAD DE SALAMANCA. *Revista de Investigación Educativa*, 413.

Hurtado, I., & Toro, J. (2007). *Paradigmas y métodos de investigación en tiempos de cambio.* Caracas: CEC SA.

J. Alcalá-Fdez; L. Sánchez; S. García; M.J. del Jesus; S. Ventura; J.M. Garrell; J. Otero; C. Romero; J. Bacardit; V.M. Rivas; J.C. Fernández; F. Herrera. (2009). KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. *Soft Computing 13*, 307-318.

Jiménez, Á., & Álvarez, H. (2010).

Jimenez, M. N. (2008). *Como diagnósticar y mejorar los estilos de aprendizajes.* Santo Domingo: Asociacion Procompal.

Knauf, R., Sakurai, Y., Takada, K., & Tsuruta, S. (2010). Personalizing learning processes by data mining. *Proceedings - 10th IEEE International Conference on Advanced Learning Technologies*, (págs. 488-492).

Knime. (2014). Obtenido de https://www.knime.org/knime

Laal, K.-K., & Laal , M. (2014). Teaching and education; collaborative style. *5th World Conference on Educational Sciences - WCES 2013*, Procedia - Social and Behavioral Sciences 116 4057- 4061.

MacGregor, J. (1990). "Collaborative Learning: Shared Inquiry as a Process of Reform.". *New Directions for Teaching and Learning no. 42*, 19-30.

Martín, J. D. (2000). *Implementación de Redes Neuro-Difusas Para Per Aplicadas en Problemas de Clasificación y Modelación.* USA: Dissertation.

Martinez, F. (2009). *Sistemas Inteligentes de Marketing para modelado casual.* Madrid: Delta.

Martínez-Estudillo A; Martínez-Estudillo F; Hervás-Martínez C; García-Pedrajas N. (2006). Evolutionary product unit based neuralnetworks for regression. *Neural Netw 19*, 477–486.

Osmanbegovic, E., Agic, H., & Suljic, M. (2014). Prediction of Students'Success by Applying. *Journal of Theoretical and Applied Information Technology*, 381-387.

Parack, Suhem; Zahid, Zain; Merchant,Fatima. (2012). Application of data mining in educational databases for predicting academic trends and patterns. *In Technology Enhanced Education (ICTEE), 2012 IEEE International Conference*, (págs. 1-4).

Pérez, C., & Santín, D. (2008). *Míneria de Datos Técnicas y Herramientas.* Madrid: Paraninfo.

Pérez, G. (2004). *Modelos de investigación cualitativa.* Madrid: NARCEA SA.

Priya , K. S., & Kumar, A. (2013). Improving the Student's Performance Using Educational Data Mining. *Int. J. Advanced Networking and Applications* , 1680-1685.

Puente, P. (1996). *La teoria de subconjuntos borrosos: Aplicaciones.* Asturias: Universidad de Oviedo.

Rapidminer. (2014). Obtenido de www.rapidminer.com/products/rapidminer-studio

Rob, P., & Coronel, C. (2004). Minería de Datos. En R. Peter, & C. Coronel, *Sistemas de bases de datos: Diseño, implementación y administración* (pág. 654). Mexico.

Rodríguez. (2005). *Metodología de la investigación.* Univ. J. Autónoma de Tabasco.

Rodríguez, F. y. (2014).

Ross, T. (2004). *Fuzzy Logic with Engineering Applications.* Nuevo Mexico: Wiley.

Ruiz Herrero, J. (2011). Rendimiento academico y ambiente social. *Politica y Sociedad*, 3-9.

Sakurai, Y., & Takada, K. (2012). A Case Study on Using Data Mining for University Curricula. *IEEE International Conference on Advanced Learning Technologies*.

Singh, Chandrani; Gopal, Arpita; Mishra, Santosh. (2011). Extraction and analysis of faculty performance of management discipline from student feedback using clustering and association rule mining techniques. *In Electronics Computer Technology (ICECT), 2011 3rd International Conference*, (págs. 94-96).

Siraj, Fadzilah; Abdoulha, Mansour Ali. (2009). Uncovering hidden information within university's student enrollment data using data mining. *Third Asia International Conference*, (págs. 413-418).

Soler, V. (2007). *Lógica difusa aplicada a conjuntos imbalanceados.*

Tejedor, F. J., & García, A. (2007). Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos). Propuestas de mejora en el marco del EEES. *Revista de Educación*, 443-473.

University of Waikato. (2014). Obtenido de http://www.cs.waikato.ac.nz/ml/weka/

Vargas, G. (2006). Factores asociados al rendimiento académico en estudiantes universitarios. *Educación*.

Vialardi, C., Chue, J., Peche, J. P., Alvarado, G., Vinatea, B., Estrella, J., & Ortigosa, Á. (2011). A data mining approach to guide students through the enrollment process based on academic performance. *Springer Science+Business Media*(21), 217-248.

Xiaolong Zhang; Guirong Liu. (2008). Score Data Analysis for Pre-Warning Students in University. *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference* , (págs. 1-4).

Zadeh, L. (1965). *Fuzzy Set Information Control.*

Zhang, J. M., & Gao, W. X. (2008). Application of association rules mining in the system of university teaching appraisal. *International Workshop on Education Technology and Training and 2008 International Workshop on Geoscience and Remote Sensing, ETT and GRS*, (págs. 26-28).

Zhang, Z. (2010). Study and analysis of data mining technology in college courses students failed. *In 2010 International Conference on Intelligent Computing and Integrated Systems*, (págs. 800-802).

Zhi-min, Y., Qing, S., & Bin, S. (2010). The Analysis of Student's Grade Based on Rough Sets. *3rd IEEE International Conference on Ubi-Media Computing*, (págs. 345-349).

Zwilling, M., & Natek, S. (2013). Data Mining for small Student Data Set-Knowledge Management System for Higher Education Teachers. *Expert Systems with Applications, Vol. 41*, 1380-1388.

CAPÍTULO VII: ANEXOS

ANEXO 1. Matriz de relación entre los problemas y los objetivos.

PROBLEMA	OBJETIVO
¿Cómo utilizar técnicas de análisis inteligente de datos para la obtención de patrones de comportamiento socioeconómico de los estudiantes de la UTEQ?	Obtener patrones socioeconómicos y su incidencia sobre el rendimiento académico de los estudiantes de la UTEQ utilizando análisis inteligente de datos.
¿Existe fiabilidad en los datos almacenados sobre las características socioeconómicas de cada estudiante en la UTEQ?	Analizar el estado de la información almacenada sobre los estudiantes.
¿Cuáles de las técnicas difusas para el análisis inteligente de datos pueden ser utilizadas para estudiar la influencia socioeconómica sobre el rendimiento estudiantil?	Aplicar las técnicas difusas de análisis inteligente de datos que se ajusten a la información con las características de la universidad
¿Cuáles son las características socioeconómicas que más influyen en rendimiento académico de los estudiantes de la UTEQ?	Evaluar los resultados obtenidos de las técnicas difusas de análisis inteligente de datos para la obtención de patrones