ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE INFORMÁTICA Y ELECTRÓNICA

ESCUELA DE INGENIERÍA EN SISTEMAS

ESTUDIO COMPARATIVO DE ALGORITMOS DE PREDICCIÓN PARA LA MINERÍA DE DATOS APLICADO AL ÁREA ACADÉMICA FIE-ESPOCH

Tesis de grado previo a la obtención del título de:

INGENIERO EN SISTEMAS INFORMÁTICOS

AUTOR: KATHERINE MARIBEL GALLEGOS CARRILLO

TUTOR: ING. IVÁN MENES CAMEJO

RIOBAMBA – ECUADOR

2015

AGRADECIMIENTO

Un agradecimiento especial a la Escuela Superior Politécnica de Chimborazo por el apoyo brindado a lo largo del desarrollo de este proyecto de minería de datos, en las personas del Ing. Iván Menes y la Ing. Gloria Arcos, docentes de la Facultad, al igual que al personal a cargo de los departamento de Sistemas y Telemática de la institución.

Katherine Maribel Gallegos Carrillo

DEDICATORIA

Dedicado a mis padres William y Sonia, a mis abuelitos Elsa y Luis Gualberto a mi hermano Israel por siempre brindarme su apoyo y fortaleza, por su confianza y ayuda en cada momento.

Katherine Maribel Gallegos Carrillo

FIRMAS RESPONSABLES Y NOTAS

NOMBRE	**FIRMA**	**FECHA**
Ing. Gonzalo Samaniego, Ph. D. **DECANO DE LA FACULTAD** **DE INFORMÁTICA Y** **ELECTRÓNICA**	_____	_____
Dr. Julio Santillán **DIRECTOR DE LA ESCUELA** **DE INGENIERÍA EN SISTEMAS**	_____	_____
Ing. Iván Menes Camejo **DIRECTOR DE TESIS**	_____	_____
Ing. Gloria Arcos Medina **MIEMBRO DE TESIS**	_____	_____
COORDINADOR **SISBI- ESPOCH**	_____	_____

NOTA FINAL: _____

"Yo, Katherine Maribel Gallegos Carrillo, soy responsable de las ideas, doctrinas y resultados expuestos en el presente proyecto de tesis, y el patrimonio intelectual del mismo pertenece a la Escuela Superior Politécnica Chimborazo".

_____

KATHERINE MARIBEL GALLEGOS CARRILLO

ÍNDICE DE ABREVIATURAS

AD: Árbol de Decisión

ARTXP: Modelo de Árbol con Regresión Automática (acrónimo del inglés Autoregressive Tree Model)

ARIMA: Modelo Autorregresivo Integrado de Media Móvil (acrónimo del inglés Autoregressive Integrated Moving Average)

BI: Business Intelligence

CRISP-DM: Cross Industry Standard Process for Data Mining

DM: Data Mining

DQS: Data Quality Services

DWH: Data Ware-House

ESPOCH: Escuela Superior Politécnica de Chimborazo

FIE: Facultad de Informática y Electrónica

OLAP: On-Line Analytical Processing

OLTP: On-Line Transactional Processing

RL: Regresión Logística

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La minería de datos es un proceso mediante el cual se hace uso de la información almacenada en grandes bases de datos con el propósito de encontrar información relevante para el negocio y que aporte en la toma de decisiones. La Facultad de Informática y Electrónica de la Escuela Superior Politécnica de Chimborazo, cuenta con grandes cantidades de datos sobre los cuales se ha aplicado el presente proyecto con el objetivo principal de identificar el algoritmo de mejor desempeño entre los implementados por Microsoft para la minería de datos sujeto a la función académica de la FIE.

En el capítulo I Marco Referencial, se justifica el desarrollo del presente proyecto con la existencia de las herramientas tecnológicas que permiten automatizar gran parte de la tarea de encontrar patrones de comportamiento ocultos en los repositorios de datos y dar apoyo a la toma de decisiones basado en los datos generados por el propio negocio; se definen además los objetivos y la hipótesis de investigación.

La definición de aspectos técnicos sobre lo que es la minería de datos y los algoritmos de predicción se lo realiza en el capítulo II Marco Teórico; esto con el fin de escoger las mejores alternativas para el proyecto de minería dadas su conceptualización y aplicabilidad.

En el capítulo III Análisis Comparativo, después de definir los indicadores de comparación entre los algoritmos de Regresión Logística y Árbol de Decisión de Microsoft se hace uso de técnicas de estadística descriptiva e inferencial para decidir sobre el algoritmo de mejor desempeño entre los mencionados anteriormente para proceder al desarrollo de los modelos de minería de datos.

Finalmente en el capítulo IV Implementación de Minería de Datos se resume el proceso de desarrollo del proyecto de minería usando el algoritmo de mejor desempeño entre el de Regresión Logística y Árbol de Decisión.

1. CAPÍTULO I: MARCO REFERENCIAL

1.1. ANTECEDENTES

Grandes cantidades de datos son generadas por los diferentes tipos de negocios. Esta información está almacenada en repositorios de datos que permiten la operación diaria. La falta de aplicación de herramientas adecuadas para la explotación de los datos históricos recopilados, impide el aprovechamiento del conocimiento y patrones de comportamiento que se pueden obtener a través de la aplicación de técnicas de minería de datos.

El campo de la minería de datos aplicada a las funciones académicas ha tomado el nombre de Minería de Datos Educativa, debido a su fuerte impacto e interés continuo. Los diferentes estudios realizados buscan extraer aquellos patrones que pueden predecir el comportamiento de los estudiantes a través de la aplicación de algoritmos que permitan extraer dicha información desde datos almacenados.

Un estudio de la Universidad de Minho en Portugal, tomó los datos de los estudiantes de secundaria de dos instituciones públicas del mismo país, para realizar la aplicación de técnicas de predicción. Se probaron tres diferentes propósitos de minería y cuatro métodos de DM. Los resultados obtenidos revelaron que es posible alcanzar una alta precisión en la predicción, dados los datos de dos períodos académicos (Cortez, y otros, 2006).

Kumar y Pal, condujeron un estudio sobre el rendimiento de los estudiantes basados en un grupo de 60 alumnos de diferentes carreras de la universidad de Awadh, en India. La tarea de clasificación fue usada en la base de datos de los estudiantes para predecir la división de los mismos. Información como la asistencia, pruebas, seminarios y tareas se recolectaron para predecir el rendimiento al final del período académico (Kumar, y otros, 2011).

En la Escuela Politécnica Nacional se han realizado estudios utilizando minería de datos para la predicción meteorológica en la ciudad de Quito. En la Escuela Superior Politécnica de Chimborazo, se cuenta con los datos históricos que resultan de la función académica de la institución y no se han registrado estudios sobre la aplicación de minería de datos.

Data Mining va más allá del acceso y navegación retrospectiva de los datos, hacia una entrega de información prospectiva y proactiva. Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras: La recolección masiva de datos, potentes computadoras con multiprocesadores y algoritmos de Data Mining (Vallejos, 2006). El presente estudio está enfocado en el análisis de los algoritmos de minería de datos.

Las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, que se aplican sobre un conjunto de datos para obtener determinados resultados. Se pueden identificar tres elementos primarios en un algoritmo de minería de datos: La representación, la evaluación y la búsqueda. La representación del modelo es el lenguaje usado para representar los patrones descubribles. Los criterios de evaluación del modelo son sentencias cuantitativas de cuan bien un patrón alcanza los objetivos del proceso de descubrimiento de conocimiento. La búsqueda del método consta de dos componentes: parámetros y modelos. Una vez que se han pulido los criterios de evaluación del algoritmo, el problema de minería de datos se reduce a solamente la tarea de optimización. Es decir encontrar aquellos parámetros y modelos de los datos seleccionados que optimizan el proceso de evaluación (Fayyad, y otros, 1996).

Algoritmos como de árbol de decisión y regresión logística de Microsoft, permiten la proyección de datos a partir de orígenes de información del tipo continuo. Otros algoritmos

permiten la predicción de datos a partir de datos discretos como el algoritmo Bayes Naive. Además existen otros algoritmos orientados a cumplir diferentes tareas dentro de la minería de datos que permitan explotar el potencial hallado en los patrones dentro de las bases de datos.

En el ámbito educativo, uno de los grandes retos a los que se enfrenta la educación superior es el pronóstico de las trayectorias individuales de los estudiantes y/o de los antiguos alumnos. A las instituciones académicas les gustaría saber, por ejemplo qué estudiante necesitará ayuda para graduarse, o cómo influyen los aspectos socio-económicos de un estudiante en su proceso de aprendizaje. Éstos entre algunos problemas que pueden ser resueltos con la aplicación de minería de datos. Además, interrogantes como el tiempo que tardará un estudiante en finalizar sus estudios, continúan motivando a las instituciones de educación superior para buscar alternativas y soluciones (Jing, 2010).

En la Escuela Superior Politécnica de Chimborazo, existen grandes cantidades de datos originadas de la función docencia, más específicamente en el área académica de la institución. Estos datos pueden ser analizados mediante técnicas de Minería de Datos con el objeto de encontrar información valiosa que se convierta en herramientas para contribuir a la toma de decisiones inteligente.

Tras el diagnóstico realizado, surge la siguiente interrogante: ¿Cuál es el mejor algoritmo de predicción de minería de datos aplicado al área académica en la Facultad de Informática y Electrónica de la Escuela Superior Politécnica de Chimborazo?

Para dar respuesta a la interrogante anterior es necesario que el proceso de investigación dé respuesta a las siguientes preguntas:

- ¿En qué consisten los algoritmos de minería de datos de árbol de decisión y de regresión logística?
- ¿Cuál es aquel algoritmo que se desempeña mejor en un ambiente de pruebas?

- ¿Cuál es el resultado de aplicar un algoritmo de minería de datos a la información generada por la función académica de la FIE?

1.2. JUSTIFICACIÓN

1.2.1. JUSTIFICACIÓN TEÓRICA

La planeación para el futuro es muy importante en los negocios. Son necesarias las estimaciones futuras de las variables del negocio para la toma de decisiones de los aspectos actuales de una organización (Berson, 2000). Por su parte, la educación es un elemento crucial en la sociedad. Las técnicas de minería de datos/inteligencia de negocios, permiten una extracción del conocimiento de alto nivel de los datos en bruto y ofrecen también posibilidades interesantes para el ámbito educativo. Particularmente, muchos estudios han usado métodos de BI/DM (Bussiness Intelligence/Data Mining) para mejorar la calidad de la educación y mejorar los recursos para el nivel administrativo de las instituciones (Cortez, y otros, 2006).

Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos (Vallejos, 2006).

Entre las ventajas del Data Mining se pueden mencionar las siguientes:

- La Minería de Datos es una herramienta que permite dar respuestas a preguntas complejas de Inteligencia de Negocios.
- Las herramientas tecnológicas disponibles permiten automatizar gran parte de la tarea de encontrar patrones de comportamiento ocultos en los repositorios de datos.
- Es una forma proactiva de convertir datos en información, y esta a su vez en conocimiento, para la correcta toma de decisiones (Vallejos, 2006).

1.2.2. JUSTIFICACIÓN METODOLÓGICA

Uso de la metodología CRISP-DM para guiar el proceso de minería de datos.

La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles: fase, tarea genérica, tarea especializada, e instancia de procesos. La metodología de CRISP-DM proporciona dos perspectivas para el usuario: el modelo de referencia y la guía de usuario. El modelo de referencia presenta una descripción rápida de fases, las tareas, y sus salidas, y describen que hacer en el proyecto de minería de datos. La guía de usuario da consejos más detallados y consejos para cada fase y cada tarea dentro de una fase, y representa como realizar un proyecto de minería de datos (Chapman, y otros, 2000).

Los proyectos de minería de datos que han culminado con éxito cumplen con las directrices y etapas de CRoss-Industry Standard Process for Data Mining (CRISP-DM). Al aumentar la demanda de la minería de datos y crearse más algoritmos, CRISP-DM garantiza que todo el mundo pueda seguir las prácticas recomendadas (Jing, 2010).

1.2.3. JUSTIFICACIÓN PRÁCTICA

Un estudio en minería de datos educativa, funciona como herramienta para identificar a aquellos estudiantes que necesitan una atención especial para reducir los niveles de pérdida y tomar las medidas necesarias para una evaluación en un próximo semestre (Kumar, y otros, 2011).

Una proyección sobre el rendimiento académico de los estudiantes brindará un soporte a la toma de decisiones desde el nivel administrativo de la Facultad de Informática y Electrónica de la ESPOCH.

El desarrollo del proyecto de minería de datos, que según se describe en la metodología CRIS-DM, se componen de las siguientes fases:

- Comprensión del negocio

- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Desarrollo

Durante la Fase del Modelado de se procederá a seleccionar el algoritmo más adecuado entre árbol de decisión y regresión logística de Microsoft. Para ello se elaborarán 2 ambientes de prueba basados en los datos académicos de los estudiantes de la Facultad de Informática y Electrónica.

Las mediciones tomadas permitirán determinar el algoritmo de predicción de mejor desempeño para la proyección de datos académicos y continuar con la fase de desarrollo de la aplicación BI.

La alineación del presente proyecto se presenta en la Tabla 1 como sigue:

TABLA 1: ALINEACIÓN DEL PROYECTO

ESPOCH	LINEA: Tecnologías de la información, comunicación y procesos industriales
	PROGRAMA: Programa para el desarrollo de aplicaciones de software para procesos de gestión y administración pública y privada. Educación.
PNVB	OBJETIVO: Mejorar las capacidades y potencialidades de la población
	POLÍTICA: Promover el acceso a la información y a las nuevas tecnologías de la información y comunicación para incorporar a la población a la sociedad de la información y fortalecer el ejercicio de la ciudadanía.

Fuente: Líneas y Programas de investigación de la ESPOCH; Plan Nacional del Buen Vivir, 2009

1.3. OBJETIVOS

1.3.1. OBJETIVO GENERAL

Analizar los algoritmos de predicción de minería de datos aplicado a la función académica de la FIE-ESPOCH.

1.3.2. OBJETIVOS ESPECÍFICOS

- Estudiar los algoritmos de predicción: árbol de decisión y de regresión logística.
- Establecer los parámetros de comparación para los algoritmos de minería de datos.
- Preparar un escenario para la comparación de los algoritmos de minería de datos.
- Aplicar el algoritmo de predicción de minería de datos seleccionada sobre los datos de la función académica de la FIE-ESPOCH.

1.4. HIPÓTESIS

El algoritmo de minería de datos de regresión logística tiene mejor desempeño que el algoritmo de árbol de decisión para obtener datos proyectados sobre las actividades académicas de la Facultad de Informática y Electrónica de la ESPOCH.

2. CAPÍTULO II: MARCO TEÓRICO

2.1. DATA MINING

Se sugiere que la minería de datos surge a la par del almacenamiento masivo de datos, pues es desde ese momento en el cual surge la incertidumbre con respecto a aquella información que se encuentra escondida en dichos almacenes de datos.

2.1.1. ¿QUÉ ES DATA MINING?

El proceso de minería de datos está definido por Jamie Macklien como sigue: Data Mining es el proceso de analizar datos usando metodologías automatizadas para encontrar patrones escondidos (MacLennan, 2008).

Por su parte, Kumar y Pal en su documento científico hacen referencia a la minería de datos como: la extracción de conocimiento de grandes cantidades de datos (Kumar, y otros, 2011).

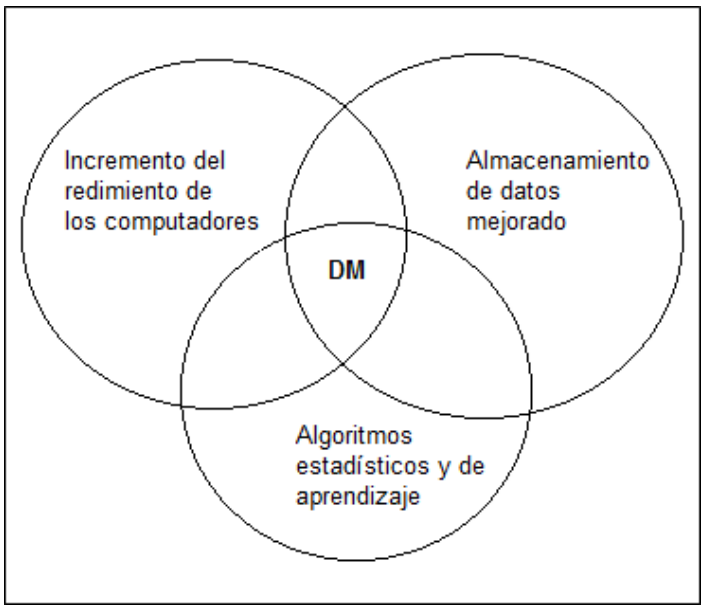
Ambos conceptos hacen alusión a la búsqueda de patrones; sin embargo, difieren en el sentido de que en la primera definición se asume el data mining como un solo proceso, mientras que en la segunda definición se la muestra como una etapa de un proceso mayor llamado descubrimiento del conocimiento.

La minería de datos se puede resumir en la búsqueda de información importante en grandes cantidades de datos que permitan la formación de patrones para tomar decisiones con respecto al negocio.

Mientras que las grandes cantidades de datos de los sistemas transaccionales y analíticos han evolucionado de manera separada, la minería de datos provee de un enlace entra ambos (UCLA ANDERSON, 1998). Lo significa que con un proceso adecuado de minería de datos es posible extraer aquella información valiosa almacenada en los repositorios de datos transaccionales y promueve la toma de decisiones inteligente a través de los sistemas analíticos del negocio.

Este proceso de extracción de datos se alimenta de tres tecnologías clave (Thearling), como se muestra en la Figura N° 1: Convergencia de tres tecnologías clave:

FIGURA N° 1: CONVERGENCIA DE TRES TECNOLOGÍAS CLAVE



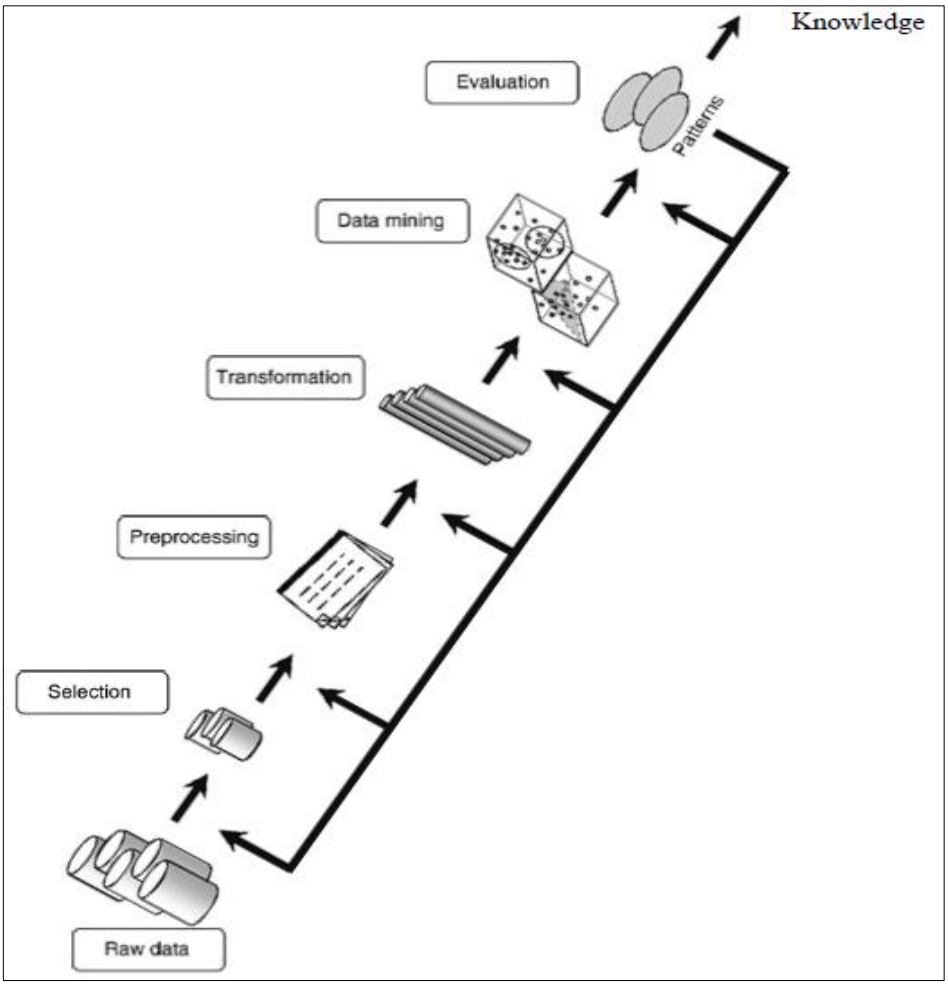
Autora: Gallegos, K. 2014

Cada uno de las tecnologías relacionadas con la inteligencia de negocios en general, son factores clave para un proceso exitoso de minería de datos y gracias a su evolución constante permiten que la minería de datos tenga un crecimiento proporcional a los avances tecnológicos, de tal forma que las constantes innovaciones en capacidad de procesamiento

de los computadores, almacenamiento en disco, y software estadístico están aumentando de manera dramática la precisión de los análisis mientras se reducen costos (UCLA ANDERSON, 1998).

La minería de datos se describe como el conjunto de diferentes fases, ver la Figura N° 2: Extracción del conocimiento desde los datos (Kumar, y otros, 2011):

FIGURA N° 2: EXTRACCIÓN DEL CONOCIMIENTO DESDE LOS DATOS.



Fuente: Kumar y Pal, 2009.

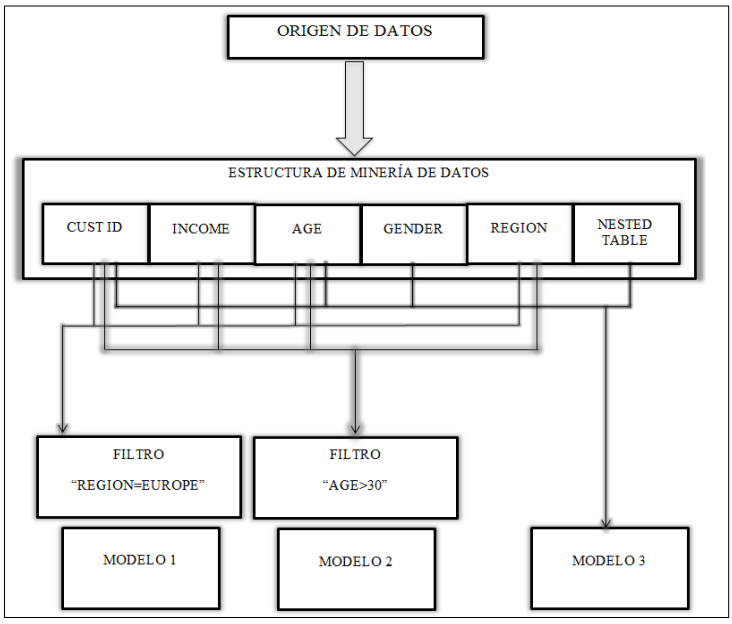
Esta perspectiva muestra que el proceso de minería de datos es, en su núcleo, un proceso de ingeniería de negocios, donde las primeras fases de selección, pre-procesamiento y transformación de datos son cruciales para la obtención de datos asegurados y consolidados para la fase próxima de minería. Finalmente la aplicación de un análisis sobre los datos da

como resultado el conocimiento o información procesada, la misma que puede ser presentada al usuario final haciendo uso de diferentes herramientas de Business Intelligence.

2.1.2. ESTRUCTURA DE MINERÍA DE DATOS

La estructura de minería de datos define los datos a partir de los cuales se generan los modelos de minería de datos: especifica la vista de datos de origen, el número y el tipo de columnas, y una partición opcional en conjuntos de entrenamiento y de pruebas. Una misma estructura de minería de datos puede admitir varios modelos que comparten el mismo dominio (Microsoft Corporation). En la Figura N° 3: Estructura genérica se presenta la condición multi-modelo de una estructura de minería de datos.

FIGURA N° 3: ESTRUCTURA GENÉRICA



Fuente: Microsoft Corporation, 2009

La configuración de una estructura de minería de datos consta de los pasos siguientes (Microsoft Corporation):

- Definir un origen de datos.- Columnas disponibles en el origen de datos existente

- Seleccionar las columnas de datos que se van a incluir en la estructura y definir una clave.
- Especificar si los datos de origen se deben separar en un conjunto de entrenamiento y en un conjunto de prueba.- Cuando se definen los datos para la estructura de minería de datos, también es posible especificar que algunos de los datos se usen para entrenamiento y otros para pruebas. Por consiguiente, ya no es necesario separar los datos antes de crear una estructura de minería de datos. En su lugar, mientras crea el modelo, puede especificar que determinado porcentaje de los datos se reserven para pruebas y que el resto se use para entrenamiento,
- Procesar la estructura.- Una estructura de minería de datos no es más que un contenedor de metadatos hasta que se procesa.

2.1.3. SELECCIÓN DE CARACTERÍSTICAS

La selección de características es un término usado habitualmente en la minería de datos para describir las herramientas y las técnicas disponibles para reducir las entradas a un tamaño apropiado para su procesamiento y análisis. La selección de características no solo implica la reducción de cardinalidades, es decir, la imposición de un límite arbitrario o predefinido en el número de atributos que se pueden considerar al crear un modelo, sino también la elección de atributos, lo que significa que el analista o la herramienta de modelado debe seleccionar o descartar activamente los atributos en función de su utilidad para el análisis. (Microsoft Corporation)

Los atributos para el proceso de minería deben ser escogidos cuidadosamente, y en base a las preguntas que serán respondidas al final del proceso de minería de datos. Otro aspecto a considerar son aquellos datos que deben ser descartados para el proceso de análisis o a su vez deben ser convertidos en un tipo representado por códigos, con el objetivo de usarlos

como parte de la dataset de la minería. Un ejemplo de este caso es la dirección domiciliaria de un estudiante, la cual, a menos que sea convertida en un dato geo-referencial o reemplazado por un código de lugares, no será de provecho para el análisis.

Dentro de la minería de datos se distinguen dos tipos de atributos, los discretos como el género, por ejemplo; y los datos continuos que son por definición numéricos y en base a los cuales se pueden realizar operaciones aritméticas.

Otra taxonomía de los atributos se da en base a su participación en el proceso de minería. Es así que un atributo puede ser de entrada, salida o ambos, y aunque superficialmente su definición resulta sencilla, en la práctica podría convertirse en un problema, ya que en términos generales un algoritmo de minería de datos usa un los datos de entrada para entender los datos de salida.

En esta etapa previa a la aplicación del análisis de minería de datos, resalta la necesidad de que el analista debe hacer énfasis en seleccionar de manera adecuada aquellos atributos que contribuirán al fin planteado, pues no todos los datos que provienen del repositorio de consolidados son útiles para un análisis u otro. Además una correcta selección también tiene que ver con los recursos del computador, pues a más atributos más complejo será el procesamiento del modelo en cualquier herramienta a utilizar. Y si bien es cierto que se cuenta con computadores eficientes, una mala selección también entorpece el proceso y puede trascender en resultados ineficientes o de baja calidad. Tres son los parámetros aconsejados a tomar en cuenta a la hora de seleccionar atributos:

- El algoritmo
- Los tipos de datos
- Parámetros propios del algoritmo

Microsoft a través de su herramienta Analysis Service propone cuatro métodos de selección de características:

- La puntuación Interestingness

- La entropía de Shannon
- Bayesiano con prioridad K2
- Equivalente Dirichlet bayesiano con prioridad uniforme

Puntuación de grado de interés

Una característica es interesante si ofrece información útil. Dado que la definición de lo que es útil varía dependiendo del escenario, el sector de la minería de datos ha desarrollado diversas maneras de medir la cualidad interestingness (Microsoft Corporation). La medida de la cualidad interestingness puede estar basada en la entropía.

Entropía de Shannon

La entropía de Shannon mide la incertidumbre de una variable aleatoria para un determinado resultado (Microsoft Corporation). Por ejemplo, la entropía de lanzar una moneda al aire para decidir algo a cara o cruz se puede representar como una función de la probabilidad de que salga cara.

Bayesiano con prioridad K2

Analysis Services proporciona dos puntuaciones de selección de características basadas en las redes bayesianas. Una red bayesiana es un gráfico dirigido o acíclico de estados y de transiciones entre ellos; esto significa que algunos estados siempre son anteriores al estado actual y otros son posteriores, y que el gráfico no se repite ni realiza bucles (Microsoft Corporation).

Equivalente Dirichlet bayesiano con prioridad uniforme

El método Equivalente Dirichlet bayesiano con prioridad uniforme (BDEU) considera un caso especial de la distribución Dirichlet, en el que se utiliza una constante matemática para crear una distribución fija o uniforme de estados anteriores. La puntuación BDE también

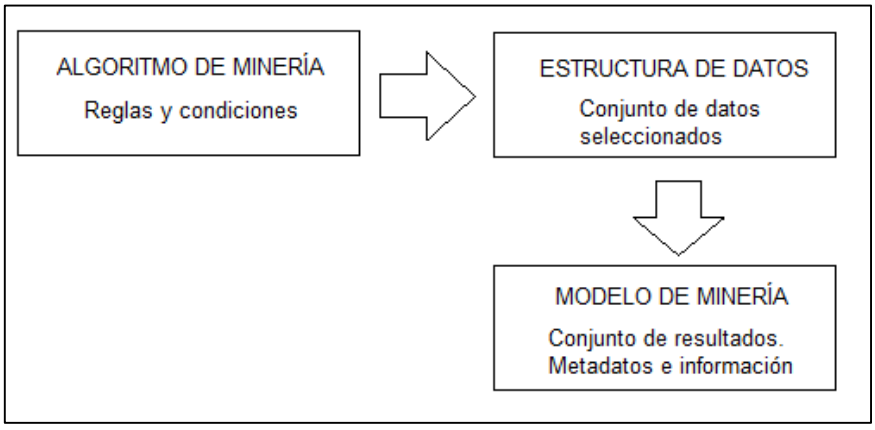
considera la equivalencia de probabilidad; esto significa que no es de esperar que los datos diferencien estructuras equivalentes (Microsoft Corporation).

2.1.4. MODELOS DE MINERÍA DE DATOS

Un modelo de minería de datos es obtenido a través de la aplicación de un algoritmo a una estructura de datos. El modelo abarca un conjunto de datos, estadísticas y patrones que pueden ser aplicados a nuevos conjuntos de datos para realizar procesos de predicción, deducción y relaciones. (Microsoft Corporation)

Para crear un modelo de minería de datos se requiere de una estructura de datos y de un algoritmo de minería, se diferencia una modelo de una estructura debido a que la estructura solo contempla al conjunto de datos, mientras que el modelo hace referencia a los datos y la información derivada del procesamiento y análisis, como patrones. Un modelo no se completa si no es hasta que se proporciona los dos elementos necesarios, la estructura y el algoritmo de minería, su resultado contiene los metadatos y enlaces de los datos que permitirán alcanzar los objetivos de un proyecto de minería. (Microsoft Corporation). En la FIGURA N° 4: ESQUEMA DE MODELO DE MINERÍA se presenta un esquema de modelo.

FIGURA N° 4: ESQUEMA DE MODELO DE MINERÍA



Autora: Gallegos, K. 2014.

Los metadatos contiene información acerca del servidor que aloja el modelo, así como las columnas usadas de la estructura para obtener el modelo, los filtros y el algoritmo aplicado debido a que todos estos factores influyen en gran medida en los resultados finales, es decir que de una misma estructura de datos se pueden obtener diversos modelos dependiendo del algoritmo usado o las columnas seleccionadas. A más de las propiedades como nombre, permisos o filtros, entre otros que constituyen la metadata del modelo, se encuentran dos propiedades fundamentales Y son las siguientes:

Propiedad del algoritmo: dentro de esta propiedad se define que algoritmo se utiliza para crear el modelo y de la cual depende los valores admitidos de la estructura de datos. Es una propiedad única para cada modelo generado y puede variar dependiendo de los algoritmos implementados por la herramienta a utilizar.

La propiedad de uso: define como se usa el modelo cada columna, ya sea de entrada, de predicción o como clave de la estructura. Es una propiedad inherente a cada columna por lo que debe establecerse individualmente para cada una de ellas. Dado que una columna de la estructura no se incluya para el modelo, se le otorga el atributo Ignorado.

Un aspecto más a tomar en cuenta son los datos de entrenamiento que se proporcionan previo a la obtención de un modelo; tiene gran influencia sobre el modelo ya que una misma estructura sometida a diferentes conjuntos de entrenamiento arroja diferentes resultados. Finalmente El último aspecto a describir del modelo de minería, son las relaciones entre los datos, que representan las referencia a los datos almacenados en caché y que pertenecen a la estructura.

Definición del modelo

Para realizar el proceso de modelado de minería de datos se necesita:

- Crear la estructura que contiene los datos adyacentes al análisis.
- Seleccionar el algoritmo que se ajuste a la tarea a realizar.

- Elegir las columnas de la estructura y especificar si se trata de un una columna de entrada o es la columna a predecir.
- Modificar los parámetros del algoritmo ya que cada uno cuenta con atributos que son modificables según las necesidades del análisis

Una vez procesado, el modelo de minería de datos contiene una gran cantidad de información sobre los datos y los patrones encontrados mediante el análisis, incluyendo estadísticas, reglas y fórmulas de regresión.

2.1.5. TAREAS EN EL DATA MINING

Una tarea dentro de la minería representa el objetivo macro a alcanzar con el proceso de análisis. Autores como Usama Fayad, dividen estas tareas en dos grandes grupos: las actividades de predicción y las actividades de descripción. La predicción recurre a usar varias variables o campos en la base de datos para predecir los valores futuros desconocidos. Por su parte, la descripción se encarga de la búsqueda de patrones que describen la información y son humanamente interpretables (Fayyad, y otros, 1996).

A partir de esta clasificación macro se pueden dividir un sin número de posibles tareas o fines bajo los cuales se desarrolla el análisis de minería de datos. Una clasificación más detallada de las tareas es presentada por Microsoft (Microsoft Corporation):

- Predecir un atributo discreto
    - Ejemplo: Marcar los clientes de una lista de posibles compradores como clientes con buenas o malas perspectivas.
- Predecir un atributo continuo
    - Ejemplo: Predecir los visitantes del sitio a partir de tendencias históricas y estacionales proporcionadas.

- Predecir una secuencia
    - Ejemplo: Analizar los factores que dan como resultado errores en el servidor.
- Buscar grupos de elementos comunes en las transacciones
    - Ejemplo: Sugerir a un cliente la compra de productos adicionales.
- Buscar grupos de elementos similares
    - Ejemplo: Crear grupos de pacientes con perfiles de riesgo en función de atributos como datos demográficos y comportamientos.

2.1.6. TIPOS DE ALGORITMOS

Se define como algoritmo de minería de datos al conjunto de cálculos y reglas heurísticas que permite generar un modelo de minería de datos que cumpla con el objetivo de la tarea de análisis seleccionada. El algoritmo también se puede definir dentro de la minería de datos como la técnica utilizada para la búsqueda de tendencias y patrones dentro de los datos proporcionados.

Las tareas de extracción de datos que se pueden llevar a cabo en un repositorio ayudan a definir una clasificación para los diferentes tipos de algoritmos de minería de datos. La siguiente categorización es una propuesta de Microsoft (Microsoft Corporation):

- Algoritmos de clasificación, que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos.
- Algoritmos de regresión, que predicen una o más variables continuas, como las pérdidas o los beneficios, basándose en otros atributos del conjunto de datos.
- Algoritmos de segmentación, que dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares.

- Algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación, que pueden usarse en un análisis de la cesta de compra.
- Algoritmos de análisis de secuencias, que resumen secuencias o episodios frecuentes en los datos, como un flujo de rutas web.

En la Tabla 2: Algoritmos de Minería, se destacan algunos de los algoritmos más utilizados por cada uno de los fines o tareas de minería expuestos anteriormente.

TABLA 2: ALGORITMOS DE MINERÍA DE DATOS

Tarea	**Algoritmo**
Predecir un atributo discreto	Algoritmo de árboles de decisión Algoritmo Bayes naive Algoritmo de clústeres Algoritmo de red neuronal
Predecir un atributo continuo	Algoritmo de árboles de decisión Algoritmo de serie temporal Algoritmo de regresión logística
Predecir una secuencia	Algoritmo de clústeres
Buscar grupos de elementos comunes en las transacciones	Algoritmo de asociación Algoritmo de árboles de decisión
Buscar grupos de elementos similares	Algoritmo de clústeres Algoritmo de clústeres de secuencia

Fuente: msdn.microsoft.com

Los modelos generados a partir de la aplicación de una técnica o algoritmo de minería de datos tienen las siguientes clases de propiedades (Microsoft Corporation):

- Las propiedades que se heredan de la estructura de minería de datos que define el tipo de datos y el tipo de contenido de los datos que usa el modelo.
- Las propiedades relacionadas con el algoritmo que se usa para crear el modelo de minería de datos, incluidos los parámetros de cliente.
- Las propiedades que definen un filtro en el modelo utilizado para entrenar el modelo.

2.2. ALGORITMOS DE MINERÍA DE DATOS DE MICROSOFT

Cada herramienta dedicada a la minería de datos ha implementado sus propios algoritmos para la extracción de conocimiento, dado que en este estudio se hará uso de las herramientas proporcionadas por Microsoft, se describen a continuación brevemente los algoritmos de minería implementados dentro los servicios de análisis de Microsoft.

Algoritmo de clústeres

El algoritmo de clústeres de Microsoft es un algoritmo de segmentación que usa técnicas iterativas para agrupar los casos de la data dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones. Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que lógicamente no se podrían derivar a través de la observación casual. Este algoritmo está orientado a la agrupación más que a las tareas de predicción. Los requisitos para un modelo de agrupación en clústeres son los siguientes:

- Una columna clave.
- Columnas de entrada (discreta o continua).
- Una columna de predicción (opcional, discreta o continua)

Algunas observaciones importantes del algoritmo de listan a continuación:

- Admite la obtención de detalles.
- Admite el uso de modelos de minería de datos OLAP y la creación de dimensiones de minería de datos

Algoritmo de árbol de decisión

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión proporcionado por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Los requerimientos del algoritmo de árbol de decisión son:

- Una columna clave.
- Columnas de entrada (discreta o continua).
- Una columna de predicción (discreta o continua)

Como datos adicionales del algoritmo se añade que:

- Admite la obtención de detalles.
- Admite el uso de modelos de minería de datos OLAP y la creación de dimensiones de |minería de datos.

Algoritmo de regresión lineal

El algoritmo de regresión lineal de Microsoft es una variante del algoritmo de árboles de decisión de Microsoft que se orienta a calcular una relación lineal entre una variable independiente y otra dependiente y, después, utilizar esa relación para la predicción. Este algoritmo requiere:

- Una columna clave.
- Columnas de entrada (continua).
- Una columna de predicción (continua)

Algoritmo de regresión logística

El algoritmo de Regresión logística de Microsoft se ha implementado utilizando una variación del algoritmo de Red neuronal. Este algoritmo comparte cualidades de las redes neurales pero es más fácil de entrenar. Es un algoritmo muy flexible, debido a que se acopla a diferentes tipos de datos y a diferentes tareas como realizar predicciones, explorar factores que contribuyen a un resultado o clasificación. Este algoritmo requiere:

- Una columna clave.
- Columnas de entrada (discreta o continua).
- Una columna de predicción (discreta o continua)

Algoritmo Bayes Naive

El algoritmo Bayes naive de Microsoft es un algoritmo de clasificación basado en los teoremas de Bayes, la palabra naïve (ingenuo en inglés) del término Bayes naive proviene del hecho que el algoritmo utiliza técnicas Bayesianas pero no tiene en cuenta las dependencias que puedan existir.

Al ser un algoritmo con menor complejidad que otros, permite la obtención más rápida de modelos de minería de datos. Este algoritmo requiere:

- Una columna clave.
- Columnas de entrada (discreta).
- Una columna de predicción (discreta)

Algoritmo de Redes Neuronales

El algoritmo de red neuronal de Microsoft combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción, y usa los datos de entrenamiento para calcular las probabilidades que son usadas para la clasificación o la regresión. Este algoritmo admite tanto datos continuos como discretos en sus columnas de entrada como de predicción, y está orientado a tareas más complejas como la minería de texto.

Dado que en este estudio se involucran los algoritmos de: regresión logística, árbol de decisión y serie temporal se describen con más detalle a continuación.

2.2.1. EL ALGORITMO DE REGRESIÓN LOGÍSTICA

El algoritmo de regresión logística es un tipo de análisis estadístico orientado a la predicción de una variable en función de otras variables consideradas como parámetros predictores. La regresión logística (RL) forma parte del conjunto de métodos estadísticos que caen bajo tal denominación y es la variante que corresponde al caso en que se valora la contribución de diferentes factores en la ocurrencia de un evento simple. (Fernández, 2011)

Específicamente el algoritmo implementado por Microsoft resulta ser una variante del algoritmo de red neuronal. Este tipo de algoritmo debido a que acepta cualquier tipo de entrada, es considerado como flexible y se ajusta a varias tareas analíticas dentro de la minería de datos, entre las que se pueden mencionar predicción, clasificación y explorar y ponderar los factores que contribuyen a un resultado específico.

2.2.1.1. Funcionamiento

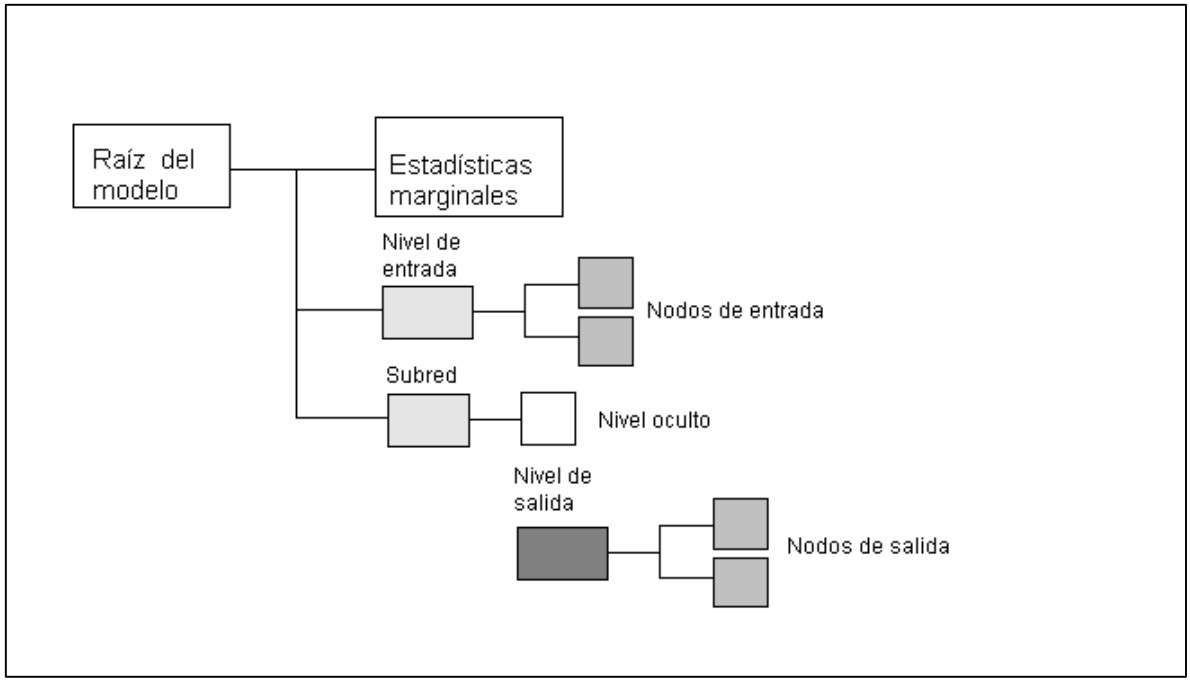
La implementación de este algoritmo usa una modificación de la red neuronal para el modelamiento de las relaciones entre las entradas y las salidas (Microsoft Corporation, 2012).

Para cada estado que aparece en el conjunto de entrenamiento, el modelo genera una entrada. Para las entradas discretas o de datos discretos, se crea una entrada adicional para

representar el estado Missing, si aparece al menos una vez en el conjunto de entrenamiento. En las entradas continuas, se crean al menos dos nodos de entrada: uno para los valores Missing, si están presentes en los datos de entrenamiento, y una entrada para todos los valores existentes o no nulos. Todos los algoritmos de minería de datos de Analysis Services utilizan automáticamente la selección de características para mejorar el análisis y reducir la carga de procesamiento. El método utilizado para la selección de características en un modelo de regresión logística depende del tipo de datos del atributo. Dado que la regresión logística se basa en el algoritmo de red neuronal de Microsoft, utiliza un subconjunto de los métodos de selección de características que se aplican a las redes neuronales. (Microsoft Corporation, 2012)

En la Figura N°5: Estructura Regresión Logística, se presentan los nodos en un modelo de regresión. Esto proporciona información adicional sobre el tipo de nodo, lo que facilita la comprensión de las relaciones entre los tipos de información del modelo.

FIGURA N° 5: ESTRUCTURA REGRESIÓN LOGÍSTICA



Fuente: msdn.microsoft.com

La denominación de los nodos en un modelo Regresión Logística proporciona información adicional sobre el tipo de nodo, lo que facilita la comprensión de las relaciones entre los tipos de información del modelo (Microsoft Corporation, 2012). Los principales tipos de nodos son conocidos en la herramienta de Analysis Services como:

- Raíz del modelo
- Nodo de estadísticas marginales
- Atributo de predicción
- Atributo de entrada
- Estado de atributo de entrada

2.2.1.2. Elementos requeridos

Para la aplicación correcta de este algoritmo en minería de datos se asume que se cuenta con un escenario con las siguientes características (Martínez, y otros, 2004):

- Atributo de entrada que admite los siguiente tipos: continuo, discreto, de datos discretos, clave, tabla,
- Atributo de predicción: continuo, discreto, de datos discretos.

Desde el punto de vista aplicado al uso de la herramienta Analysis Services, el algoritmo Regresión Logística de Microsoft admite varios parámetros que influyen en el rendimiento y la precisión del modelo de minería de datos resultante (Microsoft Corporation, 2012). Estos parámetros son personalizables y se describen en la Tabla 3: Parámetros del Algoritmo Regresión Logística:

TABLA 3: PARÁMETROS DEL ALGORITMO REGRESIÓN LOGÍSTICA

Atributo	Tipo de contenido
HOLDOUT_PERCENTAGE	Porcentaje para determinar los casos de entrenamiento. Valor predeterminado: 30
HOLDOUT_SEED	Número para inicializar el generador pseudo-aleatorio Valor predeterminado: 0
MAXIMUN_INPUT_ATTRIBUTES	Número de atributos de entrada que el algoritmo administrará antes de la selección de características. Valor predeterminado: 255
MAXIMUN_OUTPUT_ATTRIBUTES	Número de atributos de salida que el algoritmo administrará antes de la selección de características. Valor predeterminado: 255
MAXIMUN_STATES	Número máximo de estados de un atributo que admite el algoritmo. Valor predeterminado: 100
SAMPLE_SIZE	Número de casos que se van a utilizar para entrenar el modelo. El algoritmo utilizará el valor mínimo entre este parámetro y HOLDOUT_PERCENTAGE. Valor predeterminado: 10000

Fuente: msdn.microsoft.com

Al preparar los datos para su uso en el entrenamiento de un modelo de regresión logística, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan.

Los requisitos para un modelo de regresión logística son los siguientes:

- Una columna de una sola clave: cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- Columnas de entrada: cada modelo debe tener al menos una columna de entrada que contenga los valores que se utilizan como factores en el análisis. Puede tener tantas columnas de entrada como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el modelo.
- Al menos una columna de predicción: el modelo debe contener al menos una columna de predicción de cualquier tipo de datos, incluidos datos numéricos continuos. Los valores de la columna de predicción también se pueden tratar como entradas del modelo, o se puede especificar que solo se utilicen para las predicciones. No se admiten tablas anidadas en las columnas de predicción, pero se pueden usar como entradas.

El algoritmo de minería de regresión logística tiene ciertas condiciones respecto a otros algoritmos de Microsoft y se listan a continuación:

- No admite la obtención de detalles. Esto se debe a que la estructura de nodos del modelo de minería de datos no tiene por qué corresponder directamente a los datos subyacentes.
- No admite la creación de dimensiones de minería de datos.
- Admite el uso de modelos de minería de datos OLAP.

- No admite el uso del lenguaje de marcado de modelos de predicción (PMML) para crear modelos de minería de datos. (Microsoft Corporation, 2012)

2.2.2. EL ALGORITMO DE ÁRBOL DE DECISIÓN

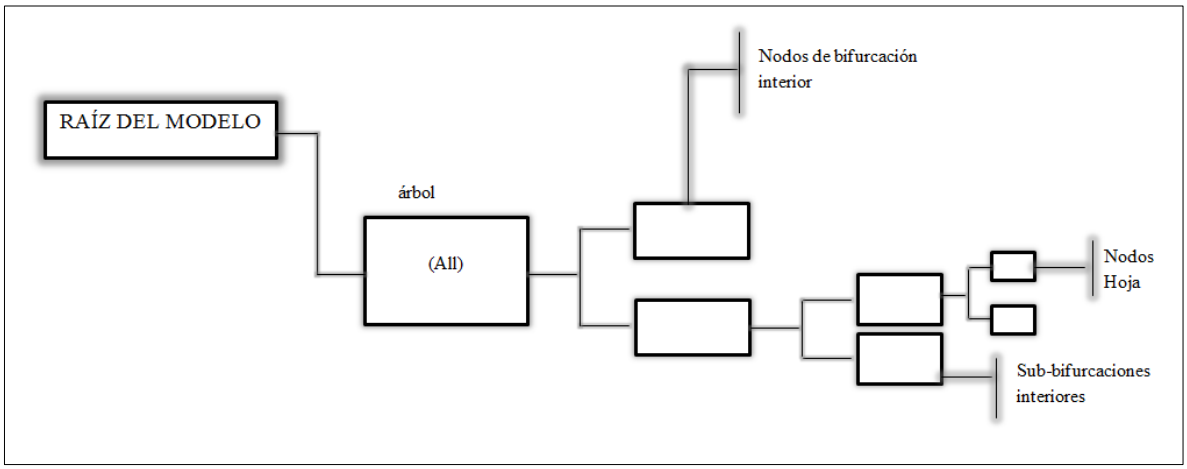
Los árboles de decisión y reglas que usan divisiones invariantes tienen una forma de representación simple, haciendo del modelo de inferencia relativamente sencillo para el entendimiento del usuario. Sin embargo, una restricción a un árbol en particular o regla de representación puede restringir de manera significativa la funcionalidad del mismo (Fayyad, y otros, 1996).

2.2.2.1. Funcionamiento

Un modelo de árboles de decisión tiene un nodo primario único que representa el modelo y sus metadatos. Debajo del nodo primario aparecen árboles independientes que representan los atributos de predicción que se seleccionan. Por ejemplo, si configura su modelo de árboles de decisión para predecir si los clientes comprarán algo y, a continuación, proporciona entradas para el género y los ingresos, el modelo creará un árbol único para el atributo de compra, con muchas bifurcaciones que se dividen en función de condiciones relacionadas con el género y los ingresos (Microsoft Corporation, 2012).

El árbol para cada atributo de predicción contiene información que describe cómo afectan las columnas de entrada elegidas al resultado de ese atributo de predicción concreto. Cada árbol está encabezado por un nodo que contiene el atributo de predicción, seguido de una serie de nodos que representan los atributos de entrada. Un atributo corresponde a una columna de nivel de caso o a valores de columnas de tabla anidada, que generalmente son los valores que aparecen en la columna Key de la tabla anidada. Los nodos interiores y los nodos hoja representan las condiciones de división. Un árbol se puede dividir varias veces por el mismo atributo. En la Figura N° 6: Estructura Árbol de Decisión se presenta gráficamente la posición de los nodos dentro del árbol.

FIGURA N° 6: ESTRUCTURA ÁRBOL DE DECISIÓN



Fuente: msdn.microsoft.com

En los modelos de árboles de decisión, se crean los tipos de nodos siguientes:

- Modelo.- Nodo raíz para el modelo.
- Árbol.- Nodo primario para los árboles de clasificación del modelo.
- Interior.- Encabezado de la bifurcación interior, que se encuentra dentro de un árbol de clasificación o de regresión.
- Distribución.- Nodo hoja, que se encuentra dentro de un árbol de clasificación o de regresión.
- Árbol de regresión.- Nodo primario para el árbol de regresión dentro del modelo.

2.2.2.2. Elementos requeridos

Los parámetros requeridos para la aplicación de un algoritmo de árboles de decisión pueden depender de la herramienta a utilizar; desde Microsoft Analisys Service se exigen los siguientes para generar resultados (Microsoft Corporation, 2012):

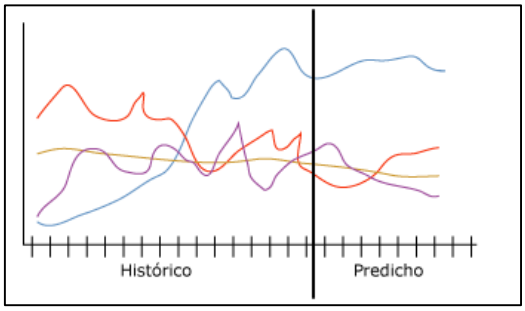
- Una columna clave: cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única.
- Una columna de predicción. Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.
- Columnas de entrada. Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

2.2.3. ALGORITMO DE SERIE TEMPORAL

El algoritmo de serie temporal de Microsoft proporciona los algoritmos de regresión que se optimizan para la previsión en el tiempo de valores continuos. Mientras que otros algoritmos de Microsoft, como por ejemplo los árboles de decisión, requieren columnas adicionales de nueva información como entrada para predecir una tendencia, los modelos de serie temporal no las necesitan. Un modelo de serie temporal puede predecir tendencias basadas únicamente en el conjunto de datos original utilizado para crear el modelo. (Microsoft Corporation, 2012)

La Figura N°7: Modelo de Serie Temporal muestra un modelo típico de previsión en el tiempo.

FIGURA N° 7: MODELO DE SERIE TEMPORAL



Fuente: msdn.microsoft.com

La información histórica aparece a la izquierda de la línea vertical y representa los datos que el algoritmo utiliza para crear el modelo. La información de la predicción aparece a la derecha de la línea vertical y representa la previsión realizada por el modelo. A la combinación de los datos de origen y los datos de la predicción se le denomina serie. (Microsoft Corporation, 2012)

2.2.3.1. Funcionamiento

El algoritmo de serie temporal de Microsoft utiliza una mezcla de los dos algoritmos al analizar patrones y realizar predicciones. El algoritmo entrena dos modelos independientes sobre los mismos datos: uno de los modelos utiliza el algoritmo ARTXP y el otro modelo utiliza el algoritmo ARIMA. A continuación, el algoritmo combina los resultados de los dos modelos para obtener la mejor predicción sobre un número variable de intervalos de tiempo. Dado que ARTXP obtiene mejores resultados en las predicciones a corto plazo, se le da mayor importancia al principio de una serie de predicciones. Sin embargo, a medida que los intervalos de tiempo que se están prediciendo se adentran en el futuro, se va dando más importancia a ARIMA. (Microsoft Corporation, 2012)

2.2.3.2. Elementos requeridos

Cada modelo de previsión debe contener una serie de casos, que es la columna que especifica los intervalos de tiempo u otras series sobre las que se produce el cambio. Los requisitos para un modelo de serie temporal son los siguientes:

- Una única columna Key Time Cada modelo debe contener una columna numérica o de fecha que se utilizará como serie de casos y que define los intervalos de tiempo que utilizará el modelo. El tipo de datos para la columna de clave temporal puede ser un tipo de datos fecha o bien numérico. Sin embargo, la columna debe contener valores continuos y éstos deben ser únicos para cada serie. La serie de casos para un modelo de serie temporal no pueden estar almacenada en dos columnas como por ejemplo una columna Año y una columna Mes.
- Una columna predecible Cada modelo debe contener por lo menos una columna predecible alrededor de la que el algoritmo generará el modelo de serie temporal. El tipo de datos de la columna predecible debe contener valores continuos. Por ejemplo, es posible predecir la manera en que los atributos numéricos tales como ingreso, ventas o temperatura, varían con el tiempo. Sin embargo, no es posible utilizar como columna predecible una columna que contenga valores discretos tales como el estado de las compras o el nivel de educación.
- Una columna de clave de serie opcional Cada modelo puede tener una columna de clave adicional que contenga valores únicos que identifiquen a una serie. La columna de clave de serie opcional debe contener valores únicos. Por ejemplo, un solo modelo puede contener ventas de muchos modelos de producto, siempre y cuando haya un solo registro para cada nombre del producto para cada intervalo de tiempo.

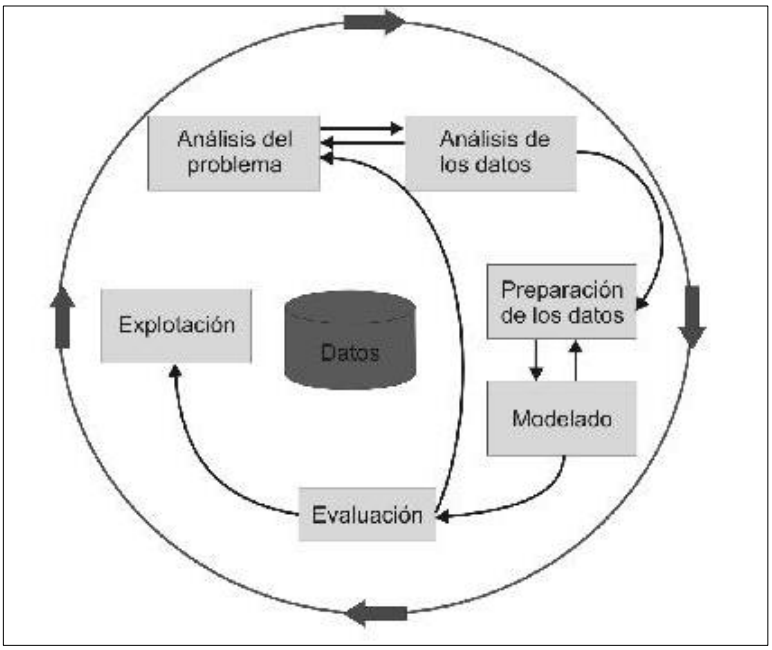
Los algoritmos de minería de datos son aplicados dependiendo del caso de estudio, en donde es necesario explorar las características que admite una u otra técnica así como es necesario saber la tarea de minería de datos a aplicar para obtener los mejores resultados.

2.3. CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

Cross Industry Standard Process for Data Mining CRISP-DM ofrece una metodología y una guía para el desarrollo de proyectos de minería de datos. A la vez que se enfoca en proporcionar un modelo de referencia que se puede adaptar a los diferentes casos y técnicas de minería de datos, ofrece un conjunto de consejos para cada una de las fases que se describen en el modelo, siendo ésta la guía del usuario que recopila las mejores prácticas de los creadores de esta metodología para los procesos de data mining. Dentro del modelo de referencia se describe el ciclo de un proyecto de minería, el mismo que está segmentado en fases, y éstas a su vez se dividen en tareas.

Las seis fases del proceso de minería de datos se muestran en la Figura N°8: Fases del Modelo de Referencia CRISP-DM. Donde la secuencia representada no es necesariamente rígida, puesto que en ocasiones se deberá alterar el orden propuesto, en consecuencia de los resultados de una fase determinada (Chapman, y otros, 2000).

FIGURA N° 8: FASES DEL MODELO DE REFERENCIA CRISP-DM



Fuente: CRISP-DM

2.3.1. FASES Y TAREAS

Las fases y tareas del ciclo de CRISP-DM son descritas en términos genéricos de tal manera que la relación con un proyecto específico pueda ser realizada enmarcándolo en un caso y recursos determinados.

Entendimiento del negocio o análisis del problema

El propósito de la fase de análisis del problema es entender los requerimientos que se pretende cubrir con el proyecto de minería desde la perspectiva del negocio. Con la fase de entendimiento del negocio se busca obtener una definición del problema y un plan preliminar diseñado para alcanzar los objetivos del proyecto. Las tareas designadas para esta fase se describen en la Tabla 4: Tareas genéricas y salidas análisis del problema.

TABLA 4: TAREAS GENÉRICAS Y SALIDAS ANÁLISIS DEL PROBLEMA

Tarea	Descripción	Salidas

Determinar los objetivos del negocio	Se pretende desde la perspectiva del negocio encontrar lo que el cliente realmente desea alcanzar con el proyecto de minería de datos	◆Información general de la institución ◆El objetivo principal del cliente en términos del negocio ◆El criterio de aceptación desde el negocio para aprobar o no los resultados del proyecto
Evaluar la situación	Determinar los recursos, restricciones, supuestos y otros factores que deberían sr considerados para el planteamiento del objetivo y el plan del proyecto de análisis	◆Lista de recursos disponibles para el proyecto ◆Lista de los requerimientos del proyecto (tiempo, tipo de resultados), supuestos y restricciones ◆Lista de los riesgos que pueden retrasar el proyecto y planes de contingencia
Determinar los objetivos de la minería de datos	El objetivo del negocio está formulado en términos propios, sin embargo para el objetivo de minería de datos deberá ser expresado en términos técnicos	◆Descripción de las salidas esperadas del proyecto que permitan alcanzar el objetivo del negocio. ◆Definición del criterio de éxito para las salidas
Generar el plan del proyecto	El plan tentativo para el desarrollo del proyecto y la consecución de los objetivos de la minería de datos.	◆Lista de las etapas a ser ejecutadas en este proceso, incluyendo duración, recursos requeridos, entradas y salidas.

Fuente: CRISP-DM 1.0, 2000

Entendimiento de los datos o análisis de los datos

La fase de análisis de datos inicia con la recolección de datos y continúa con las tareas que permiten al analista familiarizarse con los recursos de información disponible, identificar a

primera vista los problemas de calidad de datos, y lanzar hipótesis sobre los patrones que se pueden descubrir en el proceso. Esta fase contempla diferentes tareas, se presenta un resumen en la Tabla 5: Tareas genéricas y salidas del análisis de datos

TABLA 5: TAREAS GENÉRICAS Y SALIDAS DEL ANÁLISIS DE DATOS

Tarea	**Descripción**	**Salidas**
Recolectar la data inicial	Acceder a la dataset presentada como recurso dentro de la fase inicial del proyecto.	◆Lista de la dataset a la que se accede con los problemas incurridos así como la solución propuesta.
Describir los datos	Describir las propiedades de los datos adquiridos.	◆Descripción de los datos incluyendo: formato, cantidad de datos, los identificadores.
Explorar los datos	Utilizar técnicas como consultas, visualizaciones y reportes con el propósito de encontrar los atributos claves para el proceso de minería	◆Descripción de los descubrimientos realizados así como primeras hipótesis. Si se considera, agregar gráficos de las características encontradas
Verificar la calidad de los datos	Examinar los posibles errores dentro de la data, así como determinar soluciones tentativas	◆Lista de resultados de la verificación de la calidad de datos.

Fuente: CRISP-DM 1.0, 2000

Preparación de los datos

La fase de preparación de datos incluye las actividades necesarias para obtener la dataset definitiva, es decir el conjunto de datos que serán ingresados en el modelado de minería de datos. Las tareas para la preparación de los datos se resumen en la Tabla 6: Tareas genéricas y salidas preparación de los datos

TABLA 6: TAREAS GENÉRICAS Y SALIDAS PREPARACIÓN DE LOS DATOS

Tarea	**Descripción**	**Salidas**
Seleccionar los datos	Decir que datos serán usados en el análisis	◆Lista de la dataset incluida y excluida y las razones de la selección
Limpiar los datos	Elevar la calidad de los datos a la requerida para el proceso de análisis	◆Descripción de los procesos de limpieza a los que se ha incurrido así como la fundamentación de dichas acciones.
Construir los datos	Agregar columnas, filas completas, o transformar los datos en los requeridos para el análisis	◆Lista de los atributos agregados al conjunto de datos ◆Descripción de los nuevos registros generados
Integrar los datos	Usar los métodos para combinar los datos generados si es necesario	◆Lista de tablas originadas, es decir aquellas que surgen de agrupar la información en un solo objeto.
Dar formato a los datos	Describir las modificaciones sintácticas en los datos.	◆Listar aquellos cambios que modifican la forma de los datos, pero no su fondo.

Fuente: CRISP-DM 1.0, 2000.

Modelado

En la fase de modelado se seleccionan técnicas de entre las disponibles para el caso de minería de datos del proyecto, luego se aplican y se parametrizan para obtener los mejores resultados. La Tabla 7: Tareas genéricas y salidas para modelamiento, resume las actividades a realizar dentro de esta fase.

TABLA 7: TAREAS GENÉRICAS Y SALIDAS PARA MODELAMIENTO

Tarea	Descripción	Salidas
Seleccionar la técnica de modelamiento	Se puede escoger más de una técnica y se aplicará las tareas de esta fase de manera individual para cada técnica	◆Documento de la técnica a utilizar ◆Supuestos sobre los datos, que cada técnica posee
Generar el diseño de prueba	Antes de construir en modelo real es necesario medir la calidad y validez del modelo	◆Descripción del plan de pruebas.
Construir el modelo	Ejecutar la herramienta de modelado para obtener uno o más modelos	◆Configuración de los parámetros enviados al modelo. ◆Los modelos resultantes
Evaluar el modelo	Determinar la calidad de los modelos generados en términos de precisión, esta tarea incluye solamente a los modelos generados	◆Lista de tablas originadas, es decir aquellas que surgen de agrupar la información en un solo objeto.
Dar formato a los datos	Describir las modificaciones sintácticas en los datos.	◆Listar aquellos cambios que modifican la forma de los datos, pero no su fondo.

Fuente: CRISP-DM 1.0, 2000

Evaluación

En este punto del proyecto, ya se ha construido un modelo que apunta tener gran calidad, sin embargo antes de proceder al desarrollo final del modelo es necesario realizar una evaluación para asegurar que el modelo alcanzará los objetivos del negocio. La Tabla 8: Tareas genéricas y salidas para la evaluación, resume las actividades a realizar dentro de esta fase.

TABLA 8: TAREAS GENÉRICAS Y SALIDAS PARA LA EVALUACIÓN

Tarea	Descripción	Salidas
Evaluar los resultados	Esta tarea está enfocada en la evaluación del modelo en cuanto a si este alcanza o no los objetivos del negocio	◆Resumen de los resultados en cuanto al cumplimiento del criterio de aceptación del negocio
Proceso de revisión	Una vez evaluados los resultados, es necesario revisar nuevamente los procesos para detectar si algún factor importante fue omitido	◆Resultado de la revisión y resaltar los procesos que deben ser revisados nuevamente
Determinar los siguientes pasos	Se decide si continuar con la implementación o empezar un nuevo proyecto de minería de datos	◆Lista de las actividades futuras tentativas

Fuente: CRISP-DM 1.0, 2000

Explotación o Implementación

La creación del modelo no siempre representa el final de un proyecto de minería de datos, sino que concluye con una representación de los datos de tal manera que pueda ser comprendida por el usuario final. Estas representaciones pueden ocasionalmente involucran la realización de modelos en "vivo" que apoyan la toma de decisiones o simplemente reportes. La Tabla 9: Tareas genéricas y salidas para la implementación, resume las actividades a realizar dentro de esta fase.

TABLA 9: TAREAS GENÉRICAS Y SALIDAS PARA LA IMPLEMENTACIÓN

Tarea	Descripción	Salidas
Plan de implementación	Definir una estrategia para la implementación	◆Descripción de la estrategia a utilizar
Plan de	Revisión del uso de los resultados	◆Informe de los resultados

mantenimiento	de minería de datos	obtenidos en el plan de mantenimiento
Generar el informe final	Se decide si continuar con la implementación o empezar un nuevo proyecto de minería de datos	◆Resumen del proceso de minería de datos
Revisión del proyecto	Se revisa los aciertos y los problemas que ocurrieron durante el desarrollo del producto	Documento en el que se resalta la importancia de la experiencia ganada durante el proyecto

Fuente: CRISP-DM 1.0, 2000

Las fases y tareas descritas corresponden a un modelo de referencia, dependerá del caso y del propósito del proyecto de minería, el orden o el cumplimiento de las fases y sus tareas correspondientes. (Chapman, y otros, 2000).

3. CAPÍTULO III: COMPARACIÓN DE ALGORTIMOS ÁRBOL DE DECISIÓN Y REGRESIÓN LOGÍSTICA

Dentro de este apartado se realiza la comprobación de la hipótesis correspondiente al presente trabajo de investigación; a través de este proceso se cubren dos de las fases de la metodología CRISP-DM y son el modelado y la evaluación, en donde se procede a modelar una muestra de los requisitos de la minería de datos y a recolectar los valores que definirán la mejor opción entre los algoritmos para los resultados finales, en este caso haciendo uso de estadística descriptiva e inferencial.

3.1. PARÁMETROS DE COMPARACIÓN

Los algoritmos de predicción pueden ser analizados por diferentes variables. En la Tabla 10: Parámetros de comparación, se define la variable de desempeño.

TABLA 10: PARÁMETROS DE COMPARACIÓN

Variable	**Tipo**	**Concepto**
Desempeño	Compleja	Desempeño está relacionado con las características de tiempo de ejecución y respuesta, uso

		de recursos y confiabilidad de las operaciones.

Fuente: SCRUM

Debido a que se busca los mejores resultados posibles con el algoritmo seleccionado, se analizará el desempeño, ya que, por su definición no sólo se asocia a características de ejecución sino también de efectos del algoritmo sobre las entradas proporcionadas.

3.1.1. DETERMINACIÓN DE INDICADORES

Al ser el desempeño una variable de tipo compleja se deben definir los indicadores de la misma, así como los pesos que proporcionarán prioridad a cada uno de los criterios.

Tiempo de respuesta

El tiempo de respuesta corresponde a un indicador de la categoría de velocidad y estará medido en segundos. A menor tiempo de respuesta de un algoritmo frente a una misma estructura de datos mejor puntuación tendrá. Del 100% se asigna a este indicador un 10% del peso total.

Uso del CPU

El uso del CPU corresponde a la categoría de uso de recursos del computador; estará medido en porcentaje. Un algoritmo tendrá mejor desempeño con respecto al uso de CPU cuanto menor sea su valor. Se asigna a este indicador un 10% del peso total de criterios.

Uso de Memoria

El uso de memoria medido en Megabytes, al igual que el uso del CPU, será mejor en cuanto un algoritmo utilice menos memoria frente a la misma estructura de datos. Se asigna a este criterio un 10% del total.

Precisión

La precisión es el indicador más importante a la hora de decidir que algoritmo tiene un mejor desempeño por lo que se le asigna un 70% del total de la decisión final. Este valor será obtenido de la herramienta SSAS que ofrece la opción "Gráfico de precisión" y es la que ayuda a comparar los modelos calculando la efectividad del modelo a través de una población normalizada. Una mayor puntuación es mejor (Microsoft Corporation). La precisión es una magnitud adimensional.

3.2. MÉTODOS

Se ha aplicado el método científico para realizar el trabajo de investigación, el mismo que consta del planteamiento del problema, formulación de hipótesis, levantamiento de información, análisis e interpretación de resultados, comprobación de la hipótesis y difusión de resultados.

3.3. TÉCNICAS

Se ha usado la observación como recurso de recolección de los datos resultantes de las herramientas de medición.

Para el análisis de resultados se empleará técnicas de estadística descriptiva e inferencial.

3.4. INSTRUMENTOS

Los instrumentos para la medición de indicadores son:

Monitor del Sistema de Windows 8 que permite obtener estadísticas sobre la actividad y el rendimiento actuales de los procesos que se ejecutan.

Data Tools de Microsoft de Microsoft Analysis Services (SSDT) una herramienta asociada al desarrollo de bases de datos y proyectos de inteligencia de negocios.

Los indicadores del desempeño, su técnica y fuente de verificación se resumen en la Tabla 11: Indicadores de la variable:

TABLA 11: INDICADORES DE LA VARIABLE

Variable	Categoría	Indicadores	Técnica	Fuentes de verificación
Desempeño	Velocidad	Tiempo de respuesta	Observación	Microsoft SQL Data Tools
	Uso de recursos del computador	Uso del CPU	Observación	Monitor del sistema
		Uso de memoria	Observación	Monitor del sistema
	Precisión	Precisión de modelo obtenido.	Observación	Microsoft SQL Data Tools

Autora: Gallegos, K., 2014

3.5. AMBIENTES DE PRUEBA

Para la recolección de datos necesarios para la comprobación de la hipótesis, se procedió a implementar los algoritmos involucrados en el estudio en una muestra del total de requerimientos del proyecto.

El servidor sobre el que se practicaron las pruebas tiene las características de hardware y software que se describen en la Tabla 12: Especificaciones del servidor.

TABLA 12: ESPECIFICACIONES DEL SERVIDOR

Característica	Especificación
Procesador	Intel® Core™ i7-4702MQ @ 2.20GHz
Memoria RAM	8GB
Disco Duro	1TB

Sistema operativo	Windows 8.1
Motor de base de datos	Microsoft SQL Server 2012 Standard

Autora: Gallegos, K., 2014

Este ambiente de pruebas está orientado a cubrir los objetivos del proyecto de minería, los que se relacionan a los ejes del proceso académico que son ingreso, matriculación, promoción y graduación, Los requerimientos se listan a continuación:

- REQ 1: Determinar patrones de comportamiento para el ingreso de los estudiantes; por carrera y facultad.
- REQ 2: Determinar patrones de comportamiento para el ingreso directo a la carrera (sin pasar por el curso de nivelación); por carrera y facultad.
- REQ 3: Determinar patrones de comportamiento para la matriculación y selección de asignatura de los estudiantes (número de asignaturas, número de créditos, nivel y área de las asignaturas, etc.); por carrera, por nivel y por facultad.
- REQ 4: Determinar los factores que tienen influencia en los casos de deserción (retiros y pérdida de la asignatura por asistencia), por carrera, niveles, asignaturas, áreas de conocimiento y facultad.
- REQ 5: Determinar patrones de comportamiento en la promoción académica de los estudiantes; por asignatura, nivel, áreas de conocimiento, carrera y facultad.
- REQ 6: Determinar los factores que influyen en los escenarios de segunda y tercera matrícula; por asignatura, nivel, áreas de conocimiento, carrera y facultad.
- REQ 7: Determinar la proyección de graduados; por carrera y facultad.
- REQ 8: Determinar los factores que inciden en los casos de estudiantes con baja eficiencia terminal; por carrera y facultad.

3.5.1. POBLACIÓN Y TAMAÑO DE LA MUESTRA

Para realizar la medición de indicadores se tomará como población el total de los requisitos generados para el proyecto de minería. Debido a que cada uno de ellos exige diferentes niveles de detalle, esto da origen a que exista un modelo por cada segmentación requerida, es decir se deberán desarrollar un total de 528 modelos para satisfacer los requisitos del proyecto, siendo éste el número de la población, razón por la cual la medición de indicadores se realizó sobre una muestra del total. El tamaño de la muestra se define mediante:

ECUACIÓN 1: TAMAÑO DE LA MUESTRA

$$n = \frac{N * p * q * Z^2}{e^2 * (N-1) + p * q * Z^2}$$

Los valores de la fórmula anterior se muestran en la Tabla 13: Valores para determinar la muestra.

TABLA 13: VALORES PARA DETERMINAR LA MUESTRA

Variable	Definición	Valor
N	Tamaño de la población.	528
P	Variabilidad positiva	0,5
Q	Variabilidad negativa	0,5
Z	Valor obtenido mediante niveles de confianza. Es un valor constante.	1,7 => 91%
E	Límite aceptable de error que, generalmente cuando no se tiene su valor, suele utilizarse un valor que varía entre el 1% (0,01) y 9% (0,09).	9%

Autora: Gallegos, K., 2014

Con los valores definidos en la Tabla 13, el tamaño de la muestra resultante n es 76. Este es el número de modelos que serán sometidos a prueba para realizar el análisis comparativo de

los algoritmos en base al desempeño de los mismos. De la estadística descriptiva los valores que interesan para la presente investigación son:

- Media, es el promedio de todos los valores de la muestra evaluada.
- Desviación estándar, es la medida de dispersión de los valores obtenidos con respecto a la media.

Cada uno de los algoritmos será aplicado a una estructura de datos que satisface los requisitos definidos. Estas estructuras son un conjunto de datos obtenidas con lenguaje SQL e integradas a la herramienta Data Tools a través de la opción de "Vista de Datos". A continuación se presenta un ejemplo de las estructuras de datos usadas:

```
SELECT        ROW_NUMBER() OVER (ORDER BY mat.dtFechaCreada) AS numero,
[sintCodigo], mat.[strCodPeriodo], [strCodEstud], est.strNombres + ' ' +
est.strApellidos AS nombres, est.strCodSexo, est.strCodTit,

est.strNacionalidad, est.strCodInt, DATEDIFF(yy, est.dtFechaNac,
dtFechaIng) AS edadInscripcion, mat.[strCodNivel],
mat.[dtFechaAutorizada], mat.[dtFechaCreada], mat.[strCodEstado],
materiaPensum.strCodArea,

areas.strNombre AS area, materiaPensum.bytHorasPrac,
materiaPensum.bytHorasTeo, materiaPensum.strCodNivel as nivelMateria,
materiaPensum.fltCreditos, count(areas.strCodigo) as numeroMaterias

FROM            [OAS_IngSistemas].[dbo].[Matriculas] mat INNER JOIN

[dbo].[Estudiantes] est ON mat.strCodEstud = est.strCodigo INNER JOIN

[dbo].[Materias_Asignadas] materiasAsignadas ON
materiasAsignadas.sintCodMatricula = mat.sintCodigo AND
materiasAsignadas.strCodPeriodo = mat.strCodPeriodo INNER JOIN

[dbo].[Materias] materia ON materiasAsignadas.strCodMateria =
materia.strCodigo INNER JOIN

[dbo].[Periodos] per ON mat.strCodPeriodo = per.strCodigo INNER JOIN

[dbo].[Pensums] pensum ON per.strCodPensum = pensum.strCodigo INNER JOIN
```

```
[dbo].[Materias_Pensum] materiaPensum ON materiaPensum.strCodPensum =
pensum.strCodigo AND materiaPensum.strCodMateria = materia.strCodigo
INNER JOIN

[dbo].[Areas] areas ON materiaPensum.strCodArea = areas.strCodigo

WHERE mat.[strCodNivel] = '1'

GROUP BY  [sintCodigo], mat.[strCodPeriodo], [strCodEstud],
est.strNombres + ' ' + est.strApellidos , est.strCodSexo, est.strCodTit,

est.strNacionalidad, est.strCodInt,  est.dtFechaNac, dtFechaIng,
mat.[strCodNivel], mat.[dtFechaAutorizada], mat.[dtFechaCreada],
mat.[strCodEstado], materiaPensum.strCodArea,

areas.strNombre , materiaPensum.bytHorasPrac, materiaPensum.bytHorasTeo,
materiaPensum.strCodNivel , materiaPensum.fltCreditos
```

Los escenarios surgen al momento de agregar un algoritmo a la estructura de datos; es decir, se definen con la implementación del algoritmo de árbol de decisión y regresión logística respectivamente.

3.6. ESCENARIO 1 PARA REGRESIÓN LOGÍSTICA

Con la estructura de datos ingresada en la herramienta SSDT se selecciona el algoritmo de minería correspondiente, en este caso el algoritmo de regresión logística, los pasos a seguir para realizar este proceso son:

a) Seleccionar la definición el origen de la estructura de datos (relacional o multidimensional)

b) Seleccionar el algoritmo de Regresión Logística

c) Seleccionar el conjunto de datos de origen

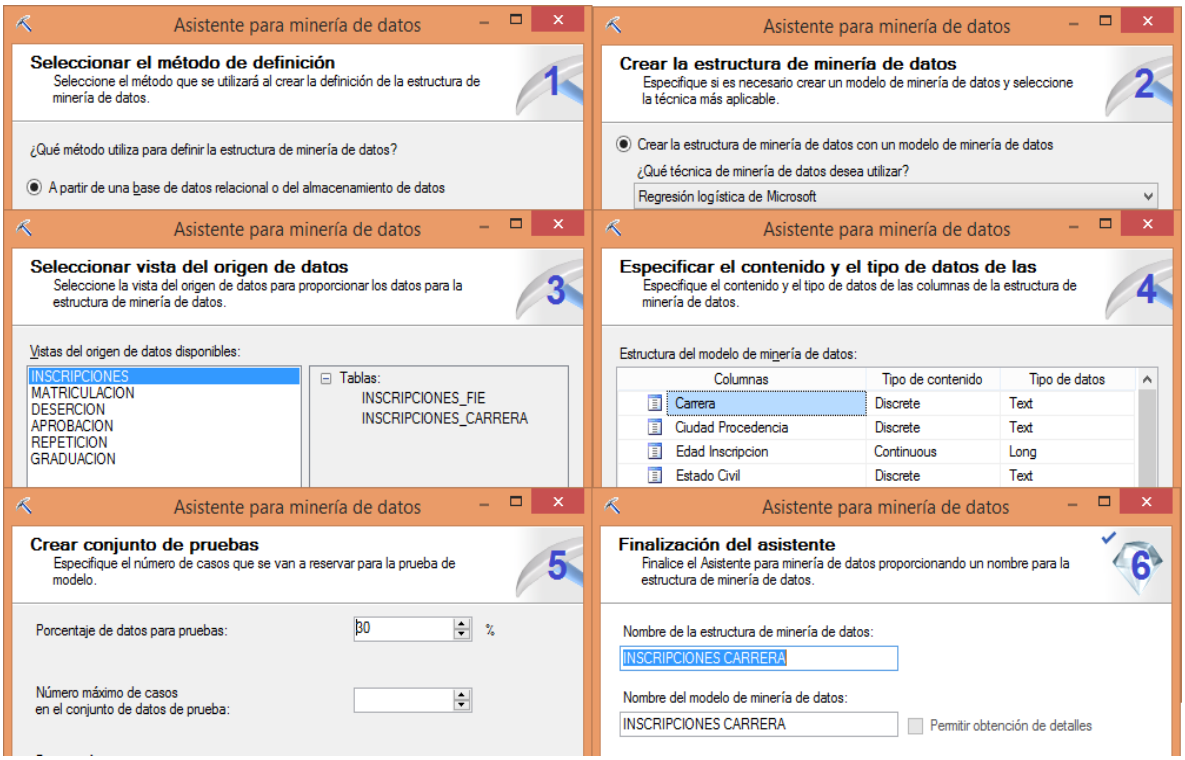
d) Especificar los tipos de datos

e) Definir el conjunto de aprendizaje y de pruebas para el algoritmo

f) Dar un nombre al modelo y estructura de minería de datos

En la Figura N°9: Configuración de escenario 1 se presentan imágenes del proceso detallado.

FIGURA N° 9: CONFIGURACIÓN DE ESCENARIO 1



Autora: Gallegos, K., 2015

3.7. ESCENARIO 2 PARA ÁRBOL DE DECISIÓN

La preparación del escenario para el algoritmo de árbol de decisión en la herramienta SSDT se lo realiza mediante los siguientes pasos:

a) Seleccionar la definición el origen de la estructura de datos (relacional o multidimensional)

b) Seleccionar el algoritmo de Árbol de Decisión

c) Seleccionar el conjunto de datos de origen

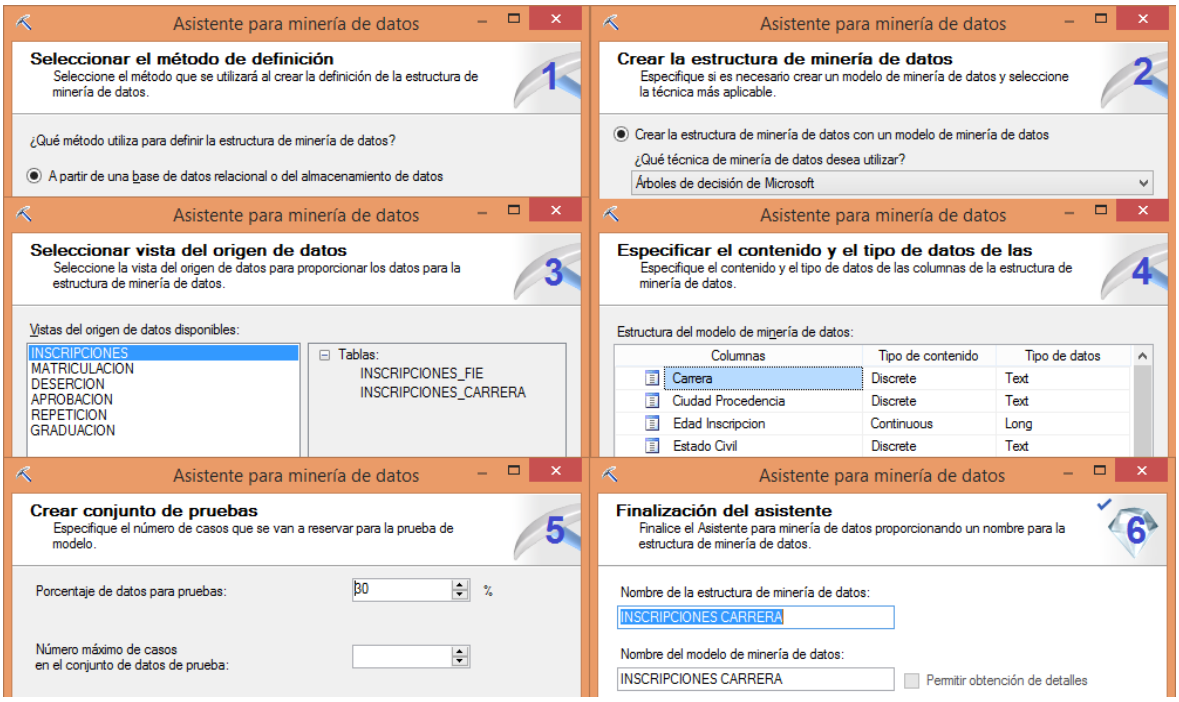
d) Especificar los tipos de datos

e) Definir el conjunto de aprendizaje y de pruebas para el algoritmo

f) Dar un nombre al modelo y estructura de minería de datos

Las diferencias entre establecer un escenario u otro, difiere en la selección del algoritmo de minería. En la Figura N°10: Configuración del escenario 2, se grafican los pasos para establecer el escenario para árbol de decisión.

FIGURA N° 10: CONFIGURACIÓN DE ESCENARIO 2



Autora: Gallegos, K, 2015

3.8. MEDICIÓN DE INDICADORES Y ANÁLISIS DE RESULTADOS

Con los escenarios listos, se procede a obtener los valores correspondientes a cada indicar de la variable Desempeño.

Los resultados de las mediciones obtenidas sobre los indicadores para el algoritmo de Regresión Logística se muestran en la Tabla 14: Medición para Regresión Logística.

TABLA 14: MEDICIÓN PARA REGRESIÓN LOGÍSTICA

ALGORITMO REGRESIÓN LOGÍSTICA				
REQUERIMIENTOS	**TIEMPO (SEG)**	**PRECISIÓN**	**USO DE CPU (%)**	**USO DE RAM (MB)**
REQ 1 (FACULTAD)	01	0,93	0,3	135,6
REQ 1 (EIS)	02	0,75	0,3	124,6
REQ 1 (EDG)	02	0,95	0,3	124,6
REQ 1 (EIECRI)	02	0,91	0,3	124,6
REQ 1 (EIETR)	02	0,86	0,3	124,6
REQ 3 (EIS -MATERIAS)	9	0,64	0,3	127,2
REQ 3 (EIS -NIVELES)	04	0,67	0,3	129,7
REQ 3 (EIS -AREAS)	05	0,53	0,7	125,7
REQ 4 (EIS)	01	0,73	0,3	82,5
REQ 4 (EIS-MATERIAS)	49	0,78	0,5	107,8
REQ 4 (EIS-AREAS)	04	0,86	0,1	92,5
REQ 5 (EIS)	05	0,42	0,4	120,7
REQ 5 (EIS-MATERIAS)	140	0,36	0,7	148,8
REQ 5 (EIS-AREAS)	05	0,70	0,6	170,4
REQ 6 (EIS)	01	0,68	0,3	173,3
REQ 6 (EIS-MATERIAS)	90	0,77	0,4	180,4
REQ 6 (EIS-AREAS)	05	0,82	0,4	182,5
REQ 8 (EIS)	01	0,70	1,1	189,1
REQ 3 (EDG-CRÉDITOS)	4	0,74	0,5	123,2
REQ 3 (EDG-MATERIAS)	13	0,91	0,3	124,5
REQ 3 (EDG -NIVELES)	3	0,47	0,5	103,1
REQ 3 (EDG -AREAS)	5	0,36	0,4	108,0
REQ 4 (EDG)	01	0,95	0,6	109,1
REQ 4 (EDG -	24	0,96	0,9	109,0

MATERIAS)				
REQ 4 (EDG -AREAS)	4	0,94	0,4	121,1
REQ 5 (EDG)	5	0,49	0,7	122,8
REQ 5 (EDG - MATERIAS)	73	0,90	0,7	127,3
REQ 5 (EDG -AREAS)	5	0,86	0,5	131,4
REQ 6 (EDG)	1	0,74	0,7	122,4
REQ 6 (EDG - MATERIAS)	9	0,93	0,7	136,4
REQ 6 (EDG -AREAS)	2	0,88	0,8	128,0
REQ 8 (EDG)	1	0,86	0,7	139,5
REQ 3 (EIECRI-CRÉDITOS)	1	0,94	0,3	122,4
REQ 3 (EIECRI - MATERIAS)	1	0,99	0,4	126,3
REQ 3 (EIECRI - NIVELES)	1	0,34	0,6	115,1
REQ 3 (EIECRI -AREAS)	4	0,32	0,6	143,9
REQ 4 (EIECRI)	4	0,95	0,7	136,6
REQ 4 (EIECRI - MATERIAS)	26	0,96	0,4	144,9
REQ 4 (EIECRI -AREAS)	2	1,00	0,6	142,1
REQ 5 (EIECRI)	3	0,53	0,3	146,0
REQ 5 (EIECRI - MATERIAS)	116	0,72	0,8	160,9
REQ 5 (EIECRI -AREAS)	20	0,67	0,5	159,9
REQ 6 (EIECRI)	1	0,65	0,7	148,8
REQ 6 (EIECRI - MATERIAS)	33	0,75	0,7	155,6
REQ 6 (EIECRI -AREAS)	5	0,75	0,6	163,4

REQ 8 (EIECRI)	1	0,37	0,6	157,9
REQ 3 (EIETR-CRÉDITOS)	1	0,87	0,2	123,9
REQ 3 (EIETR - MATERIAS)	1	1,00	1,3	135,7
REQ 3 (EIETR - NIVELES)	1	0,37	0,3	103,7
REQ 3 (EIETR -AREAS)	5	0,42	0,6	155,5
REQ 4 (EIETR)	1	0,95	0,6	146,9
REQ 4 (EIETR - MATERIAS)	14	0,95	0,9	148,2
REQ 4 (EIETR -AREAS)	1	0,99	0,6	149,5
REQ 5 (EIETR)	5	0,56	0,8	146,4
REQ 5 (EIETR - MATERIAS)	129	0,76	1,0	149,5
REQ 5 (EIETR -AREAS)	8	0,72	0,6	158,6
REQ 6 (EIETR)	1	0,69	0,3	150,7
REQ 6 (EIETR - MATERIAS)	23	0,79	0,8	156,9
REQ 6 (EIETR -AREAS)	2	0,77	0,7	165,7
REQ 8 (EIETR)	1	0,59	0,7	164,7
REQ 3 (EIS-MA-CRÉDITOS)	1	0,77	0,5	86,1
REQ 3 (EIS-MA-MATERIAS)	1	0,94	0,4	84,4
REQ 3 (EIS-MA-NIVELES)	1	0,63	0,7	84,1
REQ 3 (EIS-MA-AREAS)	1	0,96	0,7	116,1
REQ 4 (EIS-MA)	1	0,55	0,4	119,1
REQ 4 (EIS- MA-	1	0,74	0,5	121,3

MATERIAS)				
REQ 4 (EIS-MA-AREAS)	1	0,92	1,0	123,8
REQ 5 (EIS-MA)	1	0,48	0,2	122,1
REQ 5 (EIS-MA -AREAS)	1	0,95	0,6	136,9
REQ 3 (FIE-CRÉDITOS)	2	0,78	1,0	131,1
REQ 3 (FIE -MATERIAS)	2	0,62	0,7	129,9
REQ 3 (FIE -NIVELES)	5	0,26	1,0	118,3
REQ 4 (FIE)	3	0,84	1,2	134,1
REQ 5 (FIE)	56	0,46	0,5	137,1
REQ 6 (FIE)	6	0,77	0,7	137,3
REQ 8 (FIE)	1	0,81	0,6	127,5
TAMAÑO MUESTRA	76	76	76	76
MEDIA	12,80	0,74	0,58	133,68
DESVIACIÓN ESTÁNDAR	28,42	0,20	0,24	22,76

Autora: Gallegos K., 2015

Los resultados de las mediciones obtenidas sobre los indicadores para el algoritmo de Árbol de Decisión se muestran en la Tabla 15: Medición para Árbol de Decisión.

TABLA 15: MEDICIÓN PARA ÁRBOL DE DECISIÓN

ALGORITMO ÁRBOL DE DECISIÓN				
REQUERIMIENTOS	**TIEMPO (SEG)**	**PRECISIÓN**	**USO DE CPU (%)**	**USO DE RAM (MB)**
REQ 1 (FACULTAD)	10	0,98	0,4	135,0
REQ 1 (EIS)	14	0,73	0,5	120,6
REQ 1 (EDG)	14	0,94	0,5	120,6
REQ 1 (EIECRI)	14	0,90	0,5	120,6
REQ 1 (EIETR)	14	0,88	0,5	120,6
REQ 3 (EIS -MATERIAS)	9	0,64	0,3	127,2

REQ 3 (EIS -NIVELES)	10	0,93	0,4	125,6
REQ 3 (EIS -AREAS)	13	0,71	1,3	129,8
REQ 4 (EIS)	9	0,98	1,0	82,3
REQ 4 (EIS-MATERIAS)	03	0,37	1,1	166,7
REQ 4 (EIS-AREAS)	04	0,93	0,4	107,7
REQ 5 (EIS)	20	0,56	0,6	116,6
REQ 5 (EIS-MATERIAS)	07	0,59	0,7	151,4
REQ 5 (EIS-AREAS)	13	0,99	0,6	172,2
REQ 6 (EIS)	08	0,85	0,9	162,9
REQ 6 (EIS-MATERIAS)	04	0,56	1,0	177,0
REQ 6 (EIS-AREAS)	07	0,94	0,6	186,1
REQ 8 (EIS)	04	0,79	0,5	185,5
REQ 3 (EDG-CRÉDITOS)	4	0,88	0,3	101,6
REQ 3 (EDG-MATERIAS)	13	0,99	0,5	103,6
REQ 3 (EDG -NIVELES)	5	0,48	0,5	102,6
REQ 3 (EDG -AREAS)	7	0,50	0,3	101,6
REQ 4 (EDG)	7	0,97	0,6	102,0
REQ 4 (EDG - MATERIAS)	4	0,56	0,6	113,8
REQ 4 (EDG -AREAS)	5	0,97	0,3	115,9
REQ 5 (EDG)	10	0,50	0,4	121,9
REQ 5 (EDG - MATERIAS)	04	0,92	0,4	127,2
REQ 5 (EDG -AREAS)	4	1,00	0,6	125,3
REQ 6 (EDG)	6	0,86	0,5	124,5
REQ 6 (EDG - MATERIAS)	3	0,77	0,5	125,0
REQ 6 (EDG -AREAS)	5	0,91	0,9	142,0

REQ 8 (EDG)	2	0,55	0,6	135,7
REQ 3 (EIECRI-CRÉDITOS)	3	0,98	0,4	122,2
REQ 3 (EIECRI - MATERIAS)	5	0,49	0,7	125,8
REQ 3 (EIECRI - NIVELES)	11	0,38	0,4	92,3
REQ 3 (EIECRI -AREAS)	10	0,55	0,4	145,0
REQ 4 (EIECRI)	8	0,98	0,8	132,9
REQ 4 (EIECRI - MATERIAS)	4	0,41	0,7	145,6
REQ 4 (EIECRI -AREAS)	5	1,00	1,2	143,0
REQ 5 (EIECRI)	13	0,58	0,5	151,7
REQ 5 (EIECRI - MATERIAS)	5	0,65	0,5	159,3
REQ 5 (EIECRI -AREAS)	8	0,85	0,6	160,7
REQ 6 (EIECRI)	8	0,90	1,0	151,2
REQ 6 (EIECRI - MATERIAS)	3	0,63	0,5	150,1
REQ 6 (EIECRI -AREAS)	8	0,87	0,5	164,2
REQ 8 (EIECRI)	3	0,66	0,7	154,4
REQ 3 (EIETR-CRÉDITOS)	13	0,95	1,0	123,5
REQ 3 (EIETR - MATERIAS)	9	0,53	1,2	135,7
REQ 3 (EIETR - NIVELES)	8	0,46	0,3	101,7
REQ 3 (EIETR -AREAS)	10	0,72	0,8	154,6
REQ 4 (EIETR)	7	0,98	0,8	141,5
REQ 4 (EIETR -	3	0,51	0,7	148,5

MATERIAS)				
REQ 4 (EIETR -AREAS)	4	0,97	1,0	140,0
REQ 5 (EIETR)	11	0,58	0,8	142,9
REQ 5 (EIETR - MATERIAS)	4	0,69	0,8	149,0
REQ 5 (EIETR -AREAS)	9	0,96	0,9	148,9
REQ 6 (EIETR)	6	0,89	0,9	156,6
REQ 6 (EIETR - MATERIAS)	4	0,51	1,0	157,3
REQ 6 (EIETR -AREAS)	8	0,71	0,7	165,2
REQ 8 (EIETR)	6	0,58	1,1	158,4
REQ 3 (EIS-MA-CRÉDITOS)	2	0,79	0,5	84,5
REQ 3 (EIS-MA-NIVELES)	5	0,65	0,5	80,6
REQ 3 (EIS-MA-AREAS)	4	0,98	1,0	116,4
REQ 4 (EIS-MA)	5	1,00	0,5	122,5
REQ 4 (EIS- MA-MATERIAS)	2	0,34	0,8	124,8
REQ 4 (EIS-MA-AREAS)	2	0,97	0,7	124,1
REQ 5 (EIS-MA)	5	0,53	0,6	122,9
REQ 5 (EIS-MA-MATERIAS)	2	0,23	0,2	129,4
REQ 5 (EIS-MA -AREAS)	5	0.99	0,5	137,8
REQ 3 (FIE-CRÉDITOS)	7	0,82	0,8	129,9
REQ 3 (FIE -MATERIAS)	10	0,65	0,6	131,2
REQ 3 (FIE -NIVELES)	39	0,35	0,4	124,3
REQ 4 (FIE)	20	0,98	0,6	123,6
REQ 5 (FIE)	67	0,55	0,6	125,6
REQ 6 (FIE)	18	0,87	0,6	135,0

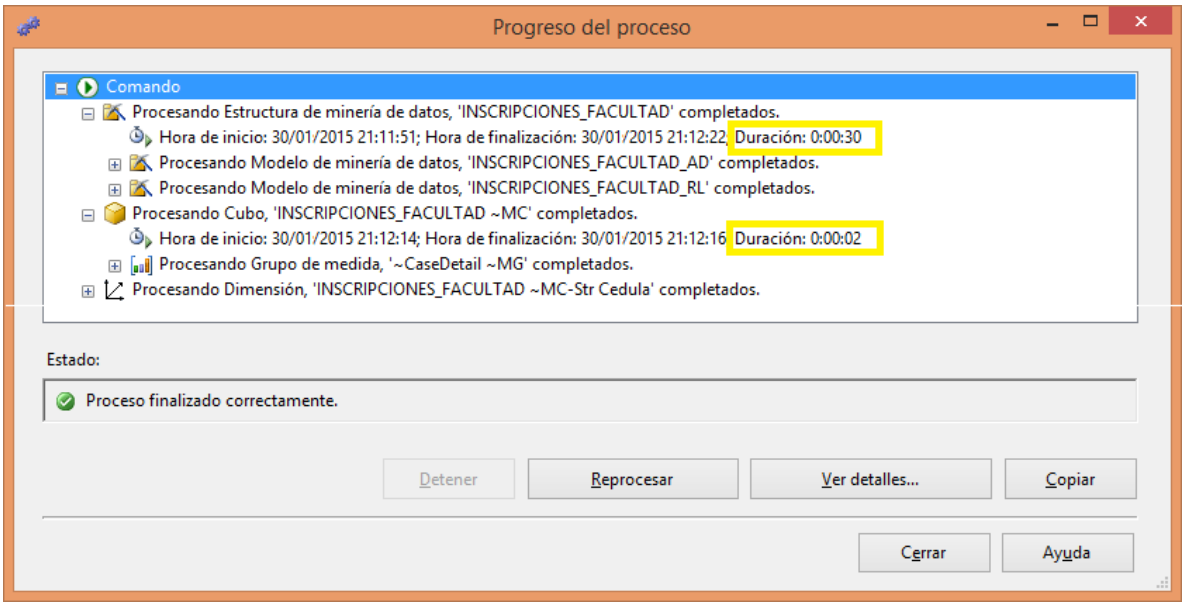
REQ 8 (FIE)	4	0,86	0,8	136,0
TAMAÑO MUESTRA	76	76	76	76
MEDIA	8,54	0,74	0,65	133,18
DESVIACIÓN ESTÁNDAR	8,77	0,21	0,25	22,77

Autora: Gallegos, K., 2015

3.8.1. TIEMPO DE RESPUESTA

El indicador tiempo de respuesta fue medido en segundos con la ayuda de la herramienta SSDT de Microsoft, la cual después de generar el modelo emite el tiempo total utilizado en el proceso de ejecución, como se muestra en la Figura N°11: Medición del tiempo.

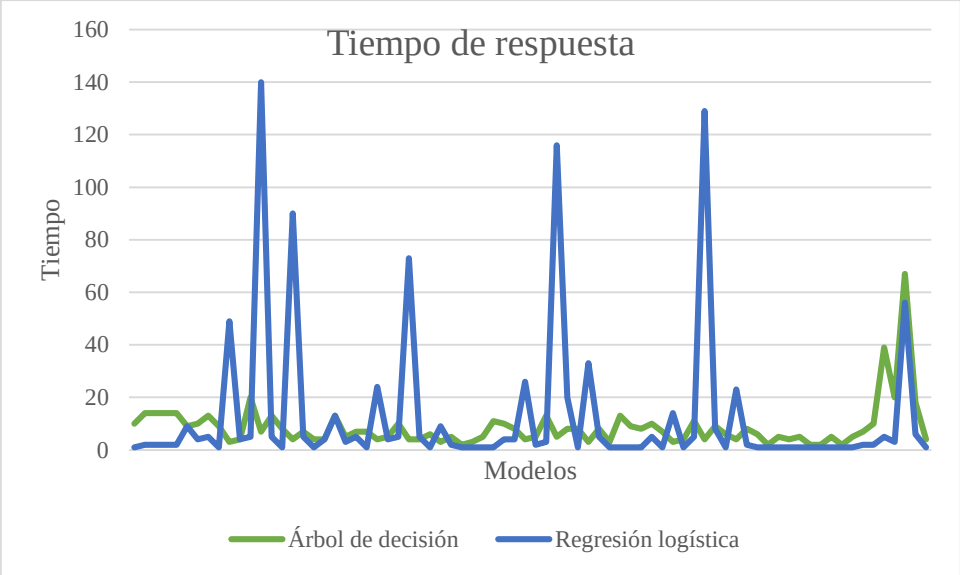
FIGURA N° 11: MEDICIÓN DEL TIEMPO



Autora: Gallegos, K., 2015

La Figura N°12: Tiempo de Respuesta, presenta un gráfico resumen de los resultados obtenidos de este proceso.

FIGURA N° 12: TIEMPO DE RESPUESTA



Autora: Gallegos K., 2015

La estadística descriptiva sobre los resultados del tiempo de respuesta de los algoritmos arrojan los resultados de la Tabla 16: Estadística Descriptiva Tiempo de Respuesta.

TABLA 16: ESTADÍSTICA DESCRIPTIVA TIEMPO DE RESPUESTA

Medida\Algoritmo	Regresión Logística	Árbol de decisión
Conteo	76	76
Media	12,80	8,54
Desviación estándar	28,42	8,77

Autora: Gallegos, K., 2015

Interpretación: Dado que un mayor tiempo de respuesta es negativo para el desempeño de un algoritmo se asigna el 10% al algoritmo de árbol de decisión y su equivalente al algoritmo de regresión logística, el mismo que es calculado con una regla de tres inversa con los valores de las medias como se muestra a continuación:

$$8{,}54 \Rightarrow \quad 10\%$$
$$12{,}8 \Rightarrow \quad x\%$$

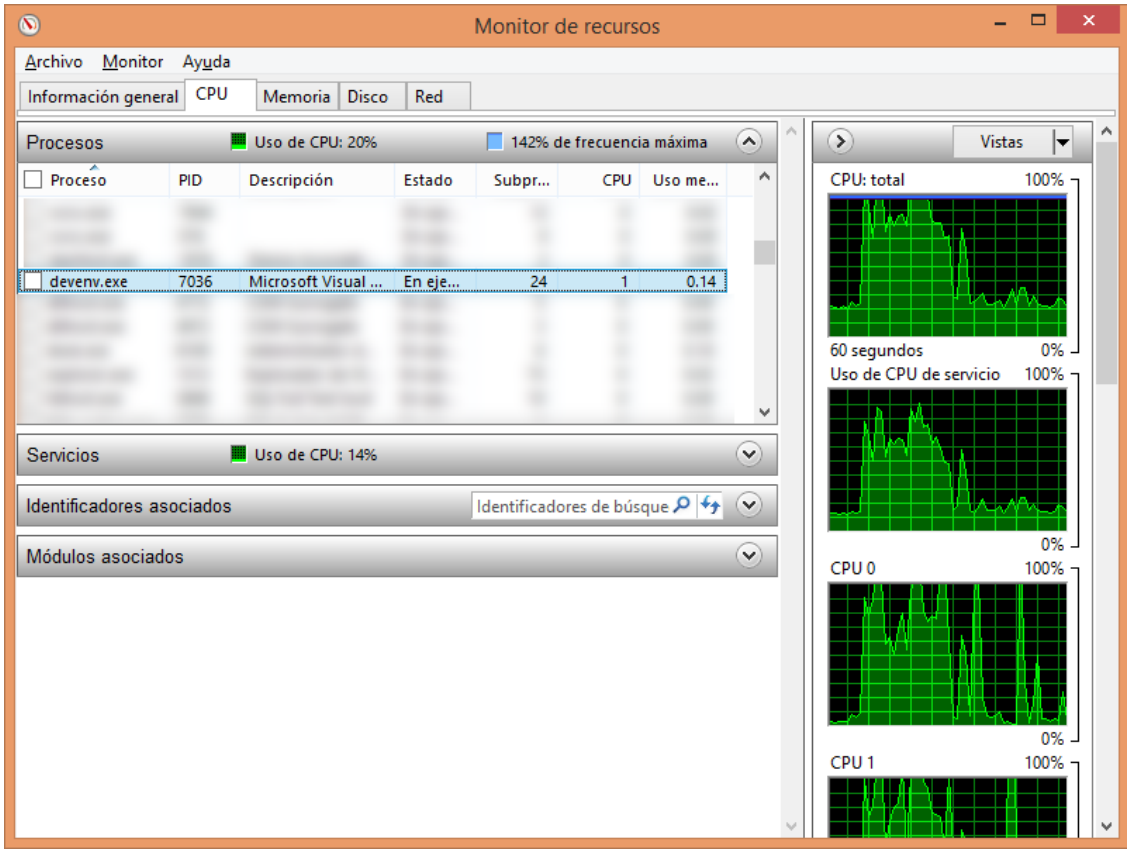
Como resultado final se obtiene:

- Algoritmo de regresión logística: 6.67%
- Algoritmo de árbol de decisión: 10%

3.8.2. USO DEL CPU

El indicador Uso del CPU fue tomado en porcentaje con la ayuda del monitor de sistema de Microsoft, como se muestra en la Figura N°13: Medición del uso del CPU.

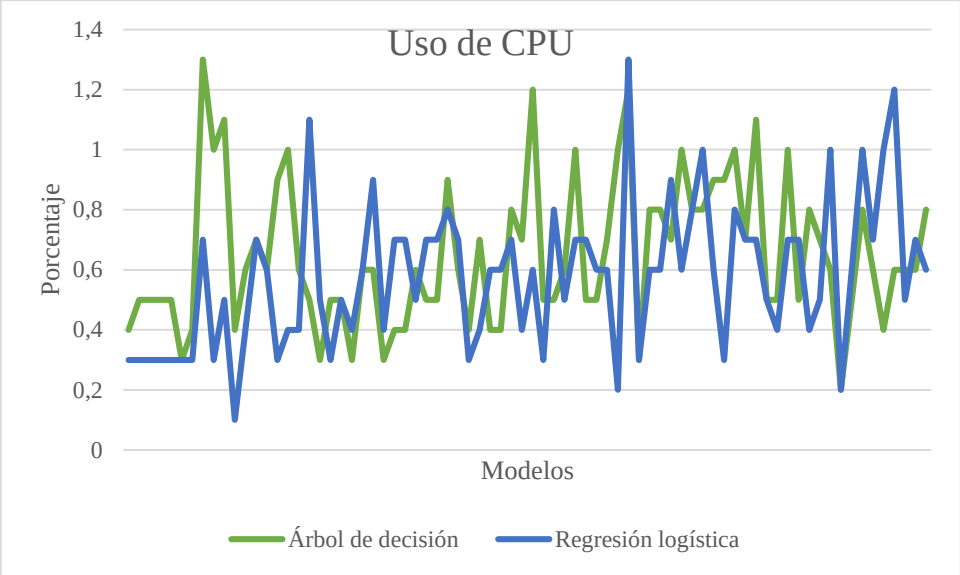
FIGURA N° 13: MEDICIÓN DEL USO DE CPU



Autora: Gallegos, K., 2015

La Figura N°14: Uso del CPU, presenta un gráfico resumen de los resultados obtenidos de este proceso.

FIGURA N° 14: USO DEL CPU



Autora: Gallegos, K., 2015

La estadística descriptiva sobre los resultados del uso del CPU de los algoritmos arrojan los resultados que se muestran en la Tabla 17: Estadística Descriptiva Uso del CPU.

TABLA 17: ESTADÍSTICA DESCRIPTIVA USO DE CPU

Medida\Algoritmo	Regresión Logística	Árbol de decisión
Conteo	76	76
Media	0,58	0,65
Desviación estándar	0,24	0,25

Autora: Gallegos, K., 2015

Interpretación: Dado que un menor porcentaje de uso del CPU es positivo para el desempeño de un algoritmo se asigna el 10% al algoritmo de regresión logística y su equivalente al algoritmo de árbol de decisión, el mismo que es calculado con una regla de tres inversa con los valores de las medias como se muestra a continuación:

0.58 => 10%

0.65 => x%

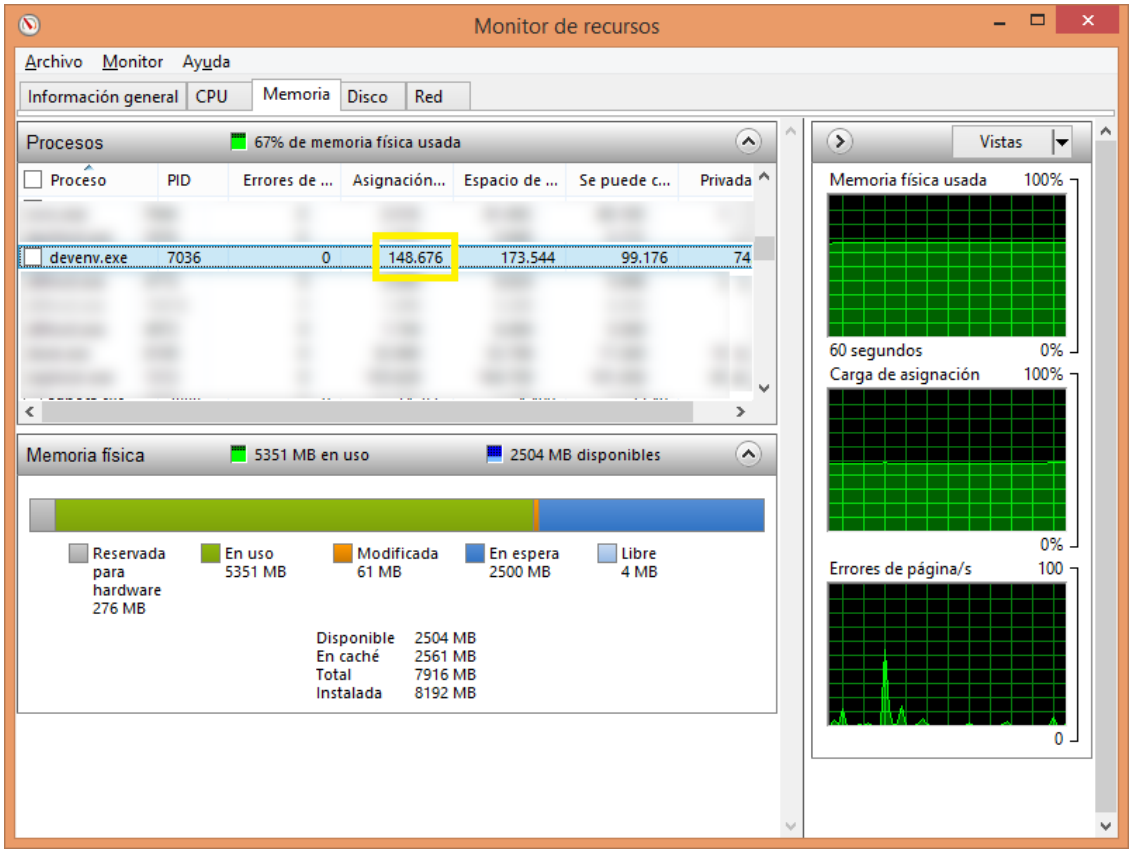
Como resultado final se obtiene:

- Algoritmo de regresión logística: 10%
- Algoritmo de árbol de decisión: 8,92%

3.8.3. USO DE RAM

El indicador Uso de RAM fue tomado en Megabytes con la ayuda del monitor de sistema de Microsoft, como se indica en la Figura N°15: Medición del uso de RAM.

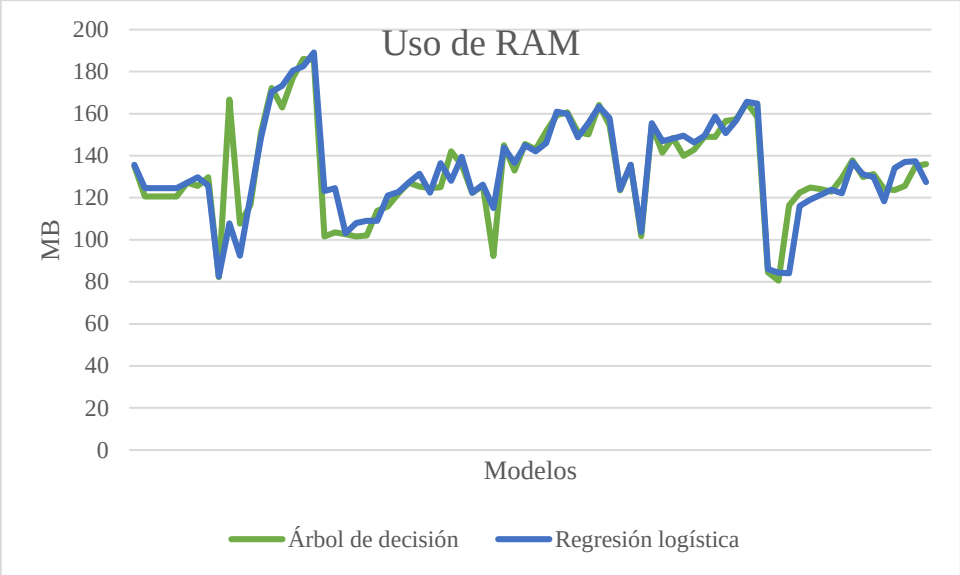
FIGURA N° 15: MEDICIÓN DEL USO DE RAM



Autora: Gallegos, K., 2015

El Figura N°16 Uso de RAM, presenta un gráfico resumen de los resultados obtenidos de este proceso.

FIGURA N° 16: USO DE RAM



Autora: Gallegos, K., 2015

La estadística descriptiva sobre los resultados del uso de RAM de los algoritmos arrojan los resultados que se muestran en la Tabla 18: Estadística Descriptiva Uso de RAM.

TABLA 18: ESTADÍSTICA DESCRIPTIVA USO DE RAM

Medida\Algoritmo	Regresión Logística	Árbol de decisión
Conteo	76	76
Media	133,68	133,18
Desviación estándar	22,76	22,77

Autora: Gallegos, K., 2015

Interpretación: Dado que un mayor porcentaje de uso de RAM es negativo para el desempeño de un algoritmo se asigna el 10% al algoritmo de árbol de decisión y su equivalente al algoritmo de regresión logística, el mismo que es calculado con una regla de tres inversa con los valores de las medias como se muestra a continuación:

133,18 => 10%

133.68 => x%

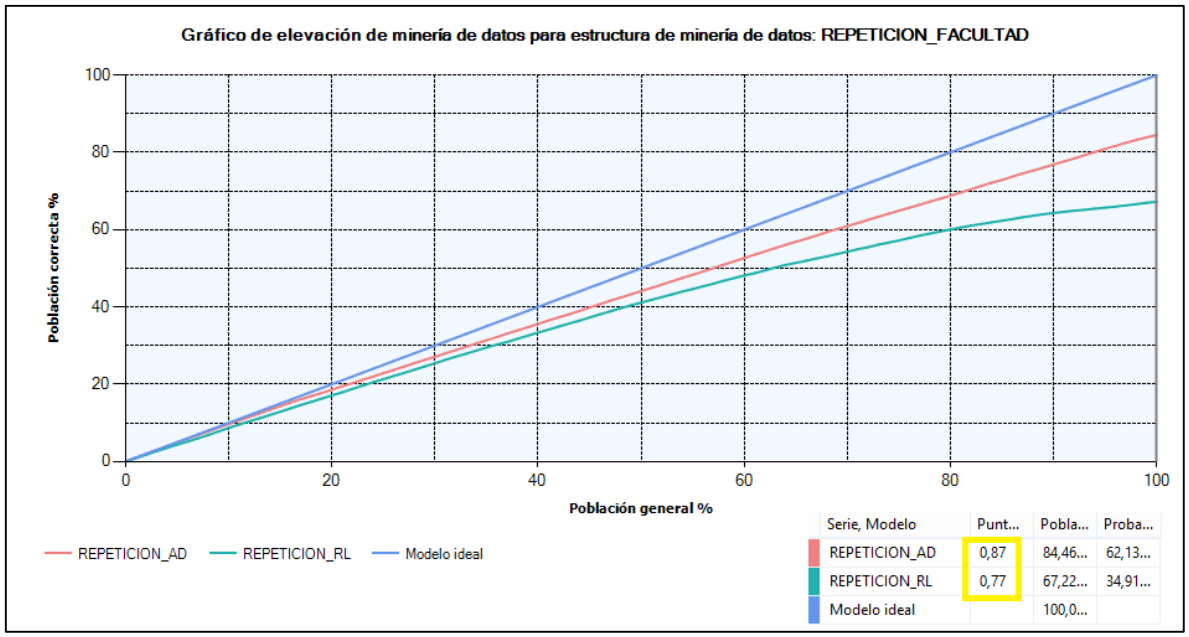
Como resultado final se obtiene:

- Algoritmo de regresión logística: 9,96%
- Algoritmo de árbol de decisión: 10 %

3.8.4. PRECISIÓN

El indicador precisión es un valor que se ha tomado de la herramienta Data Tools de SSAS de Microsoft, como se muestra la Figura N° 17: Medición de la precisión.

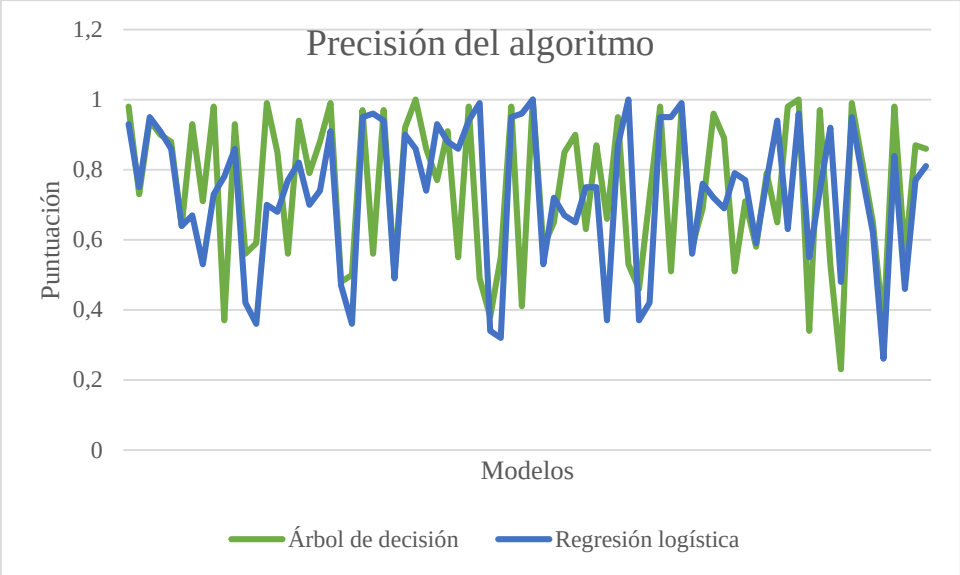
FIGURA N° 17: MEDICIÓN DE LA PRECISIÓN



Autora: Gallegos, K., 2015

La Figura N°18 Precisión, presenta un gráfico resumen de los resultados obtenidos de este proceso.

FIGURA N° 18: PRECISIÓN



Autora: Gallegos, K., 2015

La estadística descriptiva sobre los resultados de la precisión de los algoritmos arrojan los datos que se muestran en la Tabla 19: Estadística Descriptiva Precisión.

TABLA 19: ESTADÍSTICA DESCRIPTIVA PRECISIÓN

Medida\Algoritmo	Regresión Logística	Árbol de decisión
Conteo	76	76
Media	0,74	0,75
Desviación estándar	0,20	0,21

Autora: Gallegos, K., 2015

Interpretación: Dado que a mayor precisión mejor desempeño de un algoritmo, se asigna el 70% al algoritmo de árbol de decisión y su equivalente al algoritmo de regresión logística, el mismo que es calculado con una regla de tres con los valores de las medias como se muestra a continuación:

0.75 => 70%

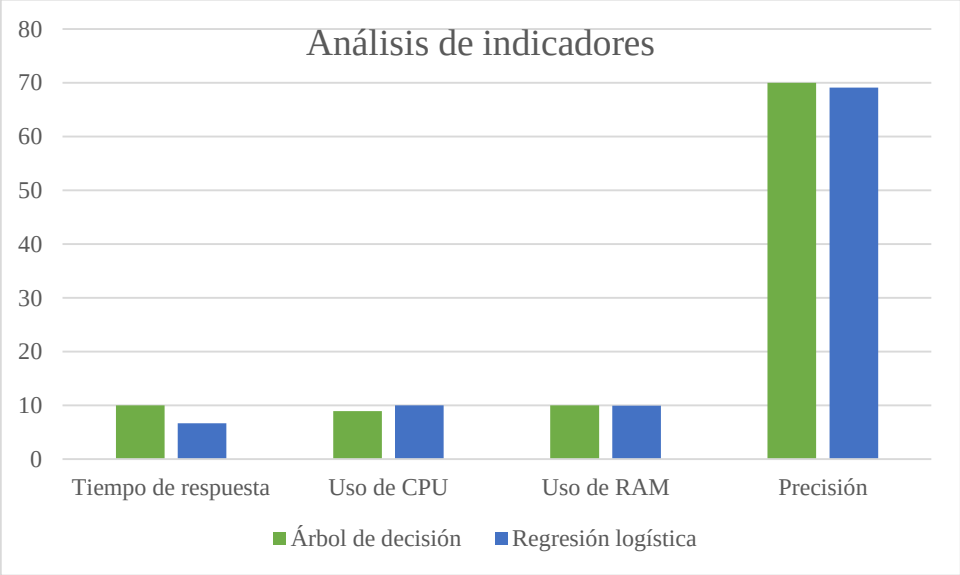
0.74 => x%

Como resultado final se obtiene:

- Algoritmo de regresión logística: 69.07%
- Algoritmo de árbol de decisión: 70%

Los resultados de los análisis de cada uno de los indicadores de desempeño se muestran en la Figura N°19: Resultados de análisis de indicadores.

FIGURA N° 19: RESULTADOS DE ANÁLISIS DE INDICADORES



Autora: Gallegos, K., 2015

En la gráfica anterior se puede apreciar que los resultados son muy parejos en cada uno de los indicadores usados para el estudio de los algoritmos. Para obtener un resultado final se suman de los datos derivados de cada indicador, ver la Tabla 20: Resultados del análisis de desempeño

TABLA 20: RESULTADOS DEL ANÁLISIS DE DESEMPEÑO

Indicador\Algoritmo	Regresión logística	Árbol de decisión
Tiempo de respuesta	6,67%	10%
Uso de CPU	10%	8,92%

Uso de RAM	9,96%	10%
Precisión	69,07%	70%
TOTAL	**95,70%**	**98,92%**

Autora: Gallegos, K., 2015

Con un 98.92% obtenido en el análisis de desempeño, el algoritmo Árbol de Decisión supera al algoritmo de Regresión Logística que obtuvo un 95,70% del desempeño.

3.9. PRUEBA DE LA HIPÓTESIS

La hipótesis de investigación, basándose en la variable de desempeño, para los algoritmos de minería de datos seleccionados es:

- El algoritmo de minería de datos de regresión logística tiene mejor desempeño que el algoritmo de árbol de decisión para obtener datos proyectados sobre la actividad académica en la Facultad de Informática y Electrónica de la ESPOCH.

Siendo la hipótesis nula:

- El algoritmo de minería de datos de regresión logística no tiene mejor desempeño que el algoritmo de árbol de decisión para obtener datos proyectados sobre la actividad académica en la Facultad de Informática y Electrónica de la ESPOCH.

Se aplicará estadística inferencial utilizando la prueba Z para dos muestras, a partir de lo cual se puede tomar decisiones sobre los resultados recogidos. La fórmula para obtener z de prueba es:

ECUACIÓN 2: PRUEBA Z

$$Zp = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Donde:

$$\bar{x}_1 = Media\ de\ la\ proporción\ 1$$

$$\bar{x}_2 = Media\ de\ la\ proporción\ 2$$

$$S_1 = Varianza\ de\ la\ proporción\ 1$$

$$S_2 = Varianza\ de\ la\ proporción\ 2$$

$$n_1 = Tamaño\ de\ la\ proporción\ 1$$

$$n_2 = Tamaño\ de\ la\ proporción\ 2$$

Tiempo de respuesta

Las hipótesis nula y de investigación para el tiempo de respuesta son:

Hi: El algoritmo de regresión logística tiene mejor de respuesta que el algoritmo árbol de decisión.

Ho: El algoritmo de regresión logística no tiene mejor tiempo de respuesta que el algoritmo árbol de decisión.

Con el nivel de significancia del 5% el Z crítico para dos colas es 1,96 lo que define que:

- Todos los valores menores a -1,96 o mayores a 1,96 caen en la zona de rechazo de la hipótesis nula
- Todos los valores mayores a -1,96 y menores 1.96 caen en la zona de aceptación de la hipótesis nula

El Z de prueba para el tiempo de respuesta es: 1.24. Con este resultado se procede a aceptar la hipótesis nula: El algoritmo de regresión logística no tiene mejor tiempo de respuesta que el algoritmo árbol de decisión.

Uso de CPU

Las hipótesis nula y de investigación para el uso del CPU son:

Hi: El algoritmo de regresión logística tiene mejor porcentaje de uso del CPU que el algoritmo árbol de decisión.

Ho: El algoritmo de regresión logística no tiene mejor porcentaje de uso del CPU que el algoritmo árbol de decisión.

Con el nivel de significancia del 5% el Z crítico para dos colas es 1,96 lo que define que:

- Todos los valores menores a -1,96 o mayores a 1,96 caen en la zona de rechazo de la hipótesis nula
- Todos los valores mayores a -1,96 y menores 1.96 caen en la zona de aceptación de la hipótesis nula

El Z de prueba para el uso del CPU es: -1,76. Con este resultado se procede a aceptar la hipótesis nula: El algoritmo de regresión logística no tiene mejor porcentaje de uso del CPU que el algoritmo árbol de decisión.

Uso de RAM

Las hipótesis nula y de investigación para el uso de RAM son:

Hi: El algoritmo de regresión logística tiene mejor uso de RAM que el algoritmo árbol de decisión.

Ho: El algoritmo de regresión logística no tienen mejor uso de RAM que el algoritmo árbol de decisión.

Con el nivel de significancia del 5% el Z crítico para dos colas es 1,96 lo que define que:

- Todos los valores menores a -1,96 o mayores a 1,96 caen en la zona de rechazo de la hipótesis nula
- Todos los valores mayores a -1,96 y menores 1.96 caen en la zona de aceptación de la hipótesis nula

El Z de prueba para el uso de RAM es: 0.13. Con este resultado se procede aceptar la hipótesis nula: El algoritmo de regresión logística no tiene mejor uso de RAM que el algoritmo árbol de decisión.

Precisión

Las hipótesis nula y de investigación para la precisión de los algoritmos:

Hi: El algoritmo de regresión logística tiene mejor precisión que el algoritmo árbol de decisión.

Ho: El algoritmo de regresión logística no tiene mejor precisión que el algoritmo árbol de decisión.

Con el nivel de significancia del 5% el Z crítico para dos colas es 1,96 lo que define que:

- Todos los valores menores a -1,96 o mayores a 1,96 caen en la zona de rechazo de la hipótesis nula
- Todos los valores mayores a -1,96 y menores 1.96 caen en la zona de aceptación de la hipótesis nula

El Z de prueba para la precisión es: -0.30. Con este resultado se procede aceptar la hipótesis nula: El algoritmo de regresión logística no tiene mejor precisión que el algoritmo árbol de decisión.

Los resultados se resumen en la Tabla 21: Resumen de comprobación de hipótesis

TABLA 21: RESUMEN DE COMPROBACIÓN DE HIPÓTESIS

Hipótesis de Trabajo	**Z tabular**	**Z calculado**	**Observación**
$H_{o\text{-}tiempo}$: El algoritmo de regresión logística no tiene mejor tiempo de respuesta que el algoritmo árbol de decisión.	-1.96< z >1,96	1,24	Se acepta
$H_{o\text{-}CPU}$: El algoritmo de regresión logística no tiene mejor porcentaje de uso del CPU que el algoritmo árbol de decisión.		-1,76	Se acepta
$H_{o\text{-}RAM}$: El algoritmo de regresión logística no tiene mejor uso de RAM que el algoritmo árbol de decisión.		0.13	Se acepta
$H_{o\text{-}precisión}$: El algoritmo de regresión logística no tiene mejor precisión que el algoritmo árbol de decisión.		-0,30	Se acepta
Ho: El algoritmo de regresión logística no tiene mejor desempeño que el algoritmo de árbol de decisión			**Se acepta**

Autora: Gallegos, K., 2015

Dado que todas las hipótesis nulas se han aceptado, se procede a aceptar la hipótesis nula con respecto al desempeño de los algoritmos seleccionados:

- El algoritmo de minería de datos de regresión logística tiene no tiene mejor desempeño que el algoritmo de árbol de decisión para obtener datos proyectados sobre la actividad académica en la Facultad de Informática y Electrónica de la ESPOCH.

Por lo tanto el algoritmo de Árbol de Decisión será utilizado en la implementación del proyecto de minería de datos sobre los datos académicos de la FIE-ESPOCH.

4. CAPÍTULO IV: IMPLEMENTACIÓN DE MINERIA DE DATOS

Con el algoritmo de minería de datos seleccionado, se procede a desarrollar el proyecto de minería de datos siguiendo la metodología Cross Industry Process for Data Mining descrita en el capítulo anterior. El análisis del negocio permite establecer las condiciones previas al desarrollo del proyecto. El análisis y la preparación de los datos consiste en una revisión y limpieza de los valores necesarios para el proyecto de minería, para después desarrollar las estructuras de datos sobre las cuales se aplicará el algoritmo de Árbol de Decisión para la obtención de los modelos con los que se cumplen los objetivos trazados para el proyecto.

Dentro de este capítulo se presenta un resumen de las actividades realizadas dentro de cada fase del proyecto de minería de datos en la Facultad de Informática y Electrónica de la ESPOCH.

4.1. ANÁLISIS DEL NEGOCIO

Del análisis del negocio se obtienen los objetivos a perseguir con la de minería de datos por lo que se plantea como primer paso dentro del proyecto. El objetivo general del proyecto es el buscar patrones de comportamiento dentro de los estudiantes de la Facultad de Informática y Electrónica de ESPOCH sobre la plataforma de datos centralizada que maneja la institución; para lo cual se han definido objetivos específicos en base a los ejes

del proceso académico que son ingreso, matriculación, promoción y graduación; estos objetivos específicos resultantes el análisis del negocio son los siguientes:

- REQ 1: Determinar patrones de comportamiento para el ingreso de los estudiantes; por carrera y facultad.
- REQ 2: Determinar patrones de comportamiento para el ingreso directo a la carrera (sin pasar por el curso de nivelación); por carrera y facultad.
- REQ 3: Determinar patrones de comportamiento para la matriculación y selección de asignatura de los estudiantes (número de asignaturas, número de créditos, nivel y área de las asignaturas, etc.); por carrera, por nivel y por facultad.
- REQ 4: Determinar los factores que tienen influencia en los casos de deserción (retiros y pérdida de la asignatura por asistencia), por carrera, niveles, asignaturas, áreas de conocimiento y facultad.
- REQ 5: Determinar patrones de comportamiento en la promoción académica de los estudiantes; por asignatura, nivel, áreas de conocimiento, carrera y facultad.
- REQ 6: Determinar los factores que influyen en los escenarios de segunda y tercera matrícula; por asignatura, nivel, áreas de conocimiento, carrera y facultad.
- REQ 7: Determinar la proyección de graduados; por carrera y facultad.
- REQ 8: Determinar los factores que inciden en los casos de estudiantes con baja eficiencia terminal; por carrera y facultad.

Con los objetivos planteados es posible realizar un análisis de los riesgos que implica el desarrollo de un proyecto de minería de datos, con el propósito de determinar si los posibles conflictos afectarían de manera radical el desarrollo del proyecto así como contar con planes de contingencia en caso de que alguno de ellos llegase a presentarse. En la Tabla

22: Matriz de riegos se presenta el listado de los posibles conflictos del desarrollo del presente proyecto de minería.

TABLA 22: MATRIZ DE RIESGOS

Código	Riesgo	Probabilidad	Impacto	Categoría
MDR01	El usuario decide no continuar con el proyecto de minería de datos	Bajo	Alto	Alto
MDR04	Los resultados del análisis de minería de datos no cumplen con los objetivos del negocio	Medio	Alto	Alto
MDR05	No se cuenta con acceso a los recursos de datos para el análisis	Medio	Alto	Alto
MDR02	Los algoritmos no proveen el porcentaje de precisión requerido por el usuario	Bajo	Medio	Medio
MDR03	No se cuenta con la infraestructura que soporte el software de minería de datos	Medio	Medio	Medio
MDR06	Los datos de origen no son actualizados	Medio	Bajo	Medio
MDR07	Los datos de origen no han pasado por un proceso de calidad	Alto	Bajo	Medio
MDR08	Los atributos de los datos de origen no son suficientes para aplicar un algoritmo de minería de datos	Bajo	Alto	Medio
MDR09	No se cuenta con conocimientos para resolver problemas dentro del desarrollo de minería de datos	Bajo	Bajo	Bajo

Autora: Gallegos, K., 2015

El contar con riesgos de nivel alto y medio implica directamente el desarrollo de acciones de contingencia para prevenir consecuencias determinantes sobre el proyecto de minería. Las acciones determinadas según los riesgos encontrados se listan en la tabla 23: Mitigación de riesgos.

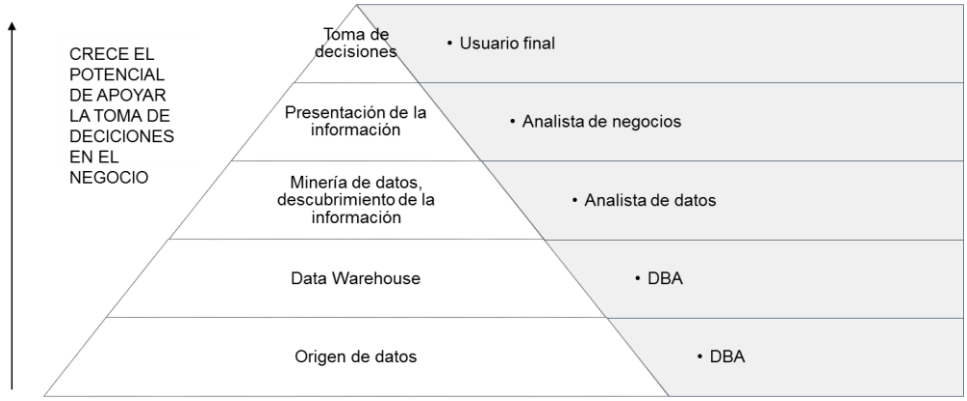
TABLA 23: MITIGACIÓN DE RIESGOS

Código	Riesgo	Mitigación
MDR01	El usuario decide no continuar con el proyecto de minería de datos	Exponer nuevamente los beneficios del proyecto de minería de datos.
MDR04	Los resultados del análisis de minería de datos no cumplen con los objetivos del negocio	Analizar nuevamente los requerimientos del usuario final, y si es necesario replantear los objetivos del negocio.
MDR05	No se cuenta con acceso a los recursos de datos para el análisis	Exponer la necesidad de acceder a los datos marcados como recursos disponibles para el avance del proyecto de minería de datos
MDR02	Los algoritmos no proveen el porcentaje de precisión requerido por el usuario	Reconfigurar los parámetros de entrada de los algoritmos
MDR03	No se cuenta con la infraestructura que soporte el software de minería de datos	Instalar servidores virtuales en un servidor provisional.
MDR06	Los datos de origen no son actualizados	Trabajar con los datos proporcionados, debido a que es posible mantener el modelado resultante con los datos no actualizados.
MDR07	Los datos de origen no han pasado por un proceso de calidad	La metodología de minería de datos destina una parte del ciclo del proyecto a la limpieza de los datos, recurrir a las herramientas oportunas para el proceso de calidad necesaria previo al modelado.
MDR08	Los atributos de los datos de origen no son suficientes para aplicar un algoritmo de minería de datos	Creación de atributos derivados de los datos, que representen un cambio de forma, mas no de fondo del origen proporcionado.

Autora: Gallegos, K., 2015

El beneficio de realizar un proyecto de minería se refleja en el apoyo a la toma de decisiones como se presenta en la Figura N°20: Potencial de Apoyo para la Toma de Decisiones.

FIGURA N° 20: POTENCIAL DE APOYO PARA LA TOMA DE DECISIONES



Fuente: Vallejos, S., 2006

Con la minería de datos se alcanza hasta un tercer nivel de la pirámide de apoyo a la toma de decisiones por parte del usuario final. El descubrimiento de la información contribuirá al análisis de los procesos académicos de la Facultad de Informática y Electrónica de la ESPOCH en base a los objetivos definidos en este análisis. Más información sobre el análisis del negocio se incluye en el Anexo 1: Informe del Análisis del Negocio.

4.2. ANÁLISIS DE LOS DATOS

El procedimiento del análisis de los datos busca identificar si existe una cantidad suficiente de filas que serán usadas para el desarrollo de las estructuras de datos. En el presente proyecto se analizaron las tablas de las bases de datos asociadas a los procesos de inscripción, de matriculación, de promoción y graduación, para determinar si existen: tanto un número de superior a 200 filas, así como al menos cinco atributos de uso potencial para el modelado (datos que no correspondan a claves autogeneradas y fechas). Los resultados se los sintetizó en la Tabla 24: Análisis de datos, que se presenta a continuación.

TABLA 24: ANÁLISIS DE DATOS

Requisito	**Segregación**	**Observaciones**	**Número de modelos**
REQ 1: Determinar patrones de comportamiento para el ingreso de los estudiantes; por carrera y facultad.	Facultad	Cumple	1
	EIS	Cumple	1
	EIETR	Cumple	1
	EIECRI	Cumple	1
	EDG	Cumple	1
	EIS ext. Macas	No cumple	0
REQ 2: Determinar patrones de comportamiento para el ingreso directo a la carrera (sin pasar por el curso de nivelación); por carrera y facultad.	Facultad	No cumple	0
	EIS	No cumple	0
	EIETR	No cumple	0
	EIECRI	No cumple	0
	EDG	No cumple	0
	EIS ext. Macas	No cumple	0
REQ 3: Determinar patrones de comportamiento para la matriculación y selección de asignatura de los estudiantes (número de asignaturas, número de créditos, nivel y área de las asignaturas, etc.); por carrera, por nivel y por facultad.	Facultad	Cumple	4
	EIS	Cumple	4
	EIS-Niveles (10)	Cumple	20
	EIETR	Cumple	4
	EIETR-Niveles(10)	Cumple	20
	EIECRI	Cumple	4
	EIECRI-Niveles(10)	Cumple	20
	EDG	Cumple	4

	EDG-Niveles(10)	Cumple	20
	EIS ext. Macas	Cumple	4
	EIS ext. Macas – Niveles (10)	No cumple	0
REQ 4: Determinar los factores que tienen influencia en los casos de deserción (retiros y pérdida de la asignatura por asistencia), por carrera, niveles, asignaturas, áreas de conocimiento y facultad.	Facultad	Cumple	3
	EIS	Cumple	3
	EIS-Niveles (10)	Cumple	30
	EIETR	Cumple	3
	EIETR-Niveles(10)	Cumple	30
	EIECRI	Cumple	3
	EIECRI-Niveles(10)	Cumple	30
	EDG	Cumple	3
	EDG-Niveles(10)	Cumple	30
	EIS ext. Macas	Cumple	3
	EIS ext. Macas – Niveles (10)	No cumple	0
REQ 5: Determinar patrones de comportamiento en la promoción académica de los estudiantes; por asignatura, nivel, áreas de conocimiento, carrera y facultad.	Facultad	Cumple	3
	EIS	Cumple	3
	EIS-Niveles (10)	Cumple	30
	EIETR	Cumple	3

	EIETR-Niveles(10)	Cumple	30
	EIECRI	Cumple	3
	EIECRI-Niveles(10)	Cumple	30
	EDG	Cumple	3
	EDG-Niveles(10)	Cumple	30
	EIS ext. Macas	Cumple	3
	EIS ext. Macas – Niveles (10)	No cumple	0
REQ 6: Determinar los factores que influyen en los escenarios de segunda y tercera matrícula; por asignatura, nivel, áreas de conocimiento, carrera y facultad.	Facultad	Cumple	3
	EIS	Cumple	3
	EIS-Niveles (10)	Cumple	30
	EIETR	Cumple	3
	EIETR-Niveles(10)	Cumple	30
	EIECRI	Cumple	3
	EIECRI-Niveles(10)	Cumple	30
	EDG	Cumple	3
	EDG-Niveles(10)	Cumple	30
	EIS ext. Macas	Cumple	3
	EIS ext. Macas –	No cumple	0

	Niveles (10)		
REQ 7: Determinar la proyección de graduados; por carrera y facultad.	Facultad	Cumple	1
	EIS	Cumple	1
	EIETR	Cumple	1
	EIECRI	Cumple	1
	EDG	Cumple	1
	EIS ext. Macas	No cumple	0
REQ 8: Determinar los factores que inciden en los casos de estudiantes con baja eficiencia terminal; por carrera y facultad.	Facultad	Cumple	1
	EIS	Cumple	1
	EIETR	Cumple	1
	EIECRI	Cumple	1
	EDG	Cumple	1
	EIS ext. Macas	No cumple	0
Total de modelos			533

Autora: Gallegos, K., 2015

Del total de 533 modelos a ser desarrollados, 5 serán sometidos al algoritmo de serie temporal y son los correspondientes al requisito 7 debido a que se realizará una predicción en base al tiempo; el resto serán modelados con el algoritmo de mejor desempeño.

4.3. PREPARACIÓN DE LOS DATOS

La preparación de los datos permite realizar una limpieza de los valores de las tablas seleccionadas en el análisis de los datos. Esta fase corresponde a un proceso de calidad de datos que permite avanzar a las siguientes fases del proceso de minería.

En algunos de los casos, los datos están respaldados por integridad referencial, en otros por la obligatoriedad de presencia de clave primaria, en el resto de casos se ha procedido a utilizar la herramienta Data Quality Services de Microsoft. Los resultados se describen a continuación:

El atributo “Nacionalidad” sin integridad referencial presente en las tablas, ha sido corregido basado en los resultados de la búsqueda de valores con la herramienta de limpieza, en donde además se detectó que un 20% de los valores son incorrectos.

Los valores nulos dentro de los atributos de tipo numérico no han sido corregidos debido a que dentro de la herramienta para minería de datos se identifican como valores ausentes y para no afectar los resultados no se han definido valores nuevos.

Los datos de la limpieza realizada se muestran en el Anexo 2: Informe de Análisis y Preparación de Datos.

4.4. MODELADO

Dentro del modelado se examinan los algoritmos disponibles para la tarea de minería de datos, y así definir los más adecuados al problema planteado para ser examinados. El análisis realizado se resume en la Tabla 25: Selección de Algoritmos de Minería.

TABLA 25: SELECCIÓN DE ALGORITMOS DE MINERÍA

Característica	**Valor esperado**	**Árbol de decisión**	**Bayes Naive**	**Clústeres**	**Red Neuronal**	**Regresión Logística**	**Regresión lineal**
Predecir atributos discretos	Sí	Sí	Sí	Sí	Sí	Sí	No
Predecir atributos continuos	Sí	Sí	No	Sí	Sí	Sí	Sí
Aceptar valores discretos	Sí	Sí	Sí	Sí	Sí	Sí	No

Aceptar valores continuos	Sí	Sí	No	Sí	Sí	Sí	Sí
Tarea de predicción	Sí	Sí	Sí	No	Sí	Sí	Sí
Tarea de agrupación	Sí	Sí	Sí	Sí	No	Sí	Sí

Fuente: msdn.microsoft.com

En este caso el algoritmo de Regresión Logística, el algoritmo de Árbol de Decisión y el algoritmo de Serie Temporal de Microsoft son los seleccionados para el proceso de evaluación sobre los datos del presente proyecto, debido a que los resultados de los análisis individuales se asemejan a los resultados esperados para los procesos de minería.

4.5. EVALUACIÓN

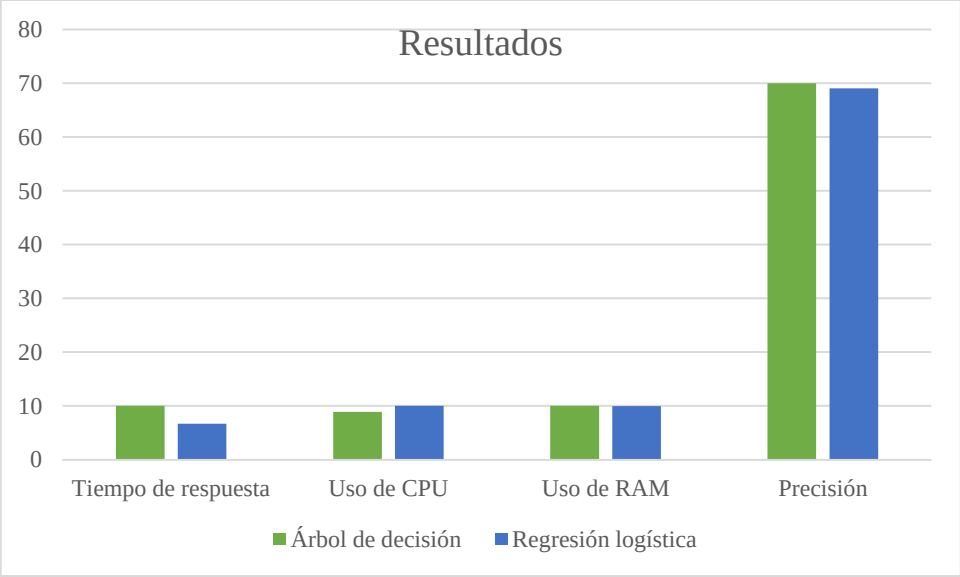
El proceso de evaluación somete a los algoritmos seleccionados a la data del proyecto de minería para determinar el más adecuado para la implementación final. En este proyecto se buscó el algoritmo de mejor desempeño entre los dos seleccionados para los requisitos planteados. Los indicadores a comparar se listan a continuación:

- Tiempo de respuesta.- Un menor tiempo de respuesta de un algoritmo frente a la misma estructura de datos favorece al mejor desempeño del mismo a nivel general.
- Uso de CPU.- El menor porcentaje de uso de CPU al modelar una estructura de datos con determinado algoritmo establece un mejor desempeño frente al otro algoritmo de estudio.

- Uso de RAM.- Menor cantidad de uso de memoria RAM permite decidir parcialmente el algoritmo de mejor desempeño frente a una misma estructura de datos.
- Precisión.- La precisión es un valor arrojado por la herramienta de minería de datos y que permite decidir que algoritmo tiene un mejor desempeño frente a una misma estructura de datos.

Bajo estas premisas se realizó la toma de medidas y la aplicación de técnicas de estadística para obtener el algoritmo de mejor desempeño entre el de Árbol de Decisión y Regresión Logística. Los resultados se muestran a continuación en Figura N°21 Resultados de evaluación de algoritmos.

FIGURA N° 21: RESULTADOS DE EVALUACIÓN DE ALGORITMOS



Autora: Gallegos, K., 2015

Los resultados arrojan que el algoritmo de Árbol de Decisión tiene un mejor desempeño que el de Regresión Logística con lo que se concluye el proceso comparativo y se procede a la implementación de la minería de datos. Más detalles sobre la evaluación de los

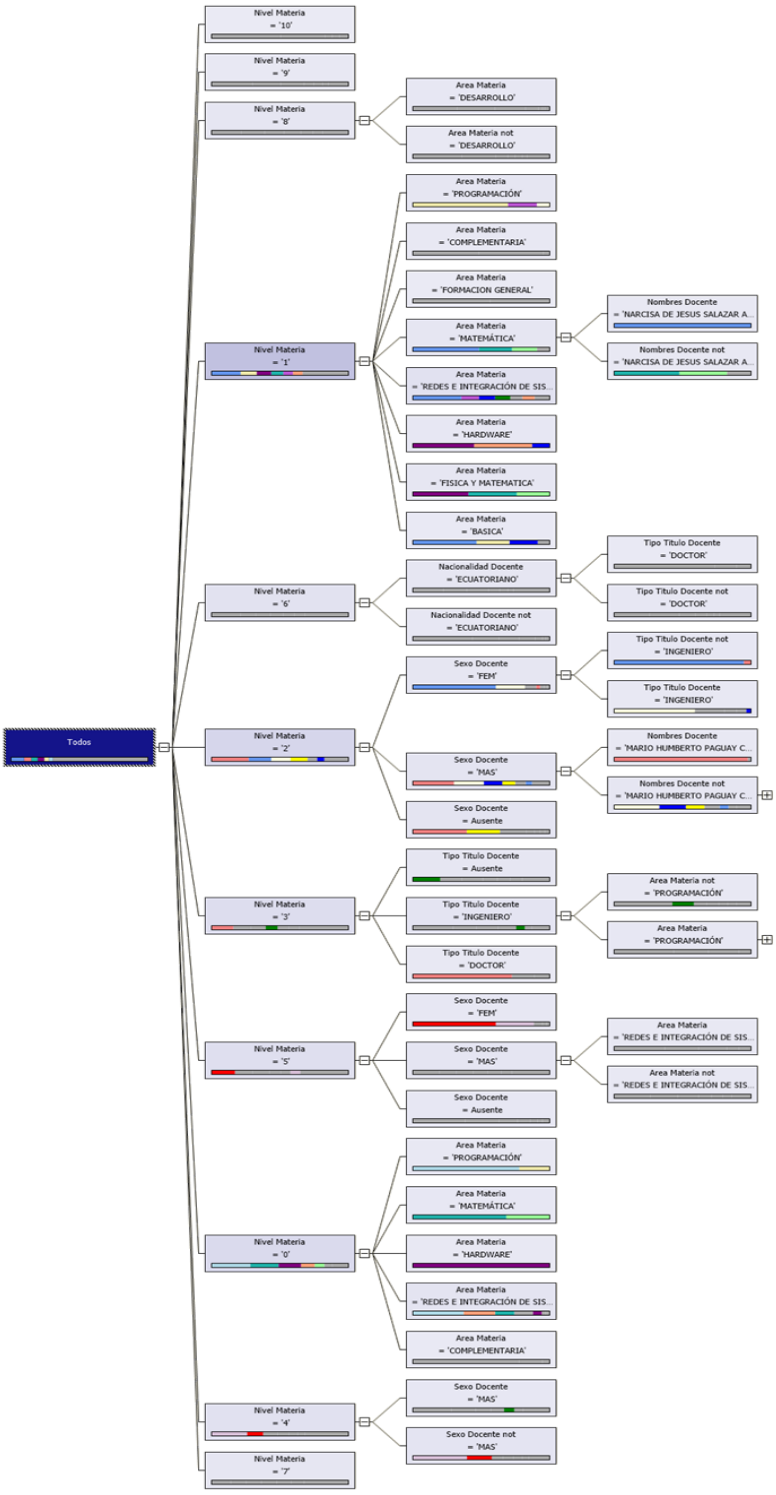
algoritmos se encuentran con en el CAPÍTULO III COMPARACIÓN DE ALGORITMOS ÁRBOL DE DECISIÓN Y REGRESIÓN LOGÍSTICA de este mismo documento.

4.6. IMPLEMENTACIÓN

La fase de implementación es la última de las seis fases planteadas por la metodología CRISP-DM. Finalmente se obtienen los modelos de minería de datos que satisfacen los requerimientos del proyecto. A continuación se presenta un ejemplo de los modelos obtenidos.

REQ 6: Determinar los factores que influyen en los escenarios de segunda y tercera matrícula; por asignatura, nivel, áreas de conocimiento, carrera y facultad.

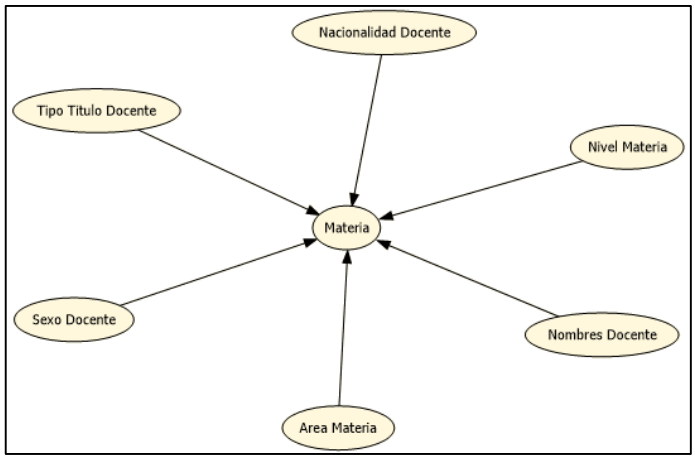
FIGURA N° 22: ÁRBOL DE DECISIÓN



Autora: Gallegos, K., 2015

La Figura N° 22: Árbol de Decisión muestra los resultados de minería sobre los datos de la Escuela de Ingeniería en Sistemas concerniente al factor de deserción de los estudiantes. Del total de deserciones, las materias con mayores registros son Física, Matemática y Fundamentos de Programación. El árbol de dependencias de la Figura N° 23: Red de Dependencias indica que los aspectos que influyen sobre la deserción en la Escuela de Ingeniería en Sistemas son: nivel de la materia, área de la materia, sexo del docente, tipo de título del docente, nacionalidad del docente y nombre del docente, en ese orden de mayor a menor.

FIGURA N° 23: RED DE DEPENDENCIAS



Autora: Gallegos, K., 2015

Como parte de la fase de implementación se sugiere también un plan de mantenimiento para el proyecto de minería que se muestra en la Tabla 26: Plan de mantenimiento preventivo y la Tabla 27: Plan de mantenimiento correctivo a continuación.

TABLA 26: PLAN DE MANTENIMIENTO PREVENTIVO

Plan de mantenimiento	
Caso	Incremento de filas en las tablas relacionadas a las estructuras de minería
Frecuencia	Semestral
Actividades	Validar nuevas entradas que afecten la limpieza de datos.

	De encontrar nuevos datos incorrectamente ingresados proceder a la limpieza y actualizar las estructuras de datos. Generar los modelos de minería

Autora: Gallegos, K., 2015

TABLA 27: PLAN DE MANTENIMIENTO CORRECTIVO

Plan de mantenimiento	
Caso	Cambios en la definición de las tablas asociadas a los datos de los procesos académicos de la FIE
Frecuencia	Semestral
Actividades	Identificar los atributos o relaciones añadidas o eliminadas. Identificar si los elementos eliminados son parte de las actuales estructuras de minería de datos Realizar un análisis sobre la afectación de los elementos eliminados y añadidos en la definición de las tablas asociadas Decidir si mantener, modificar o eliminar la estructura de datos. Implementar los nuevos modelos en los casos de no eliminar la estructura de datos.

Autora: Gallegos, K., 2015

El informe sobre la implementación de la minería de datos se encuentra en el ANEXO 3: INFORME DE IMPLEMENTACIÓN DE MINERÍA DE DATOS.

CONCLUSIONES

Los datos con los que se cuentan para realizar un análisis de minería de datos influyen directamente en la elección de los algoritmos aplicables al modelado de datos.

Los algoritmos que no aceptan datos discretos o continuos presentes en un análisis, pueden no considerar a datos relevantes del modelado de minería.

Los algoritmos de regresión logística y de árbol de decisión de Microsoft son adecuados para el análisis de los datos discretos y continuos, que se tuvieron en cuenta para el análisis de la información académica de la FIE-ESPOCH.

En un proyecto de minería de datos se debe tomar en cuenta el análisis previo para evitar el uso de datos sin integridad.

El análisis estadístico y las pruebas realizadas entre el algoritmo de Regresión Logística y el de Árbol de Decisión revela que bajo los indicadores de: uso de CPU, uso de RAM, tiempo de respuesta y precisión, el algoritmo de Árbol de Decisión tiene un mejor desempeño al de Regresión Logística.

RECOMENDACIONES

El uso de la metodología CRISP-DM para minería de datos guía los procesos a llevar a cabo a través de fases y tareas; se recomienda su uso, sobre todo bajo situación de falta de experticia en el desarrollo de proyectos de minería de datos.

Definir los algoritmos aplicables a un caso de estudio dentro de la minería de datos depende directamente de los datos y los tipos de datos con los que se cuenta para el estudio, omitir el análisis previo implicaría la obtención de modelos poco precisos o no apegados a la realidad, por lo que se recomienda realizar un estudio inicial de datos previo a la elección de los algoritmos de minería.

Contar con los datos socioeconómicos de los estudiantes de la FIE aportará con nuevos patrones dentro de la extracción de conocimiento, por lo que se recomienda desarrollar un plan para integrar dicha información al presente proyecto de minería de datos.

Se recomienda realizar estudios sobre la aplicabilidad de otras herramientas orientadas a la minería de datos y experimentar con la variabilidad en los resultados obtenidos en este proyecto.

RESUMEN

Estudio comparativo de algoritmos de minería de datos para determinar el algoritmo de mejor desempeño aplicado al área académica de la Facultad de Informática y Electrónica de la Escuela Superior Politécnica de Chimborazo.

Para comprobar la hipótesis sobre el desempeño de los algoritmos seleccionados, se utilizó el método científico y los siguientes elementos: un computador portátil y herramientas software: Microsoft SQL Server, Microsoft Data Quality, Microsoft Data Tools y Microsoft Analisys Services, y la técnica de observación para realizar las mediciones. La prueba z para dos colas y un nivel de significancia del 5%, fue aplicado sobre el uso de la Unidad Central de Proceso (CPU), uso de la Memoria de Acceso Aleatorio (RAM), tiempo de respuesta y precisión de los algoritmos. El resultado del análisis arrojó que el algoritmo de Árbol de Decisión tiene mejor desempeño que el algoritmo de Regresión Logística con 98,92% sobre 95,70% de desempeño.

Bajo la guía de la metodología Cross Industry Standard Process for Data Mining, CRISP-DM se implementó el proyecto de minería de datos sobre la información de indicadores académicos de la Facultad de Informática y Electrónica, haciendo uso del algoritmo Árbol de Decisión de Microsoft, para obtener los patrones de comportamiento requeridos.

Se concluye que el algoritmo de Árbol de Decisión tiene mejor desempeño que el Algoritmo de Regresión Logística sobre los datos académicos de la Facultad de Informática y Electrónica.

Se recomienda realizar un estudio inicial de datos previo a la elección de los algoritmos para evitar el descarte de tipos de datos existentes en un proyecto de minería de datos.

Palabras clave: <ANÁLISIS>, <MINERÍA DE DATOS>, <ALGORITMOS>, <REGRESIÓN LOGÍSTICA>, <ÁRBOL DE DECISIÓN>, <PREDICCIÓN>, <PATRONES>, < FACULTAD DE INFORMÁTICA Y ELECTRÓNICA>

SUMMARY

Comparative Study of data mining algorithms to determine the best algorithm performance applied to the academic area of the Faculty of computer science and electronics of Escuela Superior Politecnica de Chimborazo.

To validate the hypothesis on the performance of the chosen algorithms, the scientific method and the following elements such as: a portable computer and software tools were used as well as Microsoft SQL Server, Microsoft Data Quality, Microsoft Data Tools, Microsoft Analysis Services, and the observation technique to carry out the measurements. The z-test for two-tailed, and a significance level of 5% were applied on the use of a Central Processing Unit (CPU), employ of the Random Access Memory (RAM), response time and accuracy of the algorithms. The result of the analysis showed that the decision tree algorithm has better performance than the logistic regression algorithm with 98.92% on 95.70% of performance.

Under the guidance of the methodology Cross Industry Standard Process for Data Mining CRISPDM the project of data mining was implemented on the information of academic indicators of the Faculty of computer science and electronics, making use of the decision tree algorithm of Microsoft, to get the patterns of behavior required.

It is concluded that the decision tree algorithm had better performance than the logistic regression algorithm on the academic data of the Faculty of computer science and electronics.

An initial study of information prior to the choice of algorithms is recommended to avoid the discarding of data types in a data mining project.

Key Words: <ANALYSIS>, <MINING DATA>, <ALGORITHMS>, <LOGISTIC REGRESSION>, <DECISION TREE>, <PREDICTION>, <PATTERNS>, <FACULTY OF COMPUTER SCIENCE AND ELECTRONICS>

GLOSARIO

Base de datos: Es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. En este sentido, una biblioteca puede considerarse una base de datos compuesta en su mayoría por documentos y textos impresos en papel e indexados para su consulta

Bases de datos multidimensionales: Se utilizan principalmente para crear aplicaciones OLAP y pueden verse como bases de datos de una sola tabla, su peculiaridad es que por cada dimensión tienen un campo (o columna), y otro campo por cada métrica o hecho, es decir estas tablas almacenan registros cuyos campos son de la forma:

Data Mining: Es el proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y la transformándola en una estructura comprensible para su uso posterior.

Fase: Un terminó para una parte de alto nivel del modelo de proceso CRISP-DM; consiste en un conjunto de fases relacionadas.

Información: Es un conjunto de datos con un significado, o sea, que reduce la incertidumbre o que aumenta el conocimiento de algo. En verdad, la información es un mensaje con significado en un determinado contexto, disponible para uso inmediato y que proporciona orientación a las acciones por el hecho de reducir el margen de incertidumbre con respecto a nuestras decisiones

Microsoft SQL Server: Es un sistema para la gestión de bases de datos producido por Microsoft basado en el modelo relacional. Microsoft SQL Server constituye la alternativa de Microsoft a otros potentes sistemas gestores de bases de datos como son Oracle, PostgreSQL o MySQL.

Modelo de proceso: Define la estructura de proyectos de minería de datos y provee una guía para su ejecución; consiste de un modelo de referencia y una guía de usuario

Modelo de referencia: Descomposición de un proyecto de minería en fases, tareas y salidas.

Modelo: En minería de datos un modelo es un ejecutable aplicable a un conjunto de datos para predecir atributos.

OLAP: Es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia empresarial (o Business Intelligence), minería cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ello utiliza estructuras multidimensionales, que contienen datos resumidos de grandes Bases de datos o Sistemas Transaccionales

OLTP: (On-Line Transactional Processing). Se usa en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares.

Predicción: Puede referirse tanto a la «acción y al efecto de predecir, como a «as palabras que manifiestan aquello que se predice; en este sentido, predecir algo es anunciar por revelación, ciencia o conjetura algo que ha de sucede

Salida: Dentro de CRISP-DM se describe una salida como el resultado tangible de la realización de una tarea

SQL: (Por sus siglas en inglés Structured Query Language) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en estas. Una de sus características es el manejo del álgebra y el cálculo relacional permitiendo efectuar consultas con el fin de recuperar de una forma sencilla información de interés de una base de datos, así como también hacer cambios sobre ella.

Tarea genérica: Una tarea que toma lugar a través de los diferentes proyectos de minería de datos.

Tarea: Una serie de actividades para producir una o más salidas; parte de una fase.

BIBLIOGRAFÍA

DATA MINING: WHAT IS DATA MINING?. UCLA Anderson.
http://www.anderson.ucla.edu.,
2013-11-6

BERSON, Alex. Data Mining. *Building Data Mining Applications for CRM.* s.l. : McGraw-Hill Companies, 2000.

CHAPMAN, Pete and et.al. *CRISP-DM 1.0.* Washington D. C., EEUU : SPSS, 2000, 76p.

CORTEZ, Paulo and SILVA, Alice. *Using Data Minig to predict secondary school student performance.* Guimaraes, Portugal, University of Minho, 2006, 8p.

FERNÁNDEZ, Santiago. *Regresión Logístic*, Madrid, España. Universidad Autónoma de Madrid, 2011, 76p.

JING, Luan. *Aplicaciones de minería de datos en la eduación superior. New York, EEUU*. IBM Software Business Analytics, 2010, 8p.

FAYYAD, Usama, et. al. *Knowledge Discovery In Databases.* New York, EEUU, Amgerican Association for Artificial Intelligence, 1996. pp. 37-54.

KUMAR, Brijesh and SAURABH, Pal. *Mining Educational Data to analyze student´s performance.* Rajasthan, India, IJACSA, 2011. pp. 63-69.

MACLENNAN, Jamie. *Data Mining with Microsoft SQL Server 2008.* Indianapolis, EEUU, Wiley Publishing Inc. 2008. pp. 39-53.

ALGORITMO DE ÁRBOLES DE DECISIÓN. Microsoft Developer Network, 2012. https://msdn.microsoft.com/es-ec/library/ms175312.aspx
2013-11-06

ALGORITMO DE SERIE TEMPORAL DE MICROSOFT. Microsoft Developer Network. https://msdn.microsoft.com/es-es/library/ms174923.aspx.

2013-11-06

ALGORITMO REGRESIÓN LOGÍSTICA DE MICROSOFT. Microsoft Developer Network. http://msdn.microsoft.com/es-es/library/ms174806.aspx.
2013-11-06

ALGORITMOS DE MINERÍA DE DATOS. Microsoft Developer Network. http://msdn.microsoft.com.
2013-11-06

AN INTRODUCTION TO DATA MINING.
www.thearling.com.
2013-11-05

VALLEJOS, Sofia. *Minería de Datos.* Corrientes, Argentina, Universidad Nacional de Noreste, 2006, pp. 11-16.

WEISS, Sholom. *Predictive Data Mining. A practical guide.* San Francisco, EEUU, Morgan Kaufmann, 1997, pp. 1-24.

ANEXOS

ANEXO 1: INFORME DEL ANÁLISIS DEL NEGOCIO

Proyecto: Minería de datos aplicado al área académica de la FIE-ESPOCH

Director: Ing. Iván Menes

Analista: Katherine Gallegos

INFORME

El análisis del entorno en donde se aplicará el proyecto, representa la obtención del objetivo del negocio al aplicar minería de datos sobre los repositorios de información. Una vez señalado el propósito del cliente, se lista los recursos con lo se cuenta para el desarrollo del proyecto, así como los riesgos podría afectar el desarrollo del proyecto.

Objetivos

Esta sección está enfocada a mostrar los objetivos del presente informe.

Objetivo General

- Generar un informe sobre el estudio previo realizado en la FIE-ESPOCH, institución en la que se llevará a cabo el proyecto de minería de datos.

Objetivos Específicos

- Describir los objetivos de minería de datos.
- Listar los recursos disponibles para el desarrollo del proyecto.
- Identificar riesgos del proyecto de minería de datos.

Descripción del negocio

Institución: Facultad de Informática y Electrónica de la Escuela Superior Politécnica de Chimborazo.

Objetivos de minería

El principal objetivo del negocio para aplicar minería de datos es analizar si existen características que intervienen en los resultados académicos de los estudiantes de la Facultad de Informática y Electrónica; en los procesos de ingreso, matriculación, promoción y graduación. Por lo tanto los objetivos del negocio se describen como sigue:

- Determinar patrones de comportamiento para el ingreso de los estudiantes; por carrera y facultad.
- Determinar patrones de comportamiento para el ingreso directo a la carrera (sin pasar por el curso de nivelación); por carrera y facultad.
- Determinar patrones de comportamiento para la matriculación y selección de asignatura de los estudiantes (número de asignaturas, número de créditos, nivel y área de las asignaturas, etc.); por carrera, por nivel y por facultad.
- Determinar los factores que tienen influencia en los casos de deserción (retiros y pérdida de la asignatura por asistencia), por carrera, niveles, asignaturas, áreas de conocimiento y facultad.
- Determinar patrones de comportamiento en la promoción académica de los estudiantes; por asignatura, nivel, áreas de conocimiento, carrera y facultad.
- Determinar los factores que influyen en los escenarios de segunda y tercera matrícula; por asignatura, nivel, áreas de conocimiento, carrera y facultad.

- Determinar la proyección de graduados; por carrera y facultad.
- Determinar los factores que inciden en los casos de estudiantes con baja eficiencia terminal; por carrera y facultad.

Criterios de aceptación

Para determinar el éxito de los resultados obtenidos después de aplicar el análisis de minería de datos se establece que:

- Los porcentajes de error admitidos para los resultados serán menores o iguales al 25%
- Se espera, al menos cinco factores que influyan en el desempeño estudiantil como resultado del proceso de análisis.

Estos son los factores que determinarán la validez de los resultados.

Evaluación de situación actual

Los recursos disponibles para el presente proyecto se listan a continuación:

- La Facultad de Informática y Electrónica cuenta con sus datos dentro del Sistema Académico de Información OASIS, este repositorio será sobre el cual se aplicará el análisis de minería de datos.
- Se cuenta con un servidor Servidor HP ProLiant DL360 G7 para alojar el motor de base de datos.

De entre los supuestos planteados para este proyecto se listan:

- Los datos almacenados en la base de datos han pasado por un proceso de verificación de la calidad.

- Se cuenta con los datos históricos de al menos cinco años atrás.

Riesgos del proyecto

Se han identificado los siguientes riesgos en el marco del desarrollo del presente proyecto de minería de datos. Ver la tabla a continuación.

Tabla: Riesgos del proyecto

Código	**Riesgo**	**Probabilidad**	**Impacto**	**Categoría**
MDR01	El usuario decide no continuar con el proyecto de minería de datos	Bajo	Alto	Alto
MDR04	Los resultados del análisis de minería de datos no cumplen con los objetivos del negocio	Medio	Alto	Alto
MDR05	No se cuenta con acceso a los recursos de datos para el análisis	Medio	Alto	Alto
MDR02	Los algoritmos no proveen el porcentaje de precisión requerido por el usuario	Bajo	Medio	Medio
MDR03	No se cuenta con la infraestructura que soporte el software de minería de datos	Medio	Medio	Medio
MDR06	Los datos de origen no son actualizados	Medio	Bajo	Medio

MDR07	Los datos de origen no han pasado por un proceso de calidad	Alto	Bajo	Medio
MDR08	Los atributos de los datos de origen no son suficientes para aplicar un algoritmo de minería de datos	Bajo	Alto	Medio
MDR09	No se cuenta con conocimientos para resolver problemas dentro del desarrollo de minería de datos	Bajo	Bajo	Bajo

Mitigación de riesgos

Los riegos con criticidad media o alta han sido analizados para obtener la mitigación que se debería aplicar en caso de que se conviertan en un problema para el proyecto. Ver tabla a continuación.

Tabla: Mitigación de riesgos

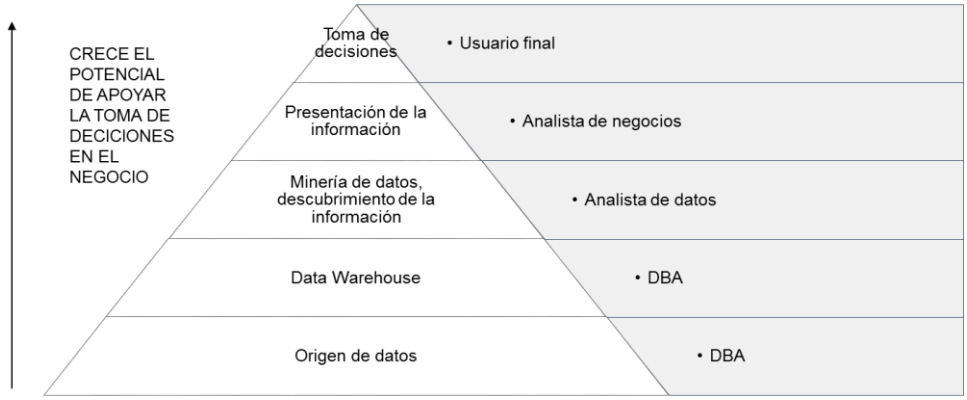
Código	**Riesgo**	**Mitigación**
MDR01	El usuario decide no continuar con el proyecto de minería de datos	Exponer nuevamente los beneficios del proyecto de minería de datos.
MDR04	Los resultados del análisis de minería de datos no cumplen	Analizar nuevamente los requerimientos del usuario final, y si es necesario replantear los

	con los objetivos del negocio	objetivos del negocio.
MDR05	No se cuenta con acceso a los recursos de datos para el análisis	Exponer la necesidad de acceder a los datos marcados como recursos disponibles para el avance del proyecto de minería de datos
MDR02	Los algoritmos no proveen el porcentaje de precisión requerido por el usuario	Reconfigurar los parámetros de entrada de los algoritmos
MDR03	No se cuenta con la infraestructura que soporte el software de minería de datos	Instalar servidores virtuales en un servidor provisional.
MDR06	Los datos de origen no son actualizados	Trabajar con los datos proporcionados, debido a que es posible mantener el modelado resultante con los datos no actualizados.
MDR07	Los datos de origen no han pasado por un proceso de calidad	La metodología de minería de datos destina una parte del ciclo del proyecto a la limpieza de los datos, recurrir a las herramientas oportunas para el proceso de calidad necesaria previo al modelado.
MDR08	Los atributos de los datos de origen no son suficientes para aplicar un algoritmo de minería de datos	Creación de atributos derivados de los datos, que representen un cambio de forma, mas no de fondo del origen proporcionado.

Beneficios del proyecto

El nombre de Data Mining deriva de las similitudes entre buscar valiosa información de negocios en grandes bases de datos. Estos procesos requieren examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente donde residen los valores. Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades de predicción automatizada de tendencias y comportamientos. Las técnicas de Data Mining pueden redituar los beneficios de automatización en las plataformas de hardware y software existentes y puede ser implementada en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados.

Figura: Potencial de apoyo para la toma de decisiones



En la figura anterior, se presenta a la minería de datos como un paso en el soporte las decisiones dentro del negocio, puesto que se obtiene acceso a un conjunto de información que se encuentra "escondida" dentro de los repositorios de datos. Así este proyecto, beneficiará a los actores del negocio al tomar en cuenta en los procesos de decisión aquella información de patrones y comportamientos que se descubren con un análisis de minería de datos.

Plan del proyecto

Para alcanzar los objetivos planteados, se presenta la organización y distribución del tiempo y recursos en la tabla a continuación:

Tabla: Planificación

Actividad	**Tiempo**	**Fecha inicio**	**Fecha fin**	**Recursos**
Análisis de los datos	10 días	06/05/2014	19/05/2014	Computador portátil Base de datos Oasis-ESPOCH
Preparación de los datos	10 días	20/05/2014	02/06/2014	Computador portátil Base de datos Oasis - ESPOCH Herramienta para calidad de datos
Modelado	15 días	03/06/2014	23/06/2014	Computador portátil Base de datos Oasis - ESPOCH SQL Analysis Services
Evaluación	10 días	24/06/2014	07/07/2014	Computador portátil Base de datos Oasis - ESPOCH

				SQL Analysis Services
Implementación	10 días	08/07/2014	21/07/2014	Servidor HP Proliant Base de datos Oasis - ESPOCH SQL Analysis Services

Donde los tiempos estimados se establecen sin contar tomar en cuenta la posibilidad de retornar al análisis para contemplar soluciones a los problemas encontrados en las fases de modelado o evaluación.

ANEXO 2: INFORME DEL ANÁLISIS Y PREPARACIÓN DE LOS DATOS

Proyecto: Minería de datos aplicado al área académica de la FIE-ESPOCH

Director: Ing. Iván Menes

Analista: Katherine Gallegos

INFORME

El análisis de los datos de origen para el proceso de minería, constituye un paso importante dentro del ciclo del proyecto, en donde se identifican los atributos y tablas de interés para la estructura final sobre la cual se obtendrán los resultados. De este análisis también se reflejan los problemas detectados dentro de la data, y constituyen el inicio de una segunda etapa denominada preparación de los datos, cuyo resultado es un dataset refinado para las siguientes fases del proyecto.

Análisis de datos

Los datos disponibles son los correspondientes a las bases de datos de sistema informático OASIS, las particiones presentes son:

- La partición DBO que contienen los datos de la tablas necesarias

Para este proyecto se hará uso de las tablas de la partición DBO, estas tablas que se han seleccionado son las que contienen los datos del estudiante y los resultados de su proceso académico dentro de la Facultad de Informática y Electrónica. Las tablas consideradas se listan a continuación:

- Areas
- Carreras
- Ciudades
- Egresados

- Escuelas
- Estudiantes
- Facultades
- Graduados
- Inscripciones
- Materias
- MateriasAprobadas
- MateriasAsignadas
- Matriculas
- Provincias
- Tesis

Esta parte del análisis incluye la verificación de que exista la suficiente cantidad de datos dentro de las tablas relacionadas con los requisitos de la minería para poder realizar el modelado. De no existir los datos suficientes se procederá a descartar el requisito, siendo que el número mínimo de registros requerido será de doscientos y al menos cinco atributos de uso potencial para el modelado (datos que no correspondan a claves autogeneradas y fechas). En la tabla a continuación, se muestra los resultados del estudio sobre los datos de interés para el proyecto de minería según los objetivos planteados.

Tabla: Análisis de datos

Requisito	**Número de registros**	**Atributos**	**Resultado**	**Modelos**

REQ 1 (Facultad)	Cumple	Cumple	Cumple	1
REQ 1 (Carrera IS)	Cumple	Cumple	Cumple	1
REQ 1 (Carrera IECRI)	Cumple	Cumple	Cumple	1
REQ 1 (Carrera IETR)	Cumple	Cumple	Cumple	1
REQ 1 (Carrera IDG)	Cumple	Cumple	Cumple	1
REQ 1 (Carrera IS ext MA)	No cumple	Cumple	No cumple	0
REQ 2 (Facultad)	No cumple	No cumple	No cumple	0
REQ 2 (Carrera IS)	No cumple	No cumple	No cumple	0
REQ 2 (Carrera IECRI)	No cumple	No cumple	No cumple	0
REQ 2 (Carrera IETR)	No cumple	No cumple	No cumple	0
REQ 2 (Carrera IDG)	No cumple	No cumple	No cumple	0

REQ 2 (Carrera IS ext MA)	No cumple	No cumple	No cumple	0
REQ 3 (Facultad)	Cumple	Cumple	Cumple	4
REQ 3 (Carrera IS)	Cumple	Cumple	Cumple	4
REQ 3 (Carrera IS Niveles)	Cumple	Cumple	Cumple	20
REQ 3 (Carrera IECRI)	Cumple	Cumple	Cumple	4
REQ 3 (Carrera IECRI Niveles)	Cumple	Cumple	Cumple	20
REQ 3 (Carrera IETR)	Cumple	Cumple	Cumple	4
REQ 3 (Carrera IETR Niveles)	Cumple	Cumple	Cumple	20
REQ 3 (Carrera IDG)	Cumple	Cumple	Cumple	4
REQ 3 (Carrera IDG Niveles)	Cumple	Cumple	Cumple	20
REQ 3 (Carrera IS ext. MA)	Cumple	Cumple	Cumple	4
REQ 3 (Carrera IS ext. MA Niveles)	No cumple	Cumple	Cumple	0
REQ 4 (Facultad)	Cumple	Cumple	Cumple	3

REQ 4 (Carrera IS)	Cumple	Cumple	Cumple	3
REQ 4 (Carrera IS Niveles)	Cumple	Cumple	Cumple	30
REQ 4 (Carrera IECRI)	Cumple	Cumple	Cumple	3
REQ 4 (Carrera IECRI Niveles)	Cumple	Cumple	Cumple	30
REQ 4 (Carrera IETR)	Cumple	Cumple	Cumple	3
REQ 4 (Carrera IETR Niveles)	Cumple	Cumple	Cumple	30
REQ 4 (Carrera IDG)	Cumple	Cumple	Cumple	3
REQ 4 (Carrera IDG Niveles)	Cumple	Cumple	Cumple	30
REQ 4 (Carrera IS ext. MA)	Cumple	Cumple	Cumple	3
REQ 4 (Carrera IS ext. MA Niveles)	No cumple	Cumple	Cumple	0
REQ 5 (Facultad)	Cumple	Cumple	Cumple	3
REQ 5 (Carrera IS)	Cumple	Cumple	Cumple	3
REQ 5 (Carrera IS Niveles)	Cumple	Cumple	Cumple	30

REQ 5 (Carrera IECRI)	Cumple	Cumple	Cumple	3
REQ 5 (Carrera IECRI Niveles)	Cumple	Cumple	Cumple	30
REQ 5 (Carrera IETR)	Cumple	Cumple	Cumple	3
REQ 5 (Carrera IETR Niveles)	Cumple	Cumple	Cumple	30
REQ 5 (Carrera IDG)	Cumple	Cumple	Cumple	3
REQ 5 (Carrera IDG Niveles)	Cumple	Cumple	Cumple	30
REQ 5 (Carrera IS ext. MA)	Cumple	Cumple	Cumple	3
REQ 5 (Carrera IS ext. MA Niveles)	No cumple	Cumple	Cumple	0
REQ 6 (Facultad)	Cumple	Cumple	Cumple	3
REQ 6 (Carrera IS)	Cumple	Cumple	Cumple	3
REQ 6 (Carrera IS Niveles)	Cumple	Cumple	Cumple	30
REQ 6 (Carrera IECRI)	Cumple	Cumple	Cumple	3
REQ 6 (Carrera IECRI Niveles)	Cumple	Cumple	Cumple	30

REQ 6 (Carrera IETR)	Cumple	Cumple	Cumple	3
REQ 6 (Carrera IETR Niveles)	Cumple	Cumple	Cumple	30
REQ 6 (Carrera IDG)	Cumple	Cumple	Cumple	3
REQ 6 (Carrera IDG Niveles)	Cumple	Cumple	Cumple	30
REQ 6 (Carrera IS ext. MA)	Cumple	Cumple	Cumple	3
REQ 6 (Carrera IS ext. MA Niveles)	No cumple	Cumple	Cumple	0
REQ 8 (Facultad)	Cumple	Cumple	Cumple	1
REQ 8 (Carrera IS)	Cumple	Cumple	Cumple	1
REQ 8 (Carrera IECRI)	Cumple	Cumple	Cumple	1
REQ 8 (Carrera IETR)	Cumple	Cumple	Cumple	1
REQ 8 (Carrera IDG)	Cumple	Cumple	Cumple	1
REQ 8 (Carrera IS ext MA)	No cumple	Cumple	No cumple	0
TOTAL DE MODELOS				528

Al tratarse de una predicción en el tiempo, como lo exige el requerimiento número siete del presente proyecto de minería de datos, el análisis que corresponde es diferente, para este caso es necesario sólo contar con el número de datos necesarios que varían en el tiempo. En la tabla a continuación se muestran los resultados.

Tabla: Análisis de datos II

Requisito	**Número de registros**	**Resultado**	**Modelos**
REQ 7 (Facultad)	Cumple	Cumple	1
REQ 7 (Carrera IS)	Cumple	Cumple	1
REQ 7 (Carrera IECRI)	Cumple	Cumple	1
REQ 7 (Carrera IETR)	Cumple	Cumple	1
REQ 7 (Carrera IDG)	Cumple	Cumple	1
REQ 7 (Carrera IS ext MA)	No cumple	No cumple	0
TOTAL DE MODELOS			5

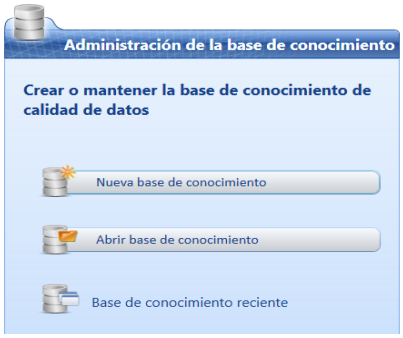
En total serán desarrollados 533 modelos de minería de datos, de los cuales 528 corresponden a la población a ser examinada entre dos algoritmos de predicción, mientras que los 5 restantes que corresponden a los modelos del requerimiento 7 de minería de

datos, serán sometidos a un solo algoritmo que será el algoritmo de serie temporal por tratarse de una predicción en base al tiempo.

Base de conocimiento para calidad de datos

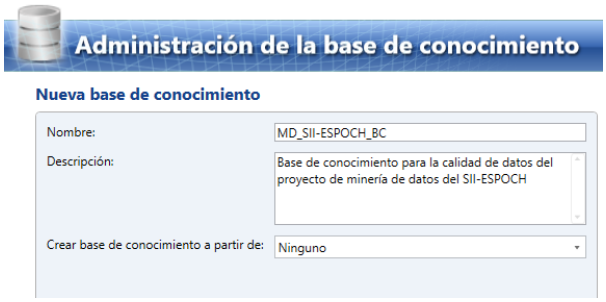
Ingresar en SQL Server Data QualityServices, e iniciar una nueva base de conocimiento. Ver la figura a continuación.

Figura: Nueva base de conocimiento



Agregar los datos de la nueva base de conocimiento como se observa en la figura a continuación. Descripción de la base de conocimiento.

Figura: Descripción de la base de conocimiento

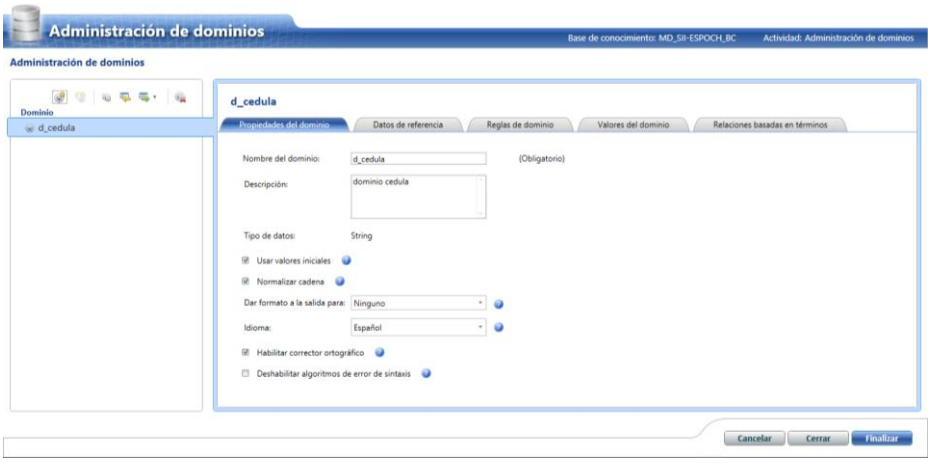


Se ha creado la base de conocimiento MD_OASIS_ESPOCH_BC, que servirá para la revisión de la calidad de datos del proyecto de Data Mining para la Facultad de Informática y Electrónica.

Dominios

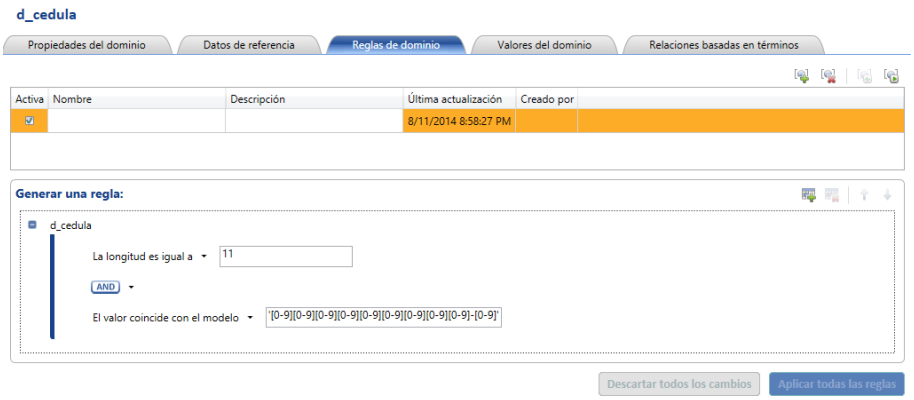
Un dominio representa el conjunto de valores que puede tomar un atributo. Para realizar el test de calidad de datos, la herramienta Data Quality Services ofrece la posibilidad de detectar los problemas dentro del conjunto de datos a través de la creación de dominios y después el descubrimiento de nuevos valores. Por lo tanto, en primer lugar es necesario crear los dominios para los atributos que serán analizados del recurso de origen. En la figura a continuación, se presenta un ejemplo de las descripciones y detalles que corresponden a un dominio.

Figura: Creación de dominios



Es posible agregar clausulas a los diferentes dominios para que la limpieza a realizar posteriormente sea más precisa, en la figura a continuación, se presenta un ejemplo de las reglas agregadas a los dominios.

Figura: Reglas del dominio



Los dominios creados corresponden a los tipos de datos que se identificaron en el conjunto de datos y servirán para identificar los problemas en la calidad de datos. En la tabla a continuación, se resumen los dominios creados para la base de conocimiento:

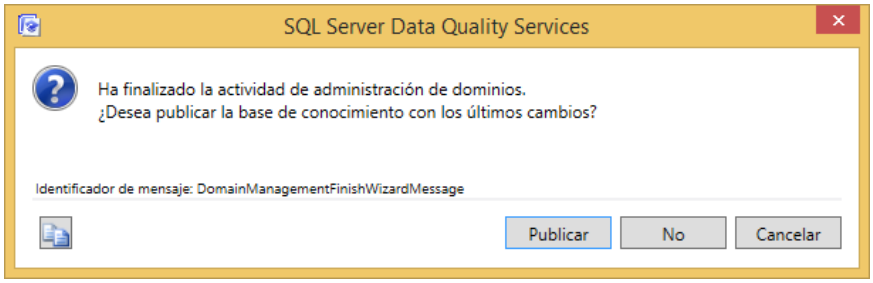
Tabla: Dominios

Nombre	**Descripción**	**Tipo**	**Restricción**
d_cedula	Dominio para los atributos de tipo cédula	Cadena	Contiene 11 caracteres, 10 números y guion
d_ciudad	Dominio para los códigos de ciudades de la base de datos	Cadena	
d_codigoestudiante	Dominio para los códigos de los estudiantes	Cadena	Sólo números
d_codigoint	Dominio para los códigos de tipo entero	Entero	Número positivos
d_codigostring	Dominio para los códigos de tipo string	Cadena	

d_descripciones	Dominio para los atributos que son descripciones	Cadena	
d_fechas	Dominio para las fechas dentro de la base de datos	Fecha	
d_nacionalidad	Dominio para las nacionalidades	Cadena	
d_nombres	Dominio para los atributos nombre y apellido	Cadena	
d_notasfloat	Dominio para los atributos que almacenan notas en formato decimal	Decimal	Números positivos
d_notasint	Dominio para los atributos que almacenan notas de tipo entero	Entero	Números positivos
d_sexo	Dominio para el atributo sexo	Cadena	Acepta las cadenas 'FEM' o 'MAS'

El siguiente paso es publicar los dominios creados. Ver la figura a continuación.

Figura: Publicación de dominios

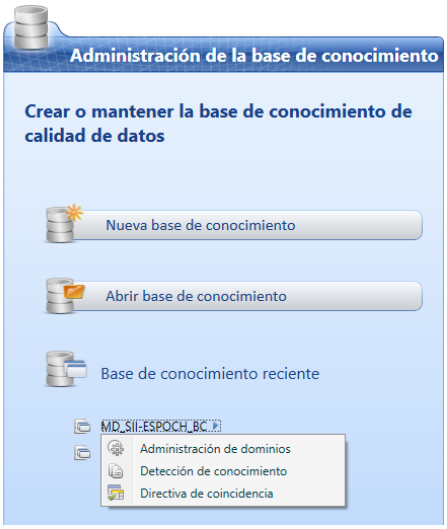


No importa si en el paso siguiente se requiere la creación de más dominios puesto que es posible seguir agregando dominios dentro del proceso consecutivo.

Detección de conocimiento

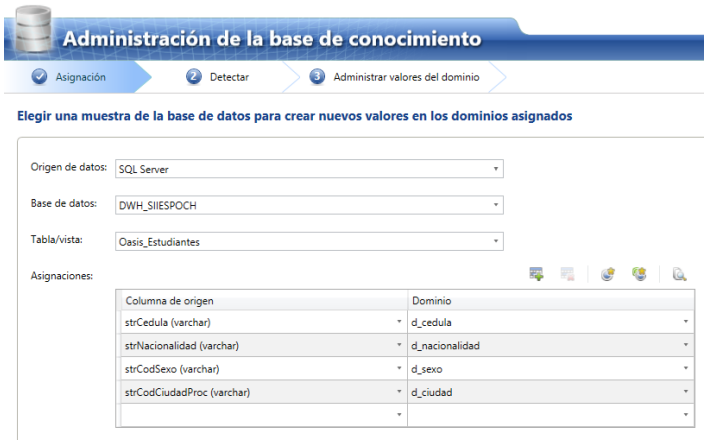
Después de la administración de dominios se procede a la detección de conocimiento que consiste en la identificación de valores para los dominios creados y así consolidar la base de conocimiento antes de proceder al perfilado y limpieza de los datos. En la figura a continuación se muestran las opciones disponibles para una base de conocimiento publicada.

Figura: Acciones de la administración de la base de conocimiento



Al seleccionar la detección de conocimiento se realiza en primera instancia la conexión a la base de datos y a la tabla de la cual se realizará la identificación de patrones. En la figura a continuación, se muestra el primer paso, en el que se relacionan los atributos de la tabla seleccionada con los dominios creados previamente. En caso de no existir el dominio a usar se procede a la creación de uno nuevo.

Figura: Asignación de dominios



El segundo paso es la detección de valores del origen de datos seleccionado. Ver la figura a continuación.

Figura: Detección de valores

Como resultado de este proceso, se han registrado nuevos valores para los dominios creados, sin embargo no todos son correctos, por lo que el siguiente paso es la administración de valores, en donde los datos son analizados para corregir los errores presentes en la data, principalmente al corregir los valores incorrectos. En la figura a continuación, se muestran los diferentes resultados obtenidos para el dominio nacionalidad.

Aquellas opciones que no son válidas son marcadas como incorrectas y se agrega la descripción por la cual se reemplazará.

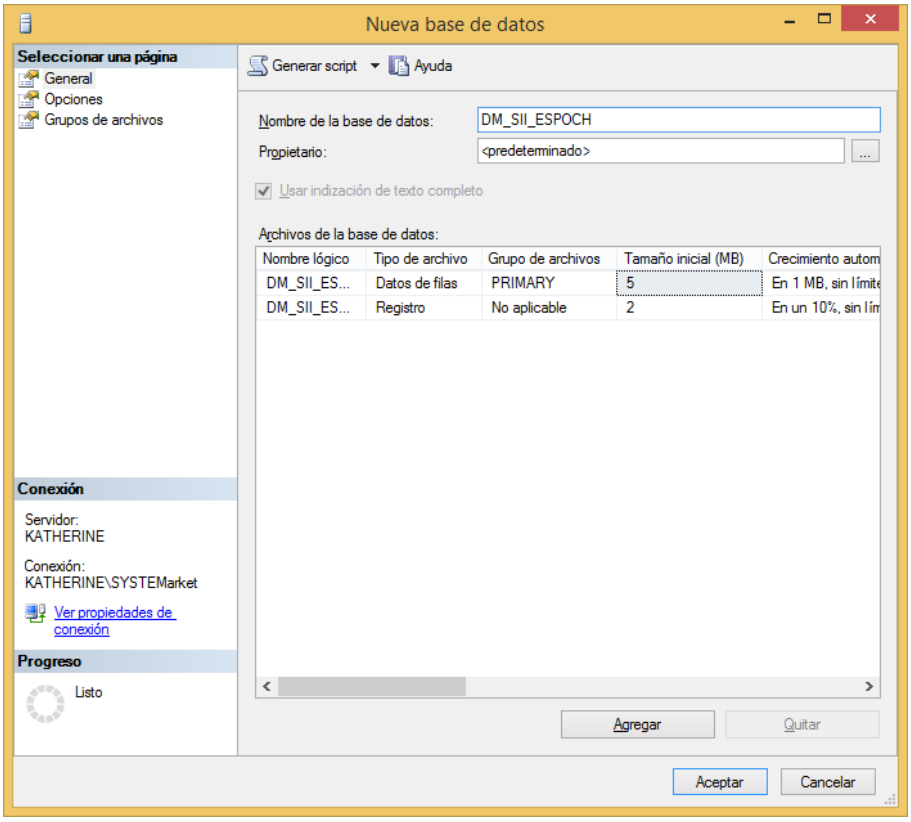
Figura: Administración de valores

Los resultados de la administración de la base de conocimiento son un conjunto de datos que permitirán continuar con el proceso de limpieza de datos.

Limpieza de datos

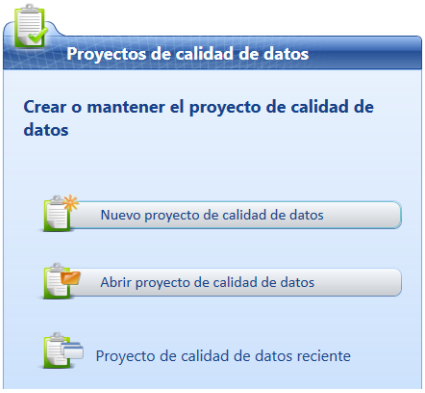
Después de crear la base de conocimiento, los datos de origen son comparados con los dominios creados para detectar los errores y completar el proceso de limpieza con la corrección de los mismos. Los resultados son almacenados en nuevas tablas, las cuales corresponden a una nueva base de datos creada específicamente para el proyecto de minería. En la figura a continuación, se presenta una gráfica del nuevo repositorio creado.

Figura: Base de datos



La base de datos DM_OASIS_ESPOCH, contiene los resultados del proyecto de calidad aplicado a cada una de las tablas identificadas como objeto de análisis. En la figura a continuación, se presenta la opción que ofrece la herramienta Data QualityServices para generar un nuevo proyecto de calidad.

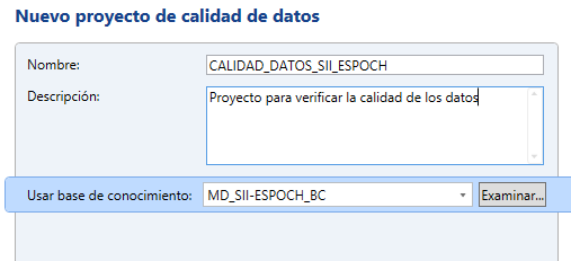
Figura: Nuevo proyecto de calidad



Un proyecto de calidad, permite asociar los campos de una tabla con los dominios de una base de conocimiento, para detectar los errores en los datos, por tal motivo un proyecto de calidad abarca una sola tabla de la base de datos.

En la Figura: Datos de proyecto, se presenta la información necesaria para un nuevo proyecto, que incluye: Nombre del proyecto, Descripción y una Base de conocimiento disponible.

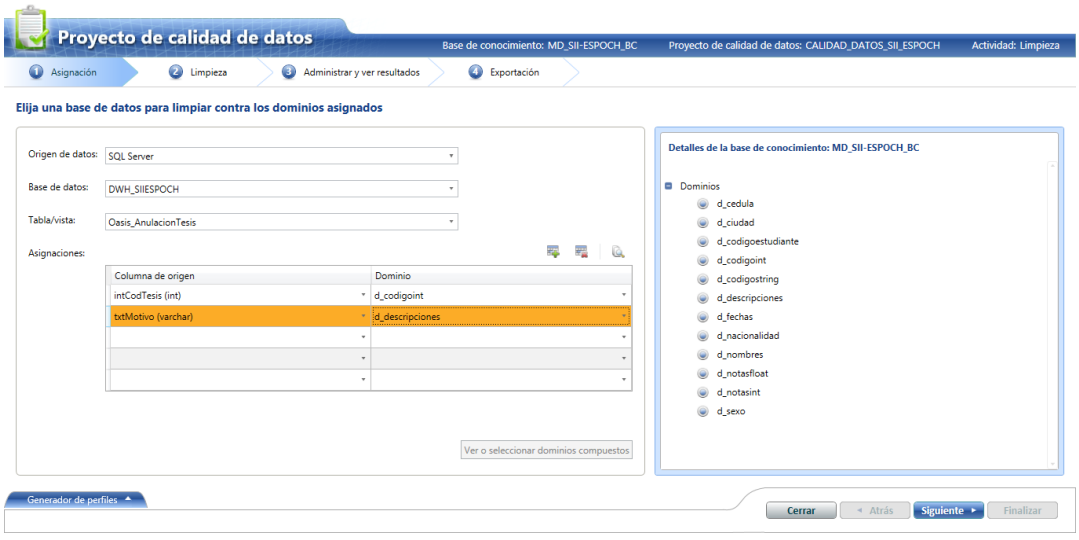
Figura: Datos de proyecto



Anulación-Tesis

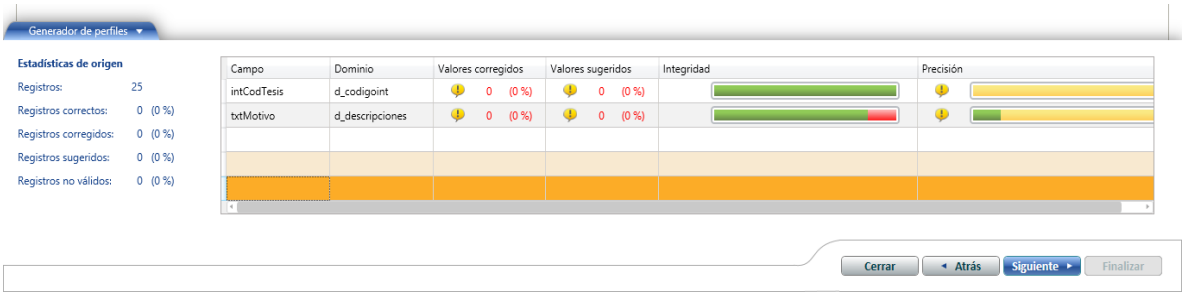
En la Figura: Asignación (Anulación-Tesis), muestra el proceso de asignación de los valores a los dominios de la base de conocimiento con los campos de la tabla seleccionada.

Figura: Asignación (Anulación-Tesis)



Lo que sigue a la asignación es comenzar el proceso de limpieza, en donde los dominios encerrarán los valores correctos e incorrectos del conjunto de datos proporcionado. La Figura: Perfil de datos (Anulación-Tesis), muestra el resultado de limpieza de los datos.

Figura: Perfil de datos (Anulación-Tesis)



Los colores que se muestran en las barras de resultados tienen un significado específico, en donde:

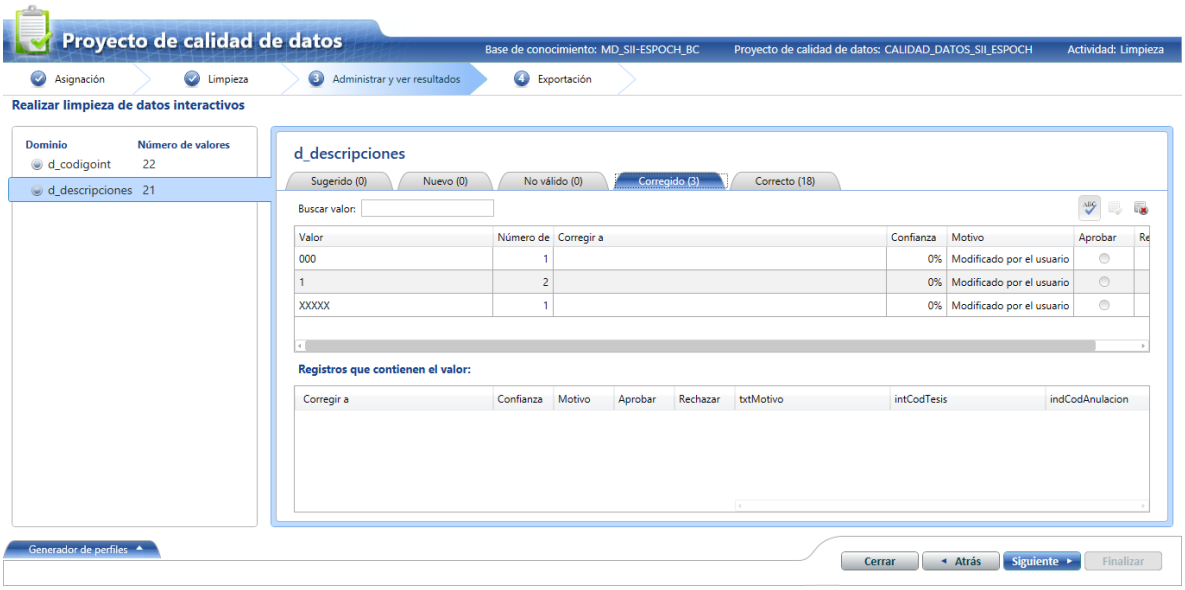
- Verde representa un valor correcto, por lo tanto cumple con todas las reglas declaradas en el dominio correspondiente.
- Amarrillo representa un dato no válido, sin embargo no puede ser considerado como incorrecto, debido a que los valores que entran dentro de este rango representan nuevas entradas para el dominio correspondiente.
- Rojo representa un valor incorrecto, estos datos serán reemplazados por nuevos valores que hayan sido definidos previamente en el dominio, o por valores que se pueden determinar en el siguiente paso del proyecto de calidad.

Como resultado del perfilado de datos para la tabla Anulación-Tesis, se obtiene un 100% de integridad en los códigos y un 88% de integridad en las descripciones.

En la Figura: Administrar resultados (Anulación-Tesis), se presenta la pantalla de administración de resultados en donde según el dominio asignado a cada atributo se presentan los valores correctos, los sugeridos en caso de corrección de textos, los valores

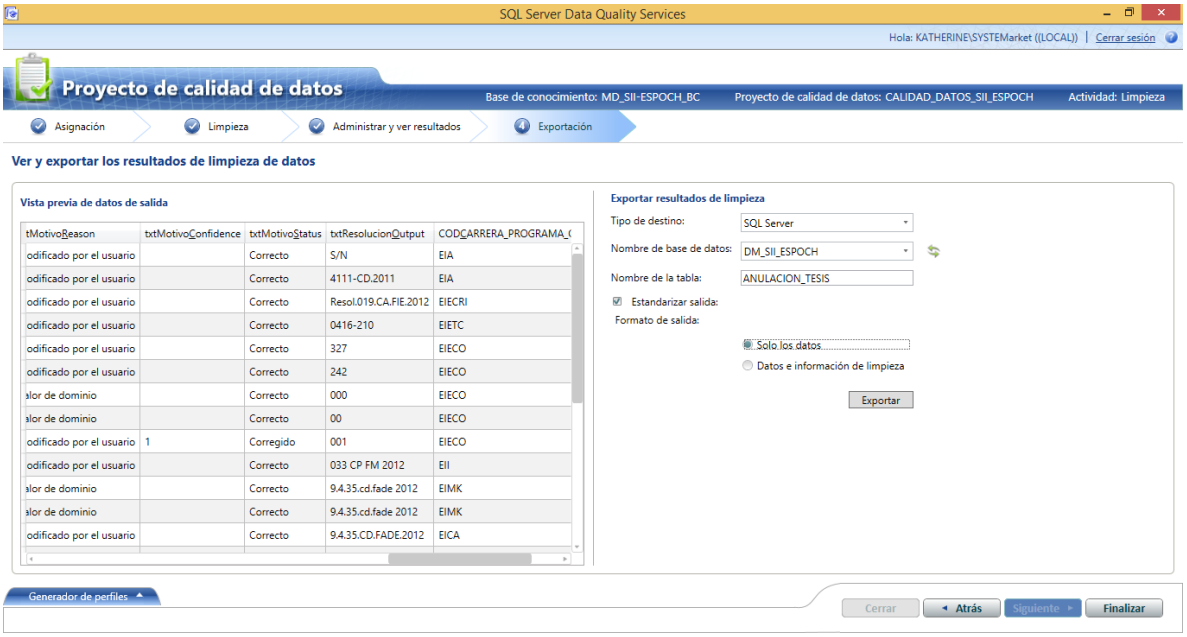
nuevos encontrados para el dominio, las correcciones realizadas y los valores correctos. Para la tabla Anulación-Tesis, se realice la corrección de tres valores, lo que representa un 14% del total de las filas analizadas.

Figura: Administrar resultados (Anulación-Tesis)



Para culminar el proceso de limpieza es necesario guardar los resultados, como se presenta en la Figura: Exportar resultados (Anulación-Tesis), la herramienta Data QualityServices ofrece las opciones de almacenar los datos solos o junto a la información de los procesos realizados durante la limpieza. Para este proyecto se almacenan solo los datos de salida, en la base de datos MD_OASIS_ESPOCH, creada previamente con el propósito de alojar los resultados del proyecto de calidad de datos.

Figura: Exportar resultados (Anulación-Tesis)



El proceso descrito con la tabla Anulación-Tesis, se siguió para las otras tablas seleccionadas, y los resultados arrojaron: problemas en la descripción de la nacionalidad y problemas de datos nulos en atributos de tipo numéricos no claves, los mismos que fueron solucionados desde los scripts de selección de datos, reemplazando los valores erróneos de la nacionalidad y asignando un valor de 0 a los nulos existentes.

Una vez culminado el proyecto de limpieza, es posible proceder a consultar los resultados dentro de la base de datos creada para el propósito de albergar los nuevos datos.

Atributos seleccionados

Tras el análisis de la fuente original de datos, la limpieza de los mismos y el almacenamiento de los resultados en nuevo repositorio, se cuenta con un nuevo conjunto de datos de los cuales se han escogido los atributos presentados en la Tabla: Atributos seleccionados para formar la estructura que servirá en el proceso de modelado de minería de datos.

Tabla: Atributos seleccionados

Atributo	**Descripción**	**Tabla**
Nombres	Nombres y apellidos de los estudiantes	Estudiante
Edad	Atributo calculado de la edad de nacimiento de cada estudiante.	Estudiante
Nacionalidad	Atributo nacionalidad de los estudiantes que representa el país de nacimiento de los estudiantes	Estudiante
Sexo	Atributo sexo de cada uno de los registros que representan a los estudiantes	Estudiantes
Graduado	Atributo calculado de que representa si un estudiante se ha graduado o no.	Graduados
Fecha de graduación	Fecha en la que el estudiante se registró como graduado en la base de datos.	Graduados
Promedio de graduación	Promedio que se registra en la graduación de un estudiante y que representa la nota final del proceso de estudios	Graduados
Promedio Notas	Atributo calculado que representa el promedio actual que registra el estudiante	Materias aprobadas

Fecha de inscripción	Fecha en la que se registra la inscripción del estudiante	Inscripciones
Materias aprobadas	Atributo calculado del número de materias aprobadas de un estudiante.	Materias aprobadas
Numero de materias con segunda matrícula	Atributo calculado del número de materias con segunda matrícula que se registra del estudiante	Materias aprobadas y asignadas
Numero de materias con tercera matrícula	Atributo calculado del número de materias con tercera matrícula que se registra del estudiante	Materias aprobadas y asignadas
Tiempo-Egresado-Tesis	Atributo calculado que representa la diferencia de tiempo entre la fecha en la estudiante egresa de su carrera hasta la fecha en la que entrega su proyecto de tesis	Terminación Tesis Egresados
Numero-Matriculas	Atributo calculado del número de matrículas que ha realizado un estudiante	Matriculas
Anulación-Tesis	Registra si un estudiante ha presentado o no una anulación de su tesis de grado	Anulación-Tesis
Egresado	Atributo que registra si un estudiante ha sido o no registrado como egresado.	Egresados

Fecha-Egresado	Fecha en la que se registra al estudiante como egresado	Egresados
Escuela	Atributo que representa la escuela a la que pertenece el estudiante	Estudiante-Carrera Escuelas
Tiempo-Inscripción-Egresado	Atributo calculado que representa la diferencia de tiempo entre la fecha de inscripción y la fecha de egresado del estudiante	Egresados Inscripciones
Tiempo-Inscripción-Egresado	Atributo calculado que representa la diferencia de tiempo entre la fecha de inscripción y la fecha de egresado del estudiante	Egresados Inscripciones
Carrera	Carrera a la que pertenece el estudiante	Carreras
Ciudad-Procedencia	Ciudad de procedencia registrada	Estudiante

Dentro de la estructura también se definen datos calculados de los datos originales, los mismos que no representan un cambio en el fondo de la información si no de forma de representación final de los datos, con el objetivo de contar con un mayor número de datos para el análisis posterior.

ANEXO 3: INFORME DE IMPLEMENTACIÓN DE MINERÍA DE DATOS

Proyecto: Minería de datos aplicado al área académica de la FIE-ESPOCH

Director: Ing. Iván Menes

Analista: Katherine Gallegos

INFORME

Tras los procesos de análisis, recolección y preparación de los datos, se realizó la evaluación de los algoritmos de minería de datos seleccionados para el proyecto, y el algoritmo de Árbol de Decisión resultó ser de mejor desempeño que el algoritmo de Regresión Logística de Microsoft para este trabajo. Bajo esa derivación, se procedió a implementar el algoritmo sobre los datos del área académica de la FIE-ESPOCH, y los resultados se resumen en el presente documento.

Resultados

El algoritmo de Árbol de Decisión presenta los resultados de la minería de datos en forma de un árbol basado con reglas que dividen a los diferentes grupos encontrados. A continuación se presenta un ejemplo de los resultados de los requisitos planteados para este proyecto.

Requisito 1

Descripción: Determinar patrones de comportamiento para el ingreso de los estudiantes; por carrera y facultad.

Figura: Árbol para requisito 1 por facultad

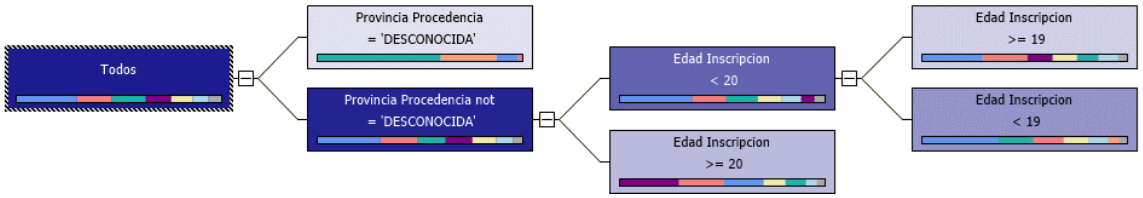
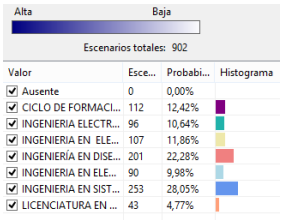
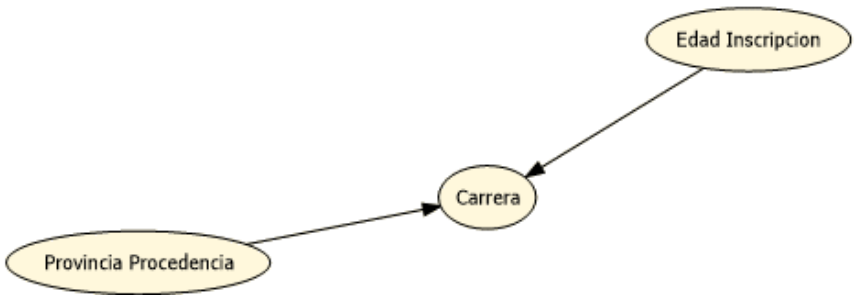


Figura: Leyenda para árbol de requisito 1



Las figuras árbol y leyenda con respecto al requisito uno, muestran que la mayoría de los estudiantes de la FIE corresponden a la carrera de Ingeniería en Sistemas, mientras que el menor porcentaje corresponde a los estudiantes inscritos para la carrera de Licenciatura en Diseño Gráfico. Para todas las carreras, la edad de los estudiantes al momento de la inscripción oscila entre 19 y 20 años de edad siendo el patrón hallado en el proceso de inscripción. El árbol de dependencias de la figura a continuación, indica que los aspectos que influyen sobre la selectividad de carrera son: provincia de procedencia y edad de inscripción, en ese orden de mayor a menor.

Figura: Red de dependencias para requisito 1



Requisito 3

Descripción: Determinar patrones de comportamiento para la matriculación y selección de asignatura de los estudiantes (número de asignaturas, número de créditos, nivel y área de las asignaturas, etc.); por carrera, por nivel y por facultad.

Figura: Árbol para requisito 3 (EDG-Áreas)

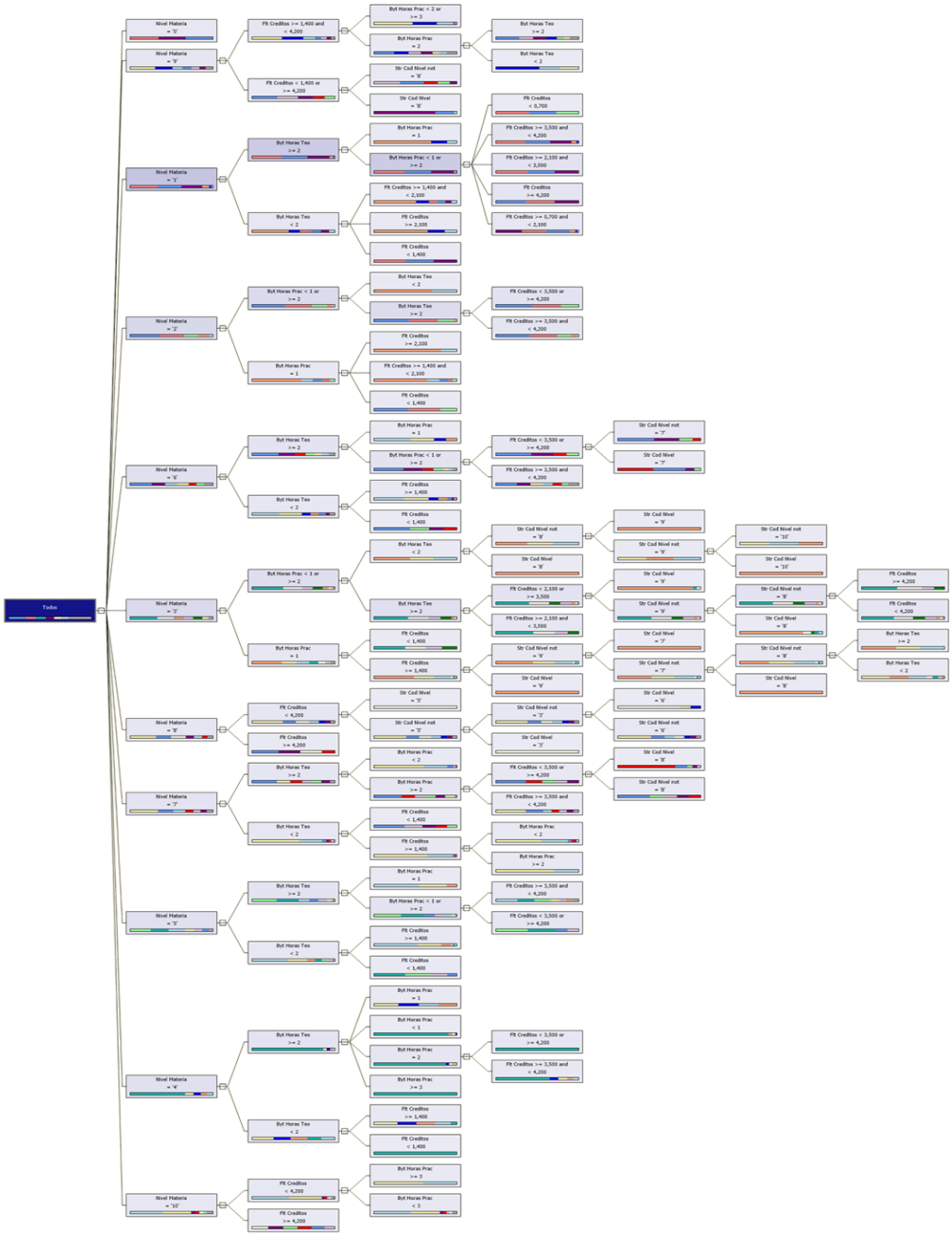
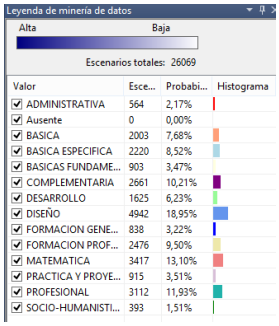
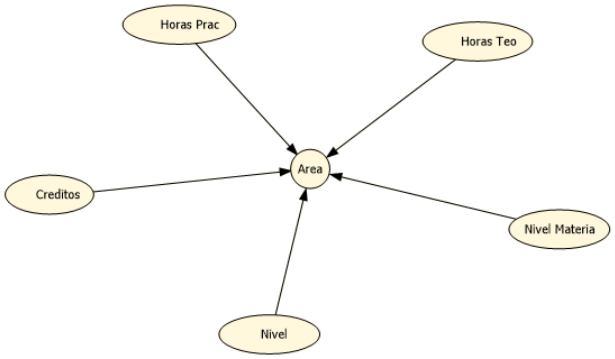


Figura: Leyenda para árbol de requisito 3



Las figuras árbol y leyenda de requisito 3, son resultados del requisito de minería sobre los datos de la Escuela de Diseño Gráfico concerniente al factor de decisión del área de las materias en el proceso de matriculación de los estudiantes. Del total de datos, existe una probabilidad de que 18,95% accedan a un área de Diseño, siendo ésta la más alta; mientras que la probabilidad de que la matrícula se asocie al área Socio-Humanística es de 1,51%, representando éste el valor más bajo. El árbol de dependencias de la figura a continuación, indica que los aspectos que influyen sobre la elección del área de la matrícula en la Escuela de Diseño Gráfico son: nivel de la matrícula, los créditos de la materia, las horas prácticas de la materia, las horas teóricas de la materia y el nivel de la materia, en ese orden de menor a mayor.

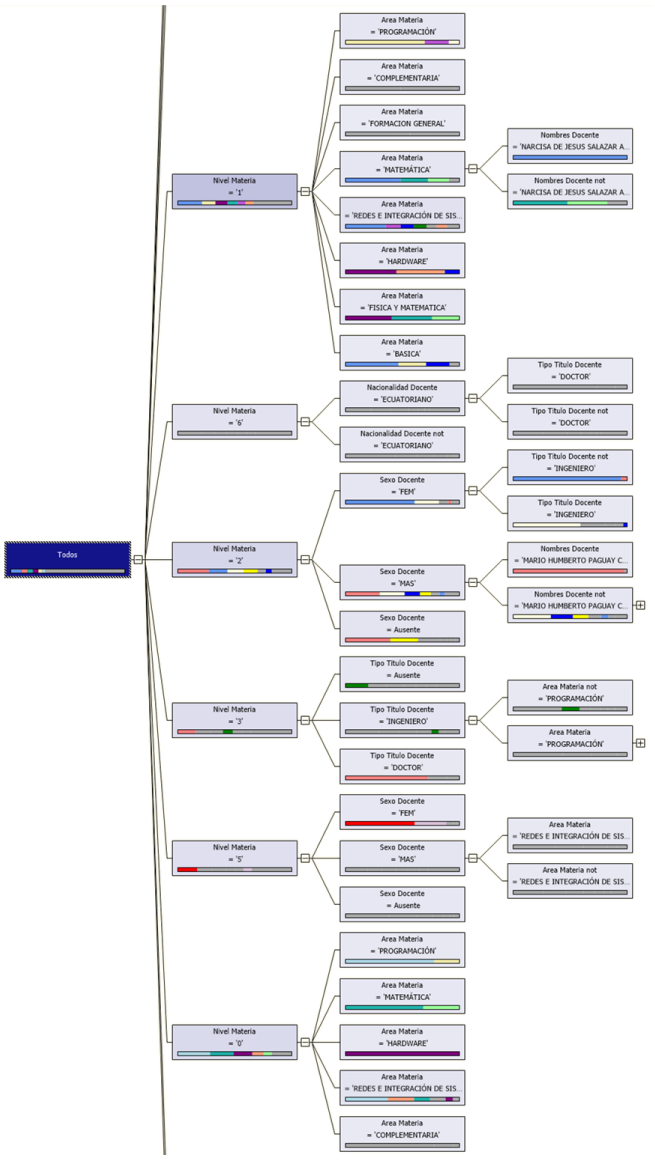
Figura: Red de dependencias para requisito 3



Requisito 4

Descripción: Determinar los factores que tienen influencia en los casos de deserción (retiros y pérdida de la asignatura por asistencia), por carrera, niveles, asignaturas y facultad.

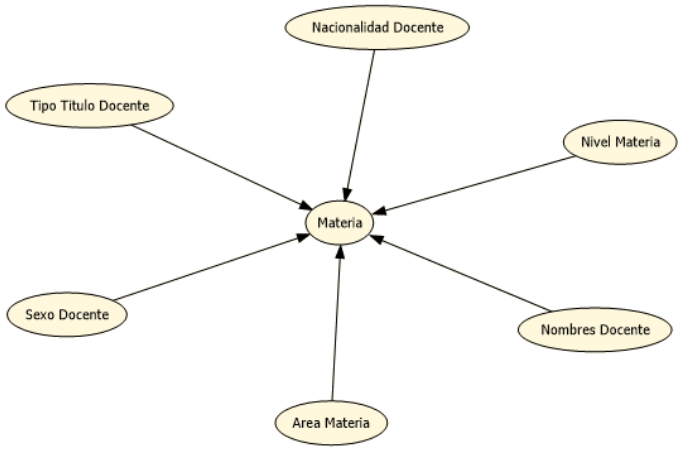
Figura: Árbol para requisito 4 (EIS)



La Figura: Árbol para requisito 4, muestra los resultados de minería sobre los datos de la Escuela de Ingeniería en Sistemas concerniente al factor de deserción de los estudiantes.

Del total de deserciones, las materias von mayores registros son Física, Matemática y Fundamentos de Programación. El árbol de dependencias de la figura a continuación, indica que los aspectos que influyen sobre la deserción en la Escuela de Ingeniería en Sistemas son: nivel de la materia, área de la materia, sexo del docente, tipo de título del docente, nacionalidad del docente y nombre del docente, en ese orden de mayor a menor.

Figura: Red de dependencias para requisito 4



Requisito 5

Descripción: Determinar patrones de comportamiento en la promoción académica de los estudiantes; por asignatura, nivel, áreas de conocimiento, carrera y facultad.

Figura: Árbol de decisión para requisito 5

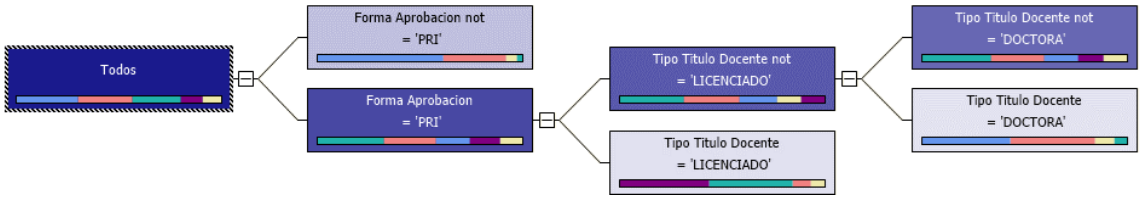
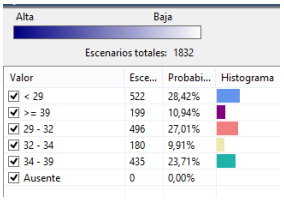
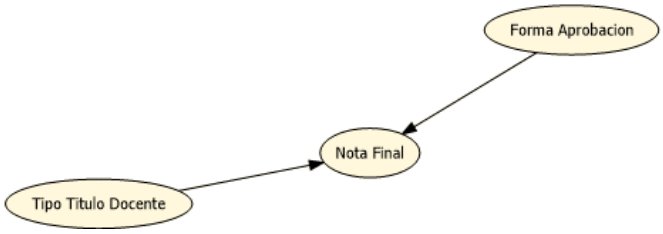


Figura: Leyenda para árbol de requisito 5



En las figuras: árbol y leyenda para requisito 5, se presentan los resultados de minería sobre la promoción de los estudiantes de la Carrera de Ingeniería en Sistemas en la Extensión Macas. Del total de registros existe un 28,42% de que los estudiantes aprueben las asignaturas con menos de 29 puntos. La probabilidad más baja es de 9,91% que los estudiantes aprueben una asignatura con un total de entre 32 y 34 puntos. La red de dependencias para este caso se presenta en la figura a continuación, y los factores que inciden en la promoción académica de los estudiantes de la Ingeniería en Sistemas extensión Macas son: el tipo de título del docente, y la forma de promoción, en ese orden de menor a mayor relevancia.

Figura: Red de dependencias para requisito 5



Requisito 6

Descripción: Determinar los factores que influyen en los escenarios de segunda y tercera matrícula; por asignatura, nivel, áreas de conocimiento, carrera y facultad.

Figura: Árbol de decisión para requisito 6

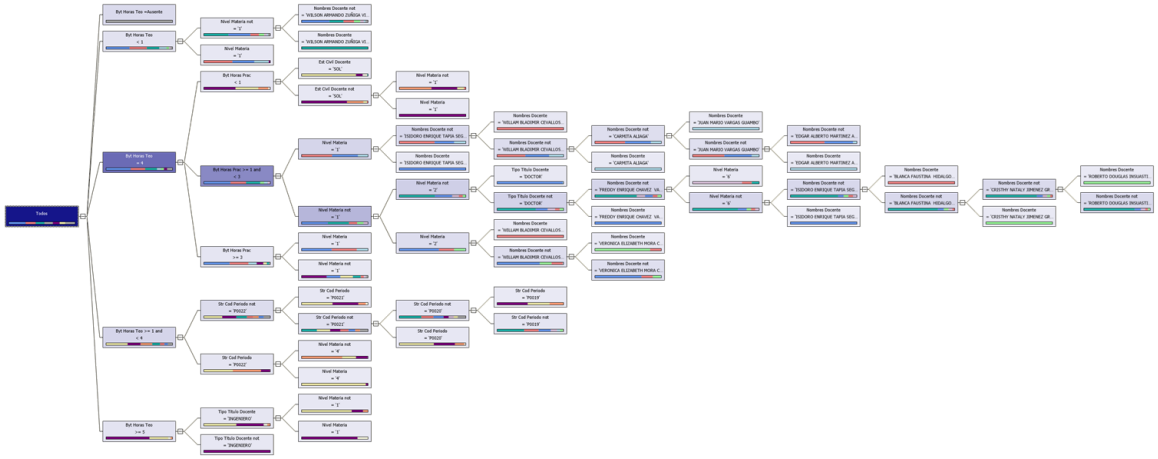
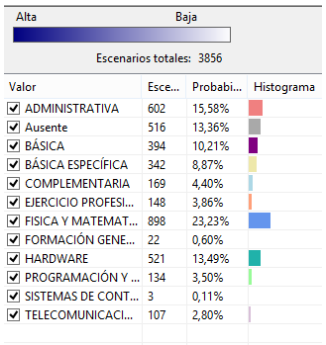


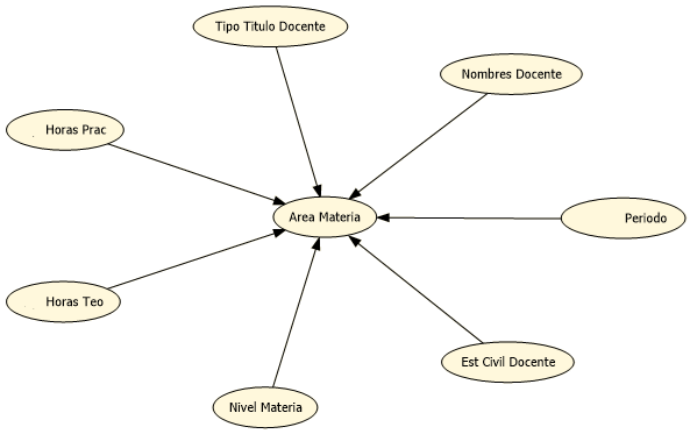
Figura: Leyenda para el árbol de requisito 6



Valor	Esce...	Probabi...	Histograma
ADMINISTRATIVA	602	15,58%	
Ausente	516	13,36%	
BÁSICA	394	10,21%	
BÁSICA ESPECÍFICA	342	8,87%	
COMPLEMENTARIA	169	4,40%	
EJERCICIO PROFESI...	148	3,86%	
FISICA Y MATEMAT...	898	23,23%	
FORMACIÓN GENE...	22	0,60%	
HARDWARE	521	13,49%	
PROGRAMACIÓN Y ...	134	3,50%	
SISTEMAS DE CONT...	3	0,11%	
TELECOMUNICACI...	107	2,80%	

Las figuras: árbol y leyenda para requisito 6, muestran los resultados de minería sobre los datos de la Escuela de Ingeniería Electrónica en Telecomunicaciones y Redes para los casos de segunda y tercera matrícula. Del total de registros, las áreas con mayor posibilidad de casos de segunda matrícula son Física y Matemática y Administrativa. Mientras que la probabilidad de segunda o tercera matrícula en las áreas de Formación general y Sistemas de control es más baja. La red de dependencias presentada en la figura a continuación indica que inciden en casos de segunda tercera matrícula los siguientes aspectos: estado civil del docente, docente, periodo, tipo del título del docente, nivel de la materia, horas teóricas y horas prácticas, en ese orden de menor a mayor trascendencia.

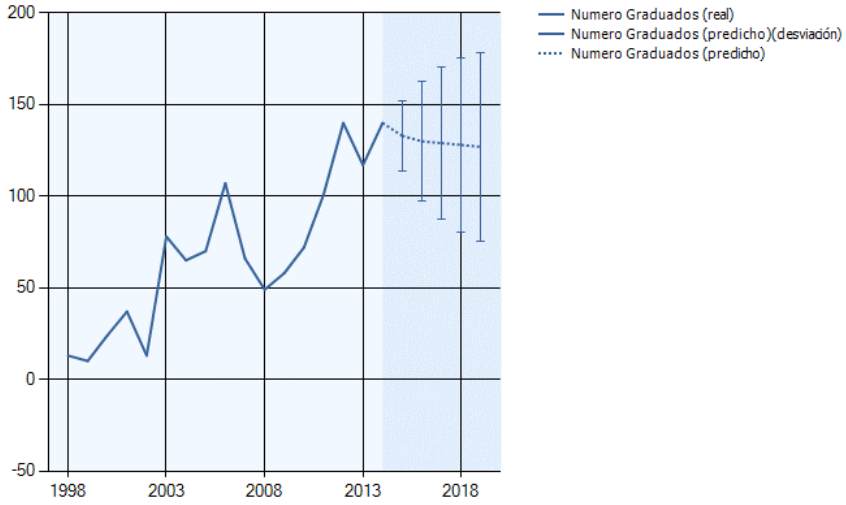
Figura: Red de dependencias para el requisito 6



Requisito 7.

Descripción: Determinar la proyección de graduados; por carrera y facultad.

Figura: Serie temporal para el requisito 7



El la Figura: Serie Temporal para el requisito 7, se muestran los resultados de minería sobre los datos de la FIE para determinar el número de graduados en el tiempo. Con los valores históricos se aplicó el algoritmo de serie temporal el cual permite obtener una proyección

del número de graduados, para el 2015 se prevé un total de 133 graduados mientras que para el 2015 130.

Requisito 8.

Descripción: Determinar los factores que inciden en los casos de estudiantes con baja eficiencia terminal; por carrera y facultad.

Figura: Árbol de decisión para requisito 8

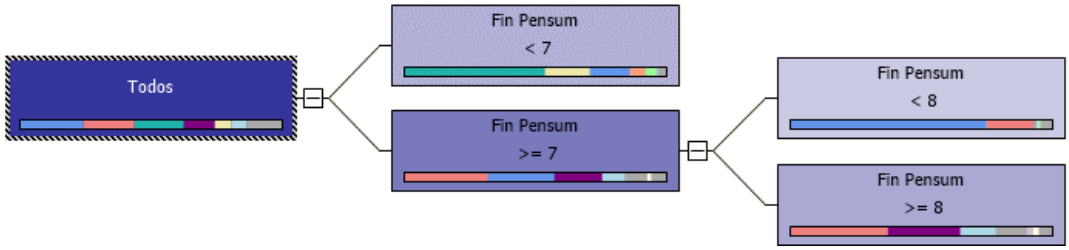
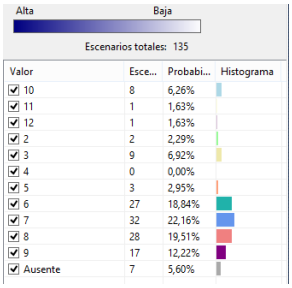


Figura: Leyenda para el árbol de requisito 8



Las figuras: árbol y leyenda para requisito 8, muestran los resultados de minería sobre los datos de la Escuela de Ingeniería Electrónica en Control y Redes Industriales para el análisis de la eficiencia terminal de los estudiantes. Del total de registros, la probabilidad de que un estudiante tenga un baja eficiencia terminal supera el 50%, es decir que un estudiante culmine sus estudios en la carrera en más de cinco años. El valor que incide significativamente sobre la eficiencia terminal es el tiempo que le toma al estudiante culminar su pensum de estudios.

Plan de Mantenimiento

El plan de mantenimiento propuesto está diseñado para permitir la actualización de los modelos de minería de datos entregados en el proyecto. El aumento de datos en los repositorios asociados, no representará un proceso de mantenimiento profuso; por otro lado, cambios en la definición de las tablas requerirá un análisis más detallado para identificar las afectaciones dentro de las estructuras y por ende en los modelos de minería de datos.

Plan de mantenimiento	
Caso	Incremento de filas en las tablas relacionadas a las estructuras de minería
Frecuencia	Semestral
Actividades	Validar nuevas entradas que afecten la limpieza de datos. De encontrar nuevos datos incorrectamente ingresados proceder a la limpieza y actualizar las estructuras de datos. Generar los modelos de minería

Plan de mantenimiento	
Caso	Cambios en la definición de las tablas asociadas a los datos de los procesos académicos de la FIE

Frecuencia	Semestral
Actividades	Identificar los atributos o relaciones añadidas o eliminadas. Identificar si los elementos eliminados son parte de las actuales estructuras de minería de datos Realizar un análisis sobre la afectación de los elementos eliminados y añadidos en la definición de las tablas asociadas Decidir si mantener, modificar o eliminar la estructura de datos. Implementar los nuevos modelos en los casos de no eliminar la estructura de datos.