Universidad Nacional Autónoma De Nicaragua
Recinto Universitario “Rubén Darío”
Facultad de Ciencias e Ingenierías
Departamento de Computación

Tema: Minería de datos

Subtema: Desarrollo de un modelo basado en minería de datos aplicando el algoritmo de **Decision Trees** en el área de **matrículas** del Instituto de Computación y Comercio (ITEC) en Managua en el II Semestre del 2015

Integrantes:

- Br. María Margarita Flores Velásquez
- Br. Mario José Cerna Suárez
- Br. Donis Santiago Montoya Poveda

Tutor:

- MSc. Luis Miguel Martínez Olivera

“Managua, 09 de Noviembre del 2015”

AGRADECIMIENTO

Damos gracias a Dios por su infinita misericordia, su fuerza, valor y sabiduría para nuestro equipo de trabajo.

A personas con ardua labor de enseñanza y dignidad humana como los son nuestros profesores, que día a día compartieron sus valiosos conocimientos, con el objetivo de formarnos como profesionales.

También a nuestros padres que con mucho empeño nos inculcaron los valores más importantes los cuales son: el amor, el respeto, la tolerancia, la confianza y la responsabilidad.

Y por supuesto a nuestra Universidad como es La Unan-Managua por habernos brindado la maravillosa oportunidad de destacarnos como profesionales en la Carrera de Licenciatura en Computación.

Autores

DEDICATORIA

Primeramente a dios por haberme permitido llegar hasta este punto y haberme dado salud, ser el manantial de vida y darme lo necesario para seguir adelante día a día para lograr mis objetivos, además de su infinita bondad y amor.

A mi madre por haberme apoyado en todo momento, por sus consejos, sus valores, por la motivación constante que me ha permitido ser una persona de bien, pero más que nada, por su amor. A mi padre por los ejemplos de perseverancia y constancia que lo caracterizan y que me ha infundado siempre, por el valor mostrado para salir adelante y por su amor. Y a todos aquellos que ayudaron directa o indirectamente a realizar este documento.

A mis maestros por su gran apoyo y motivación para la culminación de nuestros estudios profesionales, por su apoyo ofrecido en este trabajo, por haberme transmitidos los conocimientos obtenidos y haberme llevado pasó a paso en el aprendizaje.

(Mario Cerna)

Siendo este trabajo de Seminario de Graduación, el reflejo de nuestro mayor esfuerzo como universitario, se lo dedico a:

- Dios, quien nos dio el valor y la fuerza de vencer todos los obstáculos que se nos presentaron en el camino para realizar este trabajo.

- Mi madre, quien fue la fuente de inspiración, la que nos dio el ánimo de seguir adelante y poder finalizarlo, y por darnos la oportunidad de estudiar y llegar a ser alguien en la vida.

- Maestros(as), quienes con su dedicación y paciencia nos brindaron todos sus conocimientos y enseñanzas en las diferentes asignaturas que recibimos durante los años de preparación en la Carrera de Licenciatura en Computación y en la conclusión del trabajo de seminario de graduación.

(Margarita Flores)

Le agradezco al Dios de los Cielos por haber permitido clasificar, superar los retos que vinieron acompañando esta carrera y posteriormente culminar con gran esfuerzo y mucho gozo el sentirme realizado en dicha meta.

A mis compañeros de clases y amigos que jugaron un papel muy importante en poder culminar con satisfacción.

A mis profesores que tuvieron la delicadeza y paciencia de impartirnos las materias con mucho profesionalismo, humanismo y compañerismo.

A mi familia por apoyarme con consejería, paciencia, económicamente y sobre todo a mi Esposa por soportar que estuviese fuera de casa todo el tiempo que duro la carrera.

(Donis Montoya)

INDICE

I. RESUMEN **10**

II. INTRODUCCIÓN **11**

III. ANTECEDENTES **12**

IV. PLANTEAMIENTO DEL PROBLEMA **13**

4.1. CARACTERIZACIÓN DEL PROBLEMA 13
4.2. DELIMITACIÓN DEL PROBLEMA 13
4.3. FORMULACIÓN DEL PROBLEMA 13
4.4. SISTEMATIZACIÓN DEL PROBLEMA 14

V. JUSTIFICACIÓN **15**

VI. OBJETIVOS **16**

OBJETIVO GENERAL 16
OBJETIVOS ESPECÍFICOS 16

VII. MARCO TEÓRICO **17**

7.1. DATOS 18
7.1.1. Datos para la informática. *18*
7.2. INFORMACIÓN 19
7.3. CONOCIMIENTO 20
7.4. SISTEMAS DE INFORMACIÓN 21
7.4.1 Ciclo de desarrollo de un sistema de información *21*
7.5. BASES DE DATOS RELACIONALES 22
7.5.1 Base de Datos: *22*
7.5.3 Base de datos relacionales *23*
7.6. MINERÍA DE DATOS 24
7.6.1 Las bases de datos y la minería de datos: *25*
7.6.2 Las fases de la minería de datos: *25*
7.7. TÉCNICAS DE MINERÍA DE DATOS 30
7.8. ALGORITMOS DE MINERÍA DE DATOS: 31
El algoritmo que se utilizó en nuestra aplicación es: *32*
ÁRBOLES DE DECISIÓN *32*
7.9. MODELO QUE APLICAN PARA LA ELABORACIÓN DE LA APLICACIÓN 38
7.10. HERRAMIENTAS DE MINERÍA DE DATOS 39
7.10.1 SQL Server Management Studio *41*
7.10.2 Transformaciones y tareas de minería de datos en Integration Services *41*
7.11. DATAWAREHOUSE 42
7.12. NORMA ISO-IEC 9126-1 "CARACTERÍSTICAS DE CALIDAD Y SUBCARACTERÍSTICAS" (GUTIERREZ, 2014) 45
7.13. RESEÑA HISTÓRICA DE LA EMPRESA 47

VIII. HIPÓTESIS **49**

IX. DISEÑO METODOLOGICO **50**

9.1 TIPOS DE ESTUDIO 50

9.2 AREA DE ESTUDIO 50
9.3 UNIVERSO 50
9.4 MUESTRA 51
9.5 INSTRUMENTOS DE RECOLECCIÓN DE DATOS 51
9.6 ANÁLISIS DE FACTIBILIDAD TÉCNICA 51
1. En cuanto a Software 51
9.7 ANÁLISIS DE FACTIBILIDAD ECONÓMICA: 52
1. En cuanto a Hardware: 52
2. Material didáctico y muebles de trabajo 52
9.8 ANÁLISIS DE FACTIBILIDAD OPERATIVA: 53
9.9 HERRAMIENTAS UTILIZADAS PARA ELABORAR EL MODELO DE MINERÍA DE DATOS 54
9.10 FASES DE MINERIA DE DATOS 54

X. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS 55

10.1 BASE DE DATOS 56
10.2 DATAWAREHOUSE 57
10.3 INTEGRATION SERVICE 61
10.4 ANÁLISIS SERVICE 63

XI. CRONOGRAMA DE TRABAJO 73

XII. CONCLUSIONES 74

XIII. RECOMENDACIONES 75

XIV. BIBLIOGRAFIA 76

XV. COMPENDIOS 77

ÍNDICE DE FIGURAS

Figura 1: Minería de Datos 24
Figura 2: Fases de Minería de Datos 26
Figura 3: Árboles de Decisión 33
Figura 4: Histograma1 35
Figura 5: Histograma2 35
Figura 6: Formula de Regresión 36
Figura 7: Ecuación de Regresión 37
Figura 8: Datawarehouse 43
Figura 9: Base de Datos 56
Figura 10: Datawarahouse 57
Figura 11: Integration Service 61
Figura 12: Dimensiones 62
Figura 13: Árbol de Decisiones 63
Figura 14: Evaluación de la Norma ISO 71
Figura 15: Segunda Evaluación de la Norma ISO 72
Figura 16: Cronograma de Trabajo 73

ÍNDICE DE TABLAS

Tabla 1: Descripción del Software 52
Tabla 2: Reporte de Estado de estudiantes 68
Tabla 3: Reporte de Horario de docentes 69
Tabla 4: Reporte de Estudiantes 69
Tabla 5: Evaluación del modelo según Norma ISO 71
Tabla 6: Evaluación de Modelo según la Norma ISO 70

I. RESUMEN

En este trabajo se aplicó el análisis e implementación de un modelo de minería de datos utilizando el algoritmo Árboles de Decisión para el Instituto Técnico de Computación y Comercio “ITEC”.

En primera instancia se analizó la información del instituto mediante los métodos de recolección de datos, definiendo las necesidades y problemáticas que presenta el área de Matricula, lo que conllevo a aplicar el modelo de minería de datos.

Como siguiente paso se presentó una justificación en donde se detallaron los beneficios que se obtendrán mediante la aplicación del modelo en el área de matrícula para predecir el retiro de los estudiantes en el instituto. Una vez ya planteada la justificación se detallaron los objetivos a cumplir para este proyecto.

Se manejaron una serie de conceptos básicos, características y subcaracterísticas, herramientas y métodos de minería de datos que se utilizaron para la obtención del conocimiento sobre el tema a implementarse en el trabajo expuesto.

Se presentan el análisis de resultados y el desarrollo del modelo de minería de datos aplicando árboles de decisión, los cuales dieron solución a cada uno de los objetivos expuestos.

El primer resultado es el análisis sobre los procesos en el área de Matrícula así como la base de datos transaccional que utiliza el instituto, el segundo fueron los procedimientos y técnicas que se aplicaron para la implementación del algoritmo de árboles de decisión, el tercer fue la muestra de reportes para la ayuda en la toma de decisiones en el área de matrícula y último es la evaluación por parte del usuario final del método aplicado al proyecto.

II. INTRODUCCIÓN

La informatización de las empresas, organizaciones e instituciones ha generado un gran incremento de la información almacenada en las bases de datos, la cual es de gran utilidad cuando se quiere explicar el pasado, entender el presente y predecir la información futura, por lo que se hace necesario analizar la misma para la obtención de información útil para la organización.

Una de las técnicas más usadas para obtener conocimiento analizando los datos presentes en las bases de datos es la minería de datos, que permite obtener patrones o modelos a partir de los datos recopilados.

Durante la investigación se abordó inicialmente, el análisis exhaustivo, sistemático e integrado del dominio de conocimiento y el dominio de aplicación en los que se enmarca dicha investigación.

Adicionalmente, se ejecutaron las fases de un proyecto de Minería de Datos como guía en la construcción de un modelo predictivo que generó el conocimiento en situaciones problemáticas presentadas durante el proceso de matrícula en el Instituto de Computación y Comercio, y en base al modelo mencionado se implementaron mejoras, brindando el aseguramiento de la calidad de enseñanza de cada curso que ofrece el instituto técnico.

De esta manera, los resultados de este estudio revisten gran importancia, ya que la obtención de los objetivos propuestos contribuye a la investigación, al desarrollo incremental e innovación en el Instituto de Computación y Comercio (ITEC).

III. ANTECEDENTES

En la actualidad, la implantación de la Minería de Datos para recuperar información en las diversas organizaciones (bibliotecas y centros de documentación) así como en el ámbito empresarial es una técnica habitual. Tradicionalmente, quienes más han empleado las técnicas de la Minería de Datos para recuperar información han sido las relacionadas con la publicidad y con los negocios de la distribución. Sin embargo, existen multitud de áreas que han integrado en su actividad las técnicas de la Minería de Datos para Recuperar Información.

- Minería de datos aplicada a la mejora del fraccionamiento de aceite de palma en una planta de producción (Acupalma), Febrero 2011 Universidad Nacional Abierta vicerrectorado académico área de ingeniería. Acarigua, República Bolivariana de Venezuela.

- Metalco es una empresa manufacturera del sector de metalurgia, con cuarenta y seis años de existencia en el mercado costarricense. Pertenece a un grupo corporativo de capital colombiano, propietaria de más de veinte empresas en el ámbito latinoamericano.

Esta empresa se dedica a la galvanización de acero, esmaltado y formación de láminas para techo, dentro de las que destacan la teja Toledo, así como la fabricación de perfiles y tubería, tanto galvanizada como de hierro negro para múltiples usos. A través de su historia, Metalco ha sufrido cambios importantes en su composición, que inciden directamente en el manejo de la operación. Su estrategia de expansión al mercado centroamericano le llevó a crear oficinas de venta en Nicaragua en el año 2000, posteriormente en Guatemala, El Salvador y Honduras, con comunicación directa y centralizada en Costa Rica.

IV. PLANTEAMIENTO DEL PROBLEMA

4.1. Caracterización del Problema

El Instituto Técnico de computación y Comercio “ITEC” es un centro de estudio que tiene como objetivo el registro de los datos personales del estudiante en sus expedientes y llevar el control de matrículas en los diferentes cursos técnicos que brinda esta institución. Actualmente el instituto no maneja el control de los factores que conllevan a los estudiantes a no concluir sus estudios en los diferentes cursos

4.2. Delimitación del Problema

El Instituto Técnico de computación y Comercio “ITEC” necesita de gestión de reportes que le permita realizar predicciones basados en cuadros estadísticos sobre la cantidad de estudiantes que se retiran en algún tiempo determinado y quienes finalizan los cursos.

4.3. Formulación del Problema

Tomando en cuenta lo antes expuesto, se puede definir la siguiente interrogante:

- ¿De qué manera se puede aplicar minería de datos para mejorar la administración del instituto tecnológico ITEC y ofrecer un mejor servicio a sus clientes?

4.4. Sistematización del Problema

- ¿Cuál es la situación actual en que se encuentra el Instituto Técnico de computación y Comercio “ITEC” en cuanto a su matrícula?

- ¿De qué manera se pueden determinar los resultados de análisis de datos para la toma de decisión del instituto en el área de matrículas.

- ¿Cómo se podría mejorar el control de datos de los estudiantes en el área de matrículas?

- ¿Cuál sería la norma ISO correcta para evaluar aplicación de Minería de Datos?

V. JUSTIFICACIÓN

El Instituto Técnico de computación y Comercio “ITEC” actualmente cuenta con un sistema informático de registro de matrículas y calificaciones que facilita el control de estudiantes matriculados en los cursos que ofrece el instituto, por lo tanto se consideró la necesidad de crear un modelo que permitirá realizar predicciones de los estudiantes que se retiran de los cursos.

Entre los beneficios que proporcionará la aplicación a implantar serían:

- ✓ Facilitar la toma de decisiones en el área de publicidad de los cursos que tienen más demandas
- ✓ Permitir al instituto ser competitivo en forma externa
- ✓ Permitir crear estrategias para un mejor manejo de los cursos
- ✓ Disminuir el exceso de trabajo en equipo
- ✓ Proporcionar un acceso rápido a la información y por ende mejora en la atención de los usuarios
- ✓ Ayudar a mejorar la efectividad de las operaciones en el instituto

VI. OBJETIVOS

OBJETIVO GENERAL

- Desarrollar un modelo basado en minería de datos aplicando el **Algoritmo de Decisión Trees** en el área de **matrículas** del Instituto de Computación y Comercio (ITEC) en Managua en el II Semestre del 2015.

OBJETIVOS ESPECÍFICOS

- Analizar la situación actual en que se encuentra el Instituto técnico de computación y comercio “ITEC” en cuanto a su matrícula.
- Emplear técnicas de minería de datos utilizando el algoritmo de Árboles de Decisión para la obtención de información.
- Mostrar reportes para la ayuda en la toma de decisiones relacionados en el departamento de publicidad y área de matrícula.
- Evaluar el modelo según la norma ISO-9126-1 “Las característica de eficiencia”

VII. MARCO TEÓRICO

En la actualidad los sistemas informáticos son muy importantes ya que gracias a ellos las instituciones manejan la información de manera cómoda y sencilla; sin necesidad de archivar los registros en papeles u otros documentos que no van a permitir obtener información rápida de los registros diarios que se necesita en su debido momento.

Minería de Datos comenzó teniendo mucho éxito en los estudios de mercado. Junto con ella se comenzó la introducción de procesos inductivos basados en los árboles de decisión desarrollados en la Teoría de Decisión

La minería de datos es un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos.

La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining. (López & González, 2007)

Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos. (López & González, 2007)

7.1. Datos

Dato es un documento**,** una información o un testimonio que permite llegar al conocimiento de algo o deducir las consecuencias legítimas de un hecho según (Krall, 2012)

Son el conjunto básico de hechos referentes a una persona, cosa o transacción de interés para distintos objetivos, entre los cuales se encuentra la toma de decisiones. Desde el punto de vista de la computación, los datos se representan como pulsaciones o pulsos electrónicos a través de la combinación de circuitos (denominados señal digital).

Se utiliza en la toma de decisiones o en la realización de cálculos a partir de un procesamiento adecuado y teniendo en cuenta su contexto. Por lo general, el dato es una representación simbólica o un atributo de una entidad.

Este es un término que se utiliza para referirse al conjunto de información que sobre una determinada materia se ha conseguido recopilar y que puede ser utilizado por varias personas. (Krall, 2012)

7.1.1. Datos para la informática.

Para la informática**, según** (Kendall & Kendall, 2011) los datos son expresiones generales que describen características de las entidades sobre las que operan los algoritmos. Estas expresiones deben presentarse de una cierta manera para que puedan ser tratadas por una computadora. En este caso, los datos por sí solos tampoco constituyen información, sino que ésta surge del adecuado procesamiento de los datos.

7.2. Información

Consiste en un conjunto de datos que poseen un significado, de modo tal que reducen la incertidumbre y aumentan el conocimiento de quien se acerca a contemplarlos. Estos datos se encuentran disponibles para su uso inmediato y sirven para clarificar incertidumbres sobre determinados temas. (Kendall & Kendall, 2011)

Esos datos y conocimientos están estrictamente ligados con mejorar nuestra toma de decisiones.

La información está constituida por un grupo de datos ya supervisados y ordenados, que sirven para construir un mensaje basado en un cierto fenómeno o ente. La información permite resolver problemas y tomar decisiones, ya que su aprovechamiento racional es la base del conocimiento.

Por lo tanto, otra perspectiva nos indica que la información es un recurso que otorga significado o sentido a la realidad, ya que mediante códigos y conjuntos de datos, da origen a los modelos de pensamiento humano.

Los especialistas afirman que existe un vínculo indisoluble entre la información, los datos, el conocimiento, el pensamiento y el lenguaje.

Existen diversos enfoques para el estudio de la información:

- **En biología**, la información se considera como estímulo sensorial que afecta al comportamiento de los individuos
- **En computación y teoría de la información**, como una medida de la complejidad de un conjunto de datos.
- **En comunicación social y periodismo**, como un conjunto de mensajes intercambiados por individuos de una sociedad con fines organizativos concretos.

Desde el punto de vista de la **ciencia de la computación**, la información es un conocimiento explícito extraído por seres vivos o sistemas expertos como resultado de interacción con el entorno o percepciones sensibles del mismo entorno. En principio la información, a diferencia de los datos o las percepciones sensibles, tienen estructura útil que modificará las sucesivas interacciones del que posee dicha información con su entorno. (Pressman, 2005)

7.3. Conocimiento

El conocimiento es un conjunto de información almacenada mediante la experiencia o el aprendizaje, o a través de la introspección. En el sentido más amplio del término, se trata de la posesión de múltiples datos interrelacionados que, al ser tomados por sí solos, poseen un menor valor cualitativo. (Ramirez, 2010)

El conocimiento tiene su origen en la percepción sensorial, después llega al entendimiento y concluye finalmente en la razón. Se dice que el conocimiento es una relación entre un sujeto y un objeto. El proceso del conocimiento involucra cuatro elementos: sujeto**,** objeto**,** operación y representación interna.

La ciencia considera que, para alcanzar el conocimiento, es necesario seguir un método. El conocimiento científico no sólo debe ser válido y consistente desde el punto de vista lógico, sino que también debe ser probado mediante el método científico o experimental**.** (Ramirez, 2010)

La forma sistemática de generar conocimiento tiene dos etapas: la investigación básica, donde se avanza en la teoría; y la investigación aplicada, donde se aplica la información.

Cuando el conocimiento puede ser transmitido de un sujeto a otro mediante una comunicación formal, se habla de conocimiento explícito. En cambio, si el conocimiento es difícil de comunicar y se relaciona a experiencias personales o modelos mentales, se trata de conocimiento implícito. (Ramirez, 2010)

7.4. Sistemas de información

Es un conjunto de **componentes que interaccionan** entre sí para alcanzar un fin determinado, el cual es satisfacer las necesidades de información de dicha organización. Estos componentes pueden ser personas, datos, actividades o recursos materiales en general, los cuales procesan la información y la distribuyen de manera adecuada, buscando satisfacer las necesidades de la organización. (Kendall & Kendall, 2011)

El objetivo primordial de un sistema de información es apoyar la toma de decisiones y controlar todo lo que en ella ocurre. Es importante señalar que existen dos tipos de sistema de información, los formales y los informales; los primeros utilizan como medio para llevarse a cabo estructuras sólidas como ordenadores.

Un Sistema de Información realiza cuatro actividades básicas:

- **Entrada de información**: proceso en el cual el sistema toma los datos que requiere.
- **Almacenamiento de información**: pude hacerse por computadora o archivos físicos para conservar la información.
- **Procesamiento de la información**: permite la transformación de los datos fuente en información que puede ser utilizada para la toma de decisiones
- **Salida de información**: es la capacidad del sistema para producir la información procesada o sacar los datos de entrada al exterior.

7.4.1 Ciclo de desarrollo de un sistema de información

(Kendall & Kendall, 2011) define las etapas de desarrollo las cuales están constituidas por las siguientes fases:

- **Definición de Proyecto**
- **Análisis del Contexto**
- **Definición de Requerimientos**

- **Diseño del Sistema**
- **Construcción del Sistema**
- **Pruebas del Sistema**
- **Implantación del Sistema**

7.5. Bases de datos relacionales

7.5.1 Base de Datos:

Se conoce como base de datos (o database, de acuerdo al término inglés) al conjunto de los datos que pertenecen a un mismo contexto y que son almacenados de manera sistemática para que puedan utilizarse en el futuro. Estas bases de datos pueden ser estáticas (cuando los datos almacenados no varían pese al paso del tiempo) o dinámicas (los datos se modifican con el tiempo; estas bases, por lo tanto, requieren de actualizaciones periódicas). (Pressman, 2005)

Es un sistema informático a modo de almacén. En este almacén se guardan grandes volúmenes de información. La antigua gestión de datos se basaba en archivos informáticos, pero para las necesidades de hoy en día hacen falta sistemas más perfeccionados que son precisamente lo que se denomina sistema de base de datos. (Silberschatz, 2007)

7.5.3 Base de datos relacionales

En una computadora existen diferentes formas de almacenar información. Esto da lugar a distintos modelos de organización de la base de datos: jerárquico, red, relacional y orientada a objeto.

Los sistemas relacionales son importantes porque ofrecen muchos tipos de procesos de datos, como: simplicidad y generalidad, facilidad de uso para el usuario final, períodos cortos de aprendizaje y las consultas de información se especifican de forma sencilla. (Silberschatz, 2007)

Las tablas son un medio de representar la información de una forma más compacta y es posible acceder a la información contenida en dos o más tablas. Más adelante explicaremos que son las tablas.

Las bases de datos relacionales están constituidas por una o más tablas que contienen la información ordenada de una forma organizada. Cumplen las siguientes leyes básicas:

- Generalmente, contendrán muchas tablas.
- Una tabla sólo contiene un número fijo de campos.
- El nombre de los campos de una tabla es distinto.
- Cada registro de la tabla es único.
- El orden de los registros y de los campos no está determinados.
- Para cada campo existe un conjunto de valores posible.

7.6. Minería de Datos

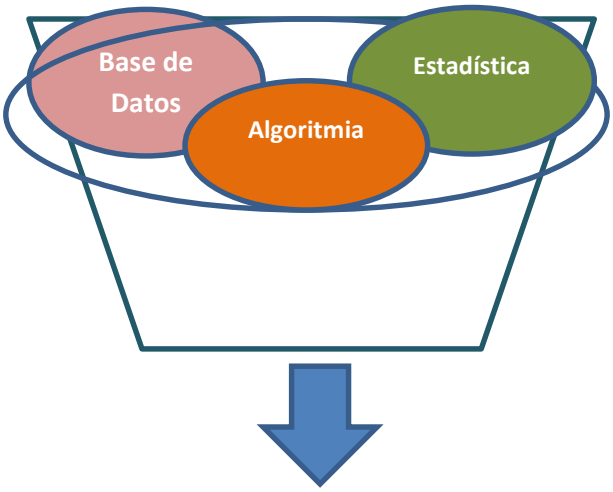


Figura 1: **Minería de Datos**

La minería de datos es el proceso de detectar la información acciónales de grandes conjuntos de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos. (López & González, 2007)

Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos. Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

- **Previsión**: calcular las ventas y predecir las cargas de servidor o el tiempo de inactividad del servidor.

- **Riesgo y probabilidad**: elegir los mejores clientes para la distribución de correo directo, determinar el punto de equilibrio probable para los escenarios de riesgo, y asignar probabilidades a diagnósticos u otros resultados.
- **Recomendaciones**: determinar los productos que se pueden vender juntos y generar recomendaciones.

- **Buscar secuencias**: analizar los artículos que los clientes han introducido en el carrito de compra y predecir los posibles eventos.

- **Agrupación**: separar los clientes o los eventos en clústeres de elementos relacionados, y analizar y predecir afinidades.

7.6.1 Las bases de datos y la minería de datos:

Las bases de datos han sido sin duda una herramienta fundamental que ha permitido la evolución de la ciencia de la minería de datos. De hecho, a veces se usa el término "KDD (Knowledge Discovery in Databases o Descubrimiento de Conocimiento en Bases de Datos) como sinónimo de minería de datos. Las bases de datos puede decirse que son una de las tres patas en que se apoya la minería de datos, y que son: (López & González, 2007)

a) Bases de datos
b) Estadística
c) Algoritmia

Una aplicación curiosa de la minería de datos es obtener imágenes representativas para realizar el análisis de datos. Esto permite mostrar lo que ocurre con miles de datos de forma gráfica.

La minería de datos es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación). Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales.

7.6.2 Las fases de la minería de datos:

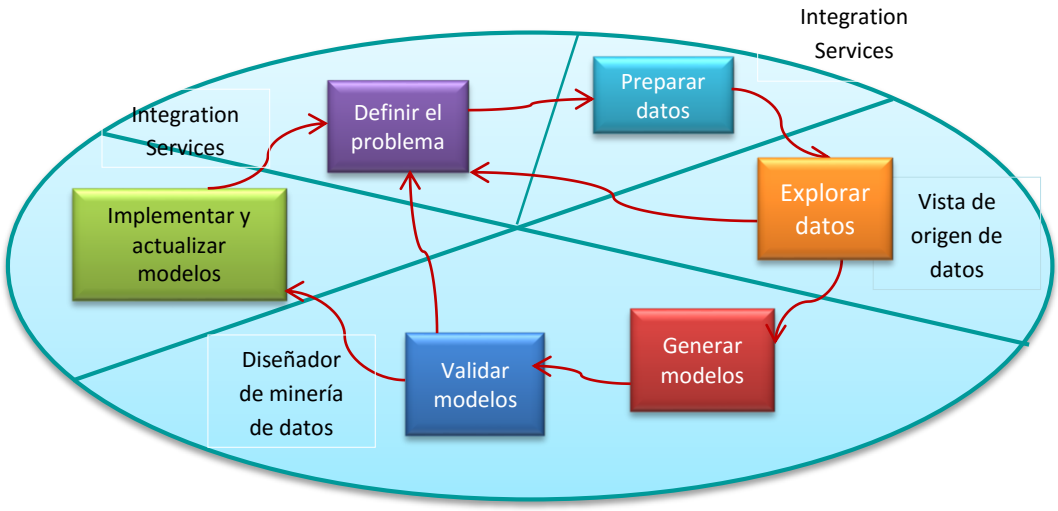


Figura 2: **Fases de Minería de Datos**

- **Definir el problema**

El primer paso del proceso de minería de datos consiste en definir claramente el problema y considerar formas de usar los datos para proporcionar una respuesta para el mismo. (López & González, 2007)

Este paso incluye analizar los requisitos empresariales, definir el ámbito del problema, definir las métricas por las que se evaluará el modelo y definir los objetivos concretos del proyecto de minería de datos. Estas tareas se traducen en preguntas como las siguientes:

- ¿Qué está buscando? ¿Qué tipos de relaciones intenta buscar?
- ¿Refleja el problema que está intentando resolver las directivas o procesos de la empresa?
- ¿Desea realizar predicciones a partir del modelo de minería de datos o solamente buscar asociaciones y patrones interesantes?
- ¿Qué resultado o atributo desea predecir?

Para responder a estas preguntas, puede que deba dirigir un estudio de disponibilidad de datos para investigar las necesidades de los usuarios de la

empresa con respecto a los datos disponibles. Si los datos no abarcan las necesidades de los usuarios, podría tener que volver a definir el proyecto. (López & González, 2007)

- **Preparar los datos**

El segundo paso del proceso de minería de datos consiste en consolidar y limpiar los datos identificados en el paso Definir el problema.

Los datos pueden estar dispersos en la empresa y almacenados en formatos distintos; también pueden contener incoherencias como entradas que faltan o incorrectas. (Ramirez, 2010)

La limpieza de datos no solamente implica quitar los datos no válidos o interpolar valores que faltan, sino también buscar las correlaciones ocultas en los datos, identificar los orígenes de datos que son más precisos y determinar qué columnas son las más adecuadas para el análisis. Por ejemplo, ¿debería utilizar la fecha de envío o la fecha de pedido? ¿Qué influye más en las ventas: la cantidad, el precio total o un precio con descuento? Los datos incompletos, los datos incorrectos y las entradas que parecen independientes, pero que de hecho están estrechamente correlacionadas, pueden influir en los resultados del modelo de maneras que no espera. (Ramirez, 2010)

- **Explorar los datos**

El tercer paso del proceso de minería de datos consiste en explorar los datos preparados. Debe conocer los datos para tomar las decisiones adecuadas al crear los modelos de minería de datos. (Orallo, Ramírez, & Ferrí, 2004)

Entre las técnicas de exploración se incluyen calcular los valores mínimos y máximos, calcular la media y las desviaciones estándar, y examinar la distribución de los datos.

Las desviaciones estándar y otros valores de distribución pueden proporcionar información útil sobre la estabilidad y exactitud de los resultados. Una desviación estándar grande puede indicar que agregar más datos podría ayudarle a mejorar el modelo.

Los datos que se desvían mucho de una distribución estándar se podrían sesgar o podrían representar una imagen precisa de un problema de la vida real, pero dificultan el ajustar un modelo a los datos. (López & González, 2007)

Al explorar los datos para conocer el problema empresarial, puede decidir si el conjunto de datos contiene datos defectuosos y, a continuación, puede inventar una estrategia para corregir los problemas u obtener una descripción más profunda de los comportamientos que son típicos de un negocio.

- **Generar modelos**

El cuarto paso del proceso de minería de datos, consiste en generar el modelo o modelos de minería de datos. Usará los conocimientos adquiridos en el paso Explorar los datos para definir y crear los modelos. (López & González, 2007)

Deberá definir qué columnas de datos desea que se usen; para ello, creará una estructura de minería de datos. La estructura de minería de datos se vincula al origen de datos, pero en realidad no contiene ningún dato hasta que se procesa. Al procesar la estructura de minería de datos, Analysis Services genera agregados y otra información estadística que se puede usar para el análisis.

Cualquier modelo de minería de datos que esté basado en la estructura puede utilizar esta información. Antes de procesar la estructura y el modelo, un modelo de minería de datos simplemente es un contenedor que especifica las columnas que se usan para la entrada, el atributo que está prediciendo y parámetros que indican al algoritmo cómo procesar los datos.

También puede utilizar los parámetros para ajustar cada algoritmo y puede aplicar filtros a los datos de entrenamiento para utilizar un subconjunto de los datos, creando resultados diferentes. Después de pasar los datos a través del modelo, el objeto de modelo de minería de datos contiene los resúmenes y modelos que se pueden consultar o utilizar para la predicción. (López & González, 2007)

Puede definir un modelo nuevo mediante el Asistente para minería de datos de SQL Server Data Tools o con el lenguaje DMX (Extensiones de minería de datos)

- **Explorar y validar los modelos**

El quinto paso del proceso de minería de datos consiste en explorar los modelos de minería de datos que ha generado y comprobar su eficacia.

Antes de implementar un modelo en un entorno de producción, es aconsejable probar si funciona correctamente. Además, al generar un modelo, normalmente se crean varios con configuraciones diferentes y se prueban todos para ver cuál ofrece los resultados mejores para su problema y sus datos.

También puede comprobar si los modelos crean predicciones correctamente mediante herramientas del diseñador como el gráfico de mejora respecto al modelo predictivo y la matriz de clasificación.

Para comprobar si el modelo es específico de sus datos o se puede utilizar para realizar inferencias en la población general, puede utilizar la técnica estadística

denominada validación cruzada para crear automáticamente subconjuntos de los datos y probar el modelo con cada uno.

- **Implementar y actualizar los modelos**

Una vez que los modelos de minería de datos se encuentran en el entorno de producción, puede llevar acabo diferentes tareas, dependiendo de sus necesidades. Las siguientes son algunas de las tareas que puede realizar:

- Use los modelos para crear predicciones que luego podrá usar para tomar decisiones comerciales.

- Crear consultas de contenido para recuperar estadísticas, reglas o fórmulas del modelo.

- Incrustar la funcionalidad de minería de datos directamente en una aplicación. Puede incluir Objetos de administración de análisis, que contiene un conjunto de objetos que la aplicación pueda utilizar para crear, cambiar, procesar y eliminar estructuras y modelos de minería de datos. También puede enviar mensajes XML directamente a una instancia de Analysis Services. (Orallo, Ramírez, & Ferrí, 2004)

7.7. Técnicas 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, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados. (López & González, 2007)

Las técnicas más representativas son:

1. **Redes neuronales**

2. **Regresión lineal**
3. **Modelos estadísticos**
4. **Agrupamiento o Clustering**
5. **Reglas de asociación**
6. **Árboles de decisión**

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Ejemplos:

- Algoritmo ID3.
- Algoritmo C4.5.

7.8. Algoritmos de minería de datos:

Un algoritmo de minería de datos es un conjunto de cálculos y reglas heurísticas que permite crear un modelo de minería de datos a partir de los datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos. (Orallo, Ramírez, & Ferrí, 2004)

A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

El modelo de minería de datos que crea un algoritmo a partir de los datos puede tomar diversas formas, incluyendo:

- Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.

- Un árbol de decisión que predice un resultado y que describe cómo afectan a este los distintos criterios.
- Un modelo matemático que predice las ventas.
- Un conjunto de reglas que describen cómo se agrupan los productos en una transacción, y las probabilidades de que dichos productos se adquieran juntos.

El algoritmo que se utilizó en nuestra aplicación es:

ÁRBOLES 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.

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. (Orallo, Ramírez, & Ferrí, 2004)

Está técnica se encuentra dentro de una metodología de aprendizaje supervisado. Su representación es en forma de árbol en donde cada nodo es una decisión, los cuales a su vez generan reglas para la clasificación de un conjunto de datos.

Los árboles de decisión son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación.

Se trata de la aplicación del conocido procedimiento del “divide y vencerás”. Sobre los datos, se van realizando sucesivas bifurcaciones hasta llegar a un resultado. Sigue unas pautas lógicas, por lo que se dice que es una “caja blanca”, o proceso comprensible por el ser humano.

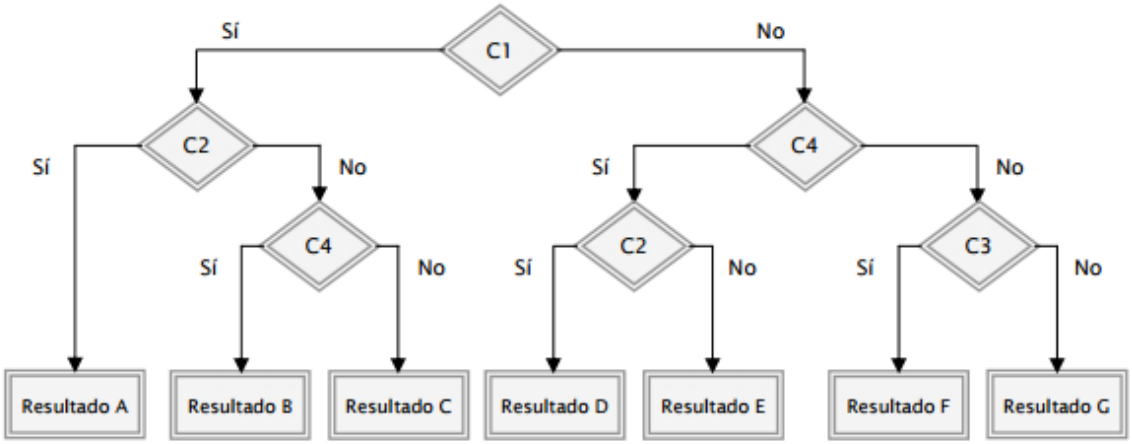


Figura 3: **Árboles de Decisión**

Cómo funciona el algoritmo

El algoritmo de árboles de decisión de Microsoft genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

El algoritmo de árboles de decisión de Microsoft utiliza la selección de características para guiar la selección de los atributos más útiles. Todos los algoritmos de minería de datos de Analysis Services utilizan la selección de características para mejorar el rendimiento y la calidad del análisis. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Si utiliza demasiados atributos de predicción o de entrada al diseñar

un modelo de minería de datos, el modelo puede tardar mucho tiempo en procesarse o incluso quedarse sin memoria. Entre los métodos que se usan para determinar si hay que dividir el árbol figuran métricas estándar del sector para la entropía y las redes Bayesianas. Para obtener más información sobre los métodos que se usan para seleccionar los atributos significativos y, a continuación, puntuarlos y clasificarlos. (Orallo, Ramírez, & Ferrí, 2004)

Un problema común de los modelos de minería de datos es que el modelo se vuelve demasiado sensible a las diferencias pequeñas en los datos de entrenamiento, en cuyo caso se dice que está sobre ajustado o sobreentrenado. Un modelo sobre ajustado no se puede generalizar a otros conjuntos de datos. Para evitar sobre ajustar un conjunto de datos determinado, el algoritmo de árboles de decisión de Microsoft utiliza técnicas para controlar el crecimiento del árbol. Para obtener una explicación más detallada de cómo funciona el algoritmo de árboles de decisión de Microsoft.

Ejemplo:

El departamento de marketing de la empresa Adventure Works Cycles desea identificar las características de los clientes antiguos que podrían indicar si es probable que realicen alguna compra en el futuro. La base de datos AdventureWorks2012 almacena información demográfica que describe a los clientes antiguos. Mediante el algoritmo de árboles de decisión que analiza esta información, el departamento puede generar un modelo que predice si un determinado cliente va a comprar productos, basándose en el estado de las columnas conocidas sobre ese cliente, como la demografía o los patrones de compra anteriores. (Orallo, Ramírez, & Ferrí, 2004)

Predecir columnas discretas

La forma en que el algoritmo de árboles de decisión de Microsoft genera un árbol para una columna de predicción discreta puede mostrarse mediante un

histograma. El siguiente diagrama muestra un histograma que traza una columna de predicción, Bike Buyers, según una columna de entrada, Age. El histograma muestra que la edad de una persona ayuda a distinguir si esa persona comprará una bicicleta.

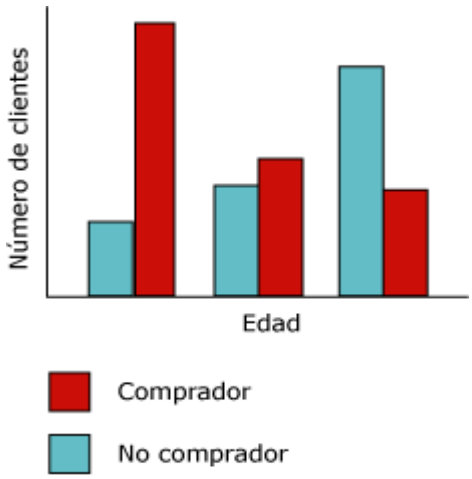


Figura 4: **Histograma1**

La correlación que aparece en el diagrama hará que el algoritmo de árboles de decisión de Microsoft cree un nuevo nodo en el modelo.

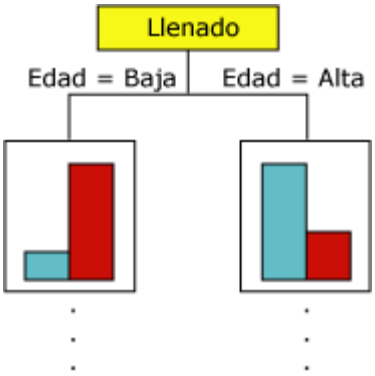


Figura 5: **Histograma2**

A medida que el algoritmo agrega nuevos nodos a un modelo, se forma una estructura en árbol. El nodo superior del árbol describe el desglose de la columna

de predicción para la población global de clientes. A medida que el modelo crece, el algoritmo considera todas las columnas.

Predecir columnas continuas

Cuando el algoritmo de árboles de decisión de Microsoft genera un árbol basándose en una columna de predicción continua, cada nodo contiene una fórmula de regresión. Se produce una división en un punto de no linealidad de la fórmula de regresión. Por ejemplo, considere el siguiente diagrama.

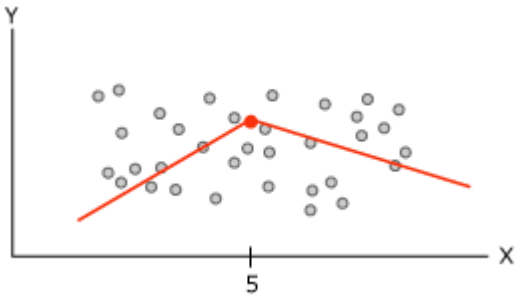


Figura 6: **Formula de Regresión**

El diagrama contiene los datos que pueden modelarse utilizando una sola línea o dos líneas conectadas. Sin embargo, una sola línea realizará un pobre trabajo en la representación de los datos. En su lugar, si se usan dos líneas, el modelo hará un mejor trabajo en la aproximación a los datos. El punto donde las dos líneas se unen es el punto de no linealidad y donde se dividiría un nodo de un modelo de árbol de decisión.

Por ejemplo, el nodo que corresponde al punto de no linealidad del gráfico anterior podría representarse mediante el siguiente diagrama. Las dos ecuaciones representan las ecuaciones de regresión de las dos líneas.

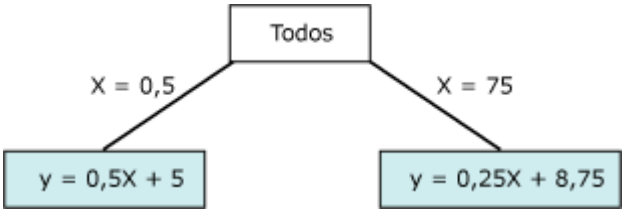


Figura 7: **Ecuación de Regresión**

Datos requeridos para los modelos de árboles de decisión

Cuando prepare los datos para su uso en un modelo de árboles de decisión, conviene que comprenda qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos se utilizan.

Los requisitos para un modelo de árboles de decisión son los siguientes:

- **Una columna key:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- **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.

El árbol es una excelente ayuda para la elección entre varios cursos de acción. Provee una estructura sumamente efectiva dentro de la cual se puede estimar, cuáles son las opciones e investigar las posibles consecuencias de seleccionar cada una de ellas.

También ayuda a construir una imagen balanceada de los riesgos y recompensas asociados con cada posible curso de acción.

Todos los arboles de decisión requieren las siguientes cuatro componentes:

- Alternativas de decisión en cada punto de decisión.
- Eventos que pueden ocurrir como resultado de cada alternativa de decisión.
- Probabilidad de que ocurran eventos posibles como resultado de las decisiones.
- Resultados de las posibles interacciones entre las alternativas de decisión y los eventos. (Casi siempre expresados en términos económicos).

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Nos ayudan a tomar la decisión más “acertada”, desde un punto de vista probabilístico

7.9. Modelo que aplican para la elaboración de la aplicación

¿Qué es un modelo de minería de datos?

La minería de datos se aplica a todo tipo de datos imaginable: desde datos numéricos a imágenes de satélite, mamografías, música, archivos de ordenador, imágenes, etc. Podemos decir que “cualquier cosa” constituye un dato. Por tanto la minería de datos tiene infinitas aplicaciones: comerciales, marketing, industria, internet, agricultura, etc. Con miles de datos, necesitamos limpiarlos (eliminar fragmentos inútiles, repetidos, etc.) y organizarlos, y una vez realizado este proceso decimos que tenemos “Información”. (Orallo, Ramírez, & Ferrí, 2004)

La información hay que tratarla con un modelo para obtener resultados o conclusiones a los que llamamos “Conocimiento”. Es decir, el conocimiento es información analizada. Para este análisis hay diferentes modelos de minería de datos. Digamos que un modelo es una forma de aplicar un tratamiento a una cantidad masiva de datos para extraer información de ellos.

¿Cómo escoger un modelo de minería de datos?

No hay un modelo óptimo de tratamiento de datos. Por tanto, el modelo a elegir depende de las circunstancias y necesidades. Factores a tener en cuenta son la efectividad del modelo para dar resultados de calidad, y el si resulta necesario o no que sea comprensible para el ser humano.

En el caso de escoger una red neuronal, las operaciones que se aplican a los datos hay que determinarlas. ¿Cómo se hace esto? Digamos que “entrenando” a la red neuronal (a esto se le llama aprendizaje automático) a través de algoritmos de optimización de forma que dados unos datos de entrada, vamos informando al sistema de si el resultado es más o menos bueno. En sucesivas iteraciones, el sistema puede alcanzar un grado de perfeccionamiento adecuado para su explotación comercial. (López & González, 2007)

7.10. Herramientas de Minería de Datos

Microsoft SQL Server Analysis Services proporciona las siguientes herramientas que puede utilizar para crear soluciones de minería de datos:

- **El Asistente para minería de datos de SQL Server Data Tools (SSDT**) facilita la creación de estructuras y de modelos de minería de datos, usando orígenes de datos relacionales o datos multidimensionales en cubos.

  En el asistente, elija los datos que desee utilizar y, a continuación, aplique técnicas de minería de datos específicas, como agrupación en clústeres, redes neurales o modelado de series temporales. (Krall, 2012)

- **SQL Server Management Studio y SQL Server Data Tools (SSDT)** disponen de visores de modelos para explorar los modelos de minería de datos una vez creados. Puede examinar los modelos mediante visores adaptados a cada algoritmo o analizar con mayor profundidad utilizando el visor de contenido del modelo. (Silberschatz, 2007)

- **El Generador de Consultas de Predicción** se proporciona en SQL Server Management Studio y SQL Server Data Tools (SSDT) para ayudarle a crear consultas de predicción. También puede probar la exactitud de los modelos respecto a un conjunto de datos de exclusión o datos externos, o utilizar validación cruzada para evaluar la calidad del conjunto de datos.

- **SQL Server Management Studio** es la interfaz en la que administra las soluciones de minería de datos implementadas en una instancia de Analysis Services. Puede volver a procesar las estructuras y modelos para actualizar los datos que contienen.

- **SQL Server Integration Services** contiene herramientas que puede utilizar para limpiar datos, automatizar tareas como la creación de predicciones y

actualización de modelos y para crear soluciones de minería de datos de texto.

7.10.1 SQL Server Management Studio

Después de crear e implementar los modelos de minería de datos en un servidor, puede utilizar SQL Server Management Studio para administrar la base de datos Analysis Services que hospeda los objetos de minería de datos. También puede seguir realizando tareas que utilizan el modelo, como explorar modelos, procesar nuevos datos y crear predicciones. (Silberschatz, 2007)

Management Studio también contiene editores de consultas que puede utilizar para diseñar y ejecutar consultas de extensiones de minería de datos (DMX) o trabajar con objetos de minería de datos utilizando XMLA.

7.10.2 Transformaciones y tareas de minería de datos en Integration Services

SQL Server Integration Services dispone de muchos componentes compatibles con la minería de datos.

Algunas herramientas de Integration Services están diseñadas para ayudar a automatizar tareas de datos comunes, incluida la predicción, la compilación de modelos y el procesamiento. Por ejemplo:

- Crear un paquete de Integration Services que actualice automáticamente el modelo cada vez que el conjunto de datos se actualice con nuevos clientes
- Realizar una segmentación personalizada o un muestreo personalizado de los registros del caso.
- Generar automáticamente modelos pasados en parámetros.
- Sin embargo, también puede utilizar la minería de datos en un flujo de trabajo de paquetes, como una entrada a otros procesos. Por ejemplo:

- Usar valores de probabilidad generados por el modelo para ponderar las puntuaciones de la minería de texto u otras tareas de clasificación.
- Generar automáticamente predicciones basadas en datos anteriores y utilizar esos valores para evaluar la validez de nuevos datos.
- Usar la regresión logística para segmentar los clientes de entrada por riesgo.

7.11. Datawarehouse

Es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. La creación de un datawarehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence.

La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales... etc). Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales)

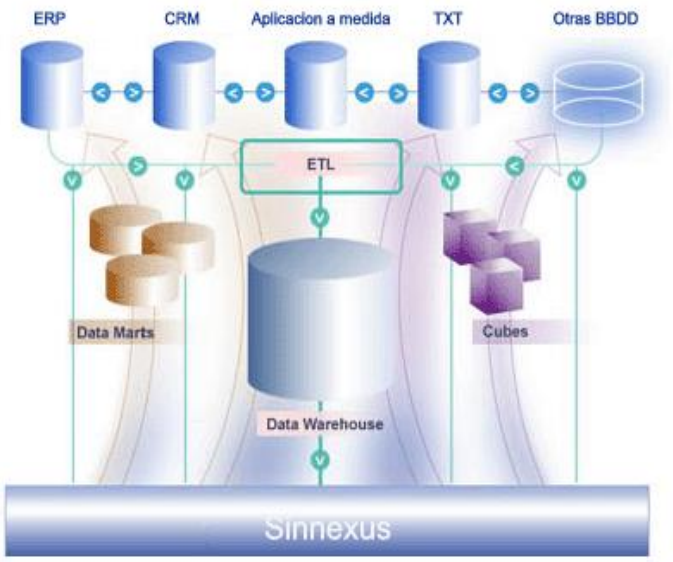


Figura 8:
Datawarehouse

El término Datawarehouse fue acuñado por primera vez por Bill Inmon, y se traduce literalmente como *almacén de datos*. No obstante, y como cabe suponer, es mucho más que eso. Según definió el propio Bill Inmon, un datawarehouse se caracteriza por ser:

- **Integrado**: los datos almacenados en el datawarehouse deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las distintas necesidades de los usuarios.
- **Temático**: sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del datawarehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.
- **Histórico**: el tiempo es parte implícita de la información contenida en un datawarehouse. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la

información almacenada en el datawarehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el datawarehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

- **No volátil**: el almacén de información de un datawarehouse existe para ser leído, pero no modificado. La información es por tanto permanente, significando la actualización del datawarehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

Otra característica del datawarehouse es que contiene metadatos, es decir, datos sobre los datos. Los metadatos permiten saber la procedencia de la información, su periodicidad de refresco, su fiabilidad, forma de cálculo.

Los metadatos serán los que permiten simplificar y automatizar la obtención de la información desde los sistemas operacionales a los sistemas informacionales.

Los objetivos que deben cumplir los metadatos, según el colectivo al que va dirigido, son:

- Dar soporte al usuario final, ayudándole a acceder al datawarehouse con su propio lenguaje de negocio, indicando qué información hay y qué significado tiene. Ayudar a construir consultas, informes y análisis, mediante herramientas de Business Intelligence como DSS, EIS o CMI.

- Dar soporte a los responsables técnicos del datawarehouse en aspectos de auditoría, gestión de la información histórica, administración del datawarehouse, elaboración de programas de extracción de la información, especificación de las interfaces para la realimentación a los sistemas operacionales de los resultados obtenidos.

7.12. Norma ISO-IEC 9126-1 “Características de Calidad y Subcaracterísticas” (Gutierrez, 2014)

Esta parte proporciona los conceptos básicos de característica, subcaracterísticas, atributo, métrica así como también muestra un modelo de calidad con seis características, las cuales a su vez se sub-dividen en subcaracterísticas.

El modelo establece diez características, seis que son comunes a las vistas interna y externa y cuatro que son propias de la métrica en uso.

Las características y subcaracterísticas propias de este estándar que se encuentran dentro de las métricas interna y externa, las cuales usaremos para evaluar el software de CMI, las cuales detallo a continuación:

1. **Funcionalidad**: capacidad del software de proveer los servicios necesarios para cumplir con los requisitos funcionales.

Subcaracterísticas:

- **Idoneidad**.- Hace referencia a que si el software desempeña las tareas para las cuales fue desarrollado.
- **Exactitud**.- Evalúa el resultado final que obtiene el software y si tiene consistencia a lo que se espera de él.
- **Interoperabilidad**.- Consiste en revisar si el sistema puede interactuar con otro sistema independiente.
- **Seguridad**.- Verifica si el sistema puede impedir el acceso a personal no autorizado.

2. **Fiabilidad**: capacidad del software de mantener las prestaciones requeridas del sistema, durante un tiempo establecido y bajo un conjunto de condiciones definidas.

Subcaracterísticas:

- **Madurez:** Se debe verificar las fallas del sistema y si muchas de estas han sido eliminadas durante el tiempo de pruebas o uso del sistema.
- **Recuperabilidad:** Verificar si el software puede reasumir el funcionamiento y restaurar datos perdidos después de un fallo ocasional.
- **Tolerancia a fallos:** Evalúa si la aplicación desarrollada es capaz de manejar errores.

3. **Usabilidad**: esfuerzo requerido por el usuario para utilizar el producto satisfactoriamente.

Subcaracterísticas:

- **Aprendizaje**: Determina que tan fácil es para el usuario aprender a utilizar el sistema.
- **Comprensión:** Evalúa que tan fácil es para el usuario comprender el funcionamiento del sistema
- **Operatividad:** Determina si el usuario puede utilizar el sistema sin mucho esfuerzo.
- **Atractividad:** Verifica que tan atractiva se ve la interfaz de la aplicación.

4. **Eficiencia**: relación entre las prestaciones del software y los requisitos necesarios para su utilización.

Subcaracterísticas:

- **Comportamiento en el tiempo**.- Verifica la rapidez en que responde el sistema
- **Comportamiento de recursos**.- Determina si el sistema utiliza los recursos de manera eficiente

5. **Mantenibilidad**: esfuerzo necesario para adaptarse a las nuevas especificaciones y requisitos del software.

Subcaracterísticas:

- **Estabilidad:** Verifica si el sistema puede mantener su funcionamiento a pesar de realizar cambios.
- **Facilidad de análisis:** Determina si la estructura de desarrollo es funcional con el objetivo de diagnosticar fácilmente las fallas.
- **Facilidad de cambio:** Verifica si el sistema puede ser fácilmente modificado.
- **Facilidad de pruebas:** Evalúa si el sistema puede ser probado fácilmente.

6. **Portabilidad**: capacidad del software ser transferido de un entorno a otro.

Subcaracterísticas:

- **Capacidad de instalación:** Verifica si el software se puede instalar fácilmente
- **Capacidad de reemplazamiento:** Determina la facilidad con la que el software puede remplazar otro software similar.
- **Adaptabilidad:** El software se puede trasladar a otros ambientes

7.13. Reseña Histórica de la Empresa

El instituto técnico de computación y comercio (ITECC) nace en el año 2011, ofreciendo cursos de computación al público en general. Inicia operaciones con un personal de tres profesores y una cajera, además del gerente del negocio.

Al paso del tiempo dicho instituto fue creciendo, ofreciendo otros tipos de cursos como, Ingles técnico y cursos de Cajero administrativo, esto ha provocado un incremento en los ingresos de matrículas y por consiguiente los ingresos económicos han aumentado debido a la atención ofrecida en el instituto.

Como política de los propietarios, se decide registrar todas las operaciones realizadas con respecto a los estudiantes de forma de que pueda tener control sobre las operaciones del instituto. Para realizar este proceso el gerente de ITECC se auxilia en hojas de cálculos de Microsoft Excel.

Mientras el negocio progresa, se incrementa el número de matrículas, aumenta la demanda de cursos y crece la competencia en el mercado, por lo cual la gerencia considera implementar un software que sea capaz de realizar las actividades de matrícula e ingreso de los pagos de correspondientes así como los ingresos de calificaciones obtenidas por los estudiantes, todo esto de una forma sencilla.

Actualmente nuestro grupo de estudiantes pertenecientes a la carrera de licenciatura en ciencias de la computación de la Universidad nacional autónoma de Nicaragua (UNAN-MANAGUA), hemos decidido desarrollar un modelo de aplicación según los requerimientos del instituto, que brinde las herramientas necesarias para generar un mejor control de los procesos operativos que lleva acabo el instituto técnico de computación y comercio “ITECC”.

VIII. HIPÓTESIS

La implantación de un modelo de Minería de Datos permitirá mejorar la toma de decisiones con respecto al proceso de matrícula en el Instituto de Computación y Comercio (ITEC) brindando una mayor demanda en los cursos que se ofrecen.

IX. DISEÑO METODOLOGICO

9.1 TIPOS DE ESTUDIO

De acuerdo (Piura López, 2010), según el diseño metodológico el tipo de estudio es **analítico**, de acuerdo a (Sampieri & Fernandez, 2010) según el tiempo de ocurrencia de los hechos y registro de la Información el estudio es **prospectivo** porque se van analizando los datos del instituto según se van generando y según el período y secuencia del estudio es **transversal** por que el estudio se realizó en un tiempo determinado el cual es en el II Semestre del 2015.

El tipo de investigación según criterios de generación y desarrollo de las tecnologías como un bien público o privado es una **Investigación Aplicada:** ya que está constituida por los trabajos llevados a cabo con la intención de desarrollar nuevos conocimientos en áreas específicas como es el Departamento de **Matrículas**, con los cuales se espera resolver problemas ya definidos.

9.2 AREA DE ESTUDIO

El área de estudio que utilizamos para desarrollar un modelo que está basado en **Minería de Datos** aplicando el algoritmo de **Árboles de Decisión** la cual se realizó en el área de matrículas del Instituto de Computación y Comercio (ITEC) en Managua en el I Semestre del 2015.

9.3 UNIVERSO

El universo que se utilizó para desarrollar el modelo de Minería de datos fue en el Instituto de Computación y Comercio (ITEC) de Rubenia en Managua

9.4 MUESTRA

La muestra que obtuvimos fue en el Departamento de Matriculas del Instituto, que consta de un personal de 3 trabajadores:

- 1 responsable de matrícula de estudiantes
- 1 cajera
- 1 auxiliar de oficina

9.5 INSTRUMENTOS DE RECOLECCIÓN DE DATOS

- Utilizamos entrevista para la recolección de datos
- Correo electrónico para recopilación de información del instituto
- Observación en el periodo de matrícula de los estudiantes

9.6 ANÁLISIS DE FACTIBILIDAD TÉCNICA

La factibilidad tecnica consistio en realizar una evaluación de la tecnología existente para la elaboración del modelo de Minería de Datos.

De acuerdo a las tecnologías necesarias para la implantación del modelo se evaluó bajo dos enfoques: **Software y Hardware**

1. **<u>En cuanto a Software</u>**:

En cuanto al software se cuenta con todas las aplicaciones y programas que empleamos para el desarrollo del modelo de Minería de Datos y funcionamiento del algoritmo de arboles de Decisión. Las estaciones de trabajo operan bajo ambiente Windows, el servidor requiere el sistema operativo Windows Seven Ultimate.

Cantidad	Descripcion
01	Sistema Operativo Windows Seven Ultimate
01	Herramientas de escritorio office 2010
01	Sistemas administrativos
01	Microsoft Sql Server 2008
01	Microsoft Visual Studio 2010
01	Diversos Antivirus(Eset Nod 32)

Tabla 1: **Descripción del Software**

9.7 ANÁLISIS DE FACTIBILIDAD ECONÓMICA:

A continuación se presenta un estudio que dio como resultado la factibilidad económica del desarrollo del nuevo modelo de Minería de Datos.

1. En cuanto a Hardware: Costo

Dos Computadora de Escritorio	U$ 1,250.00
Laptop	U$ 700.00
Total	**U$ 1,950.00**

2. Material didáctico y muebles de trabajo Costo

➢ Dos Muebles de computadora		U$ 100.00
➢ Dos Silla Secretarial		U$ 85.00
➢ Dos Memorias USB 8 GB		U$ 25.00
➢ Lapicero, papel bond, engrapadora		U$ 10.00
➢ Impresora		U$ 49.00
➢ Scaner		U$ 50.00
➢ Dispositivos Inalabricos de Internet	(mensual)	U$ 41.00
Total		**U$ 860.00**

9.8 ANÁLISIS DE FACTIBILIDAD OPERATIVA:

La factibilidad operativa permite predecir, si se pondra en marcha el modelo propuesto, aprovechando los beneficios que ofrece, a todos los usuarios involucrados con el mismo.

Por otra parte el correcto funcionamiento del modelo en cuestión siempre estará subeditado a la capacidad de los usuarios que tendran acceso a dicho modelo.

La necesidad y deseo de un cambio en las actividades de atencion a los usuarios en el instituto, expresada por los estudiantes, administrativos, docentes y el personal involucrado, llevará a la aceptación de este modelo, que de una manera mas sencilla y amigable cubra todos sus requerimientos, expectativas y proporciona la información en forma oportuna y confiable. Basandose en las entrevistas y conversaciones sostenidas con el personal se demostro que estos no representan ninguna oposición con la implementación de la aplicación, por lo que le este es factible operacionalmente.

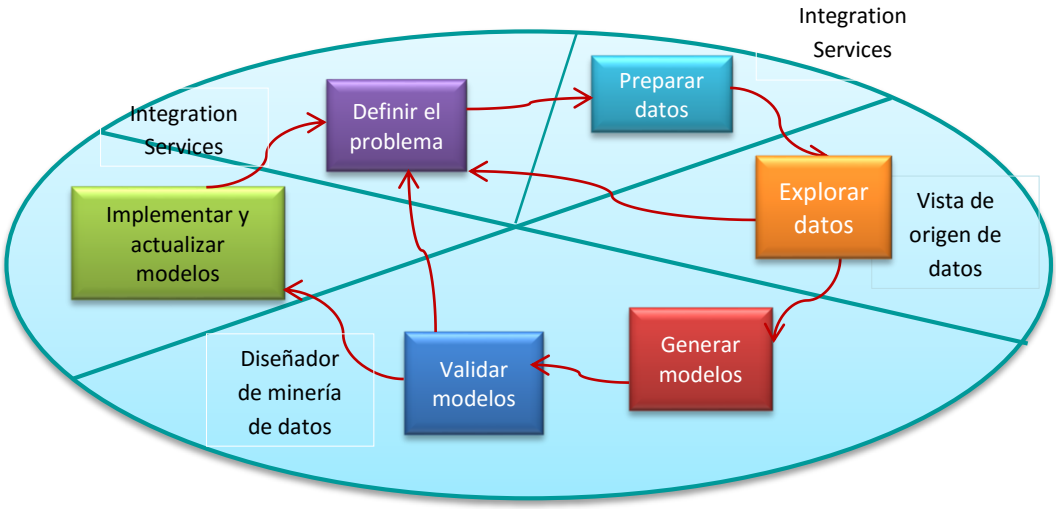
En el proceso de adiestramiento se detallan los aspectos de actualización de conocimientos y nuevas formas en el procesamiento de la información que representa el manejo del nuevo modelo en el instituto.

Con la finalidad de garantizar el buen funcionamiento del modelo y que este impactará en forma positiva a los usuarios, el mismo fue desarrollado en forma estándar, presentando una interfaz amigable al usuario, lo que se traduce en una herramienta de fácil manejo y comprensión, tanto el uso de la información veraz y consisa que se brindará. Contando con la opinión de los mismos para cualquier modificación del modelo.

9.9 HERRAMIENTAS UTILIZADAS PARA ELABORAR EL MODELO DE MINERÍA DE DATOS

1. SQL Server Management Studio 2008
2. Análisis Service
3. Integration Service SQL server 2008 y Visual Studio 2008 para limpiar datos, automatizar tareas
4. Algoritmo Decision trees
5. Reportes (Microsoft Excel)

9.10 FASES DE MINERIA DE DATOS



Ref. Figura 2 Fases de Minería de Datos

X. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

- Se desarrolló un modelo basado en minería de datos en el cual se aplicó el **Algoritmo de Decisión Trees** en el área de **matrículas** del Instituto de Computación y Comercio (ITEC) en Managua en el II Semestre del 2015, utilizando todos los requerimientos (información brindada por el Instituto) para la finalización de dicho proyecto.

- Según los datos analizados de la situación en el Instituto técnico de computación y comercio "ITEC" en cuanto a su matrícula son los siguientes:

  a. Sistema de Matricula en Sql Server 2008
  b. Base de datos
  c. Servidor Local
  d. Necesidades Encontradas:

  1. Facilitar la toma de decisiones en el área de publicidad de los cursos que tienen más demandas
  2. Permitir al instituto ser competitivo en forma externa
  3. Permitir crear estrategias para un mejor manejo de los cursos
  4. Disminuir el exceso de trabajo en equipo
  5. Proporcionar un acceso rápido a la información y por ende mejorar la atención de los usuarios
  6. Ayudar a mejorar la efectividad de las operaciones en el instituto

10.1 BASE DE DATOS

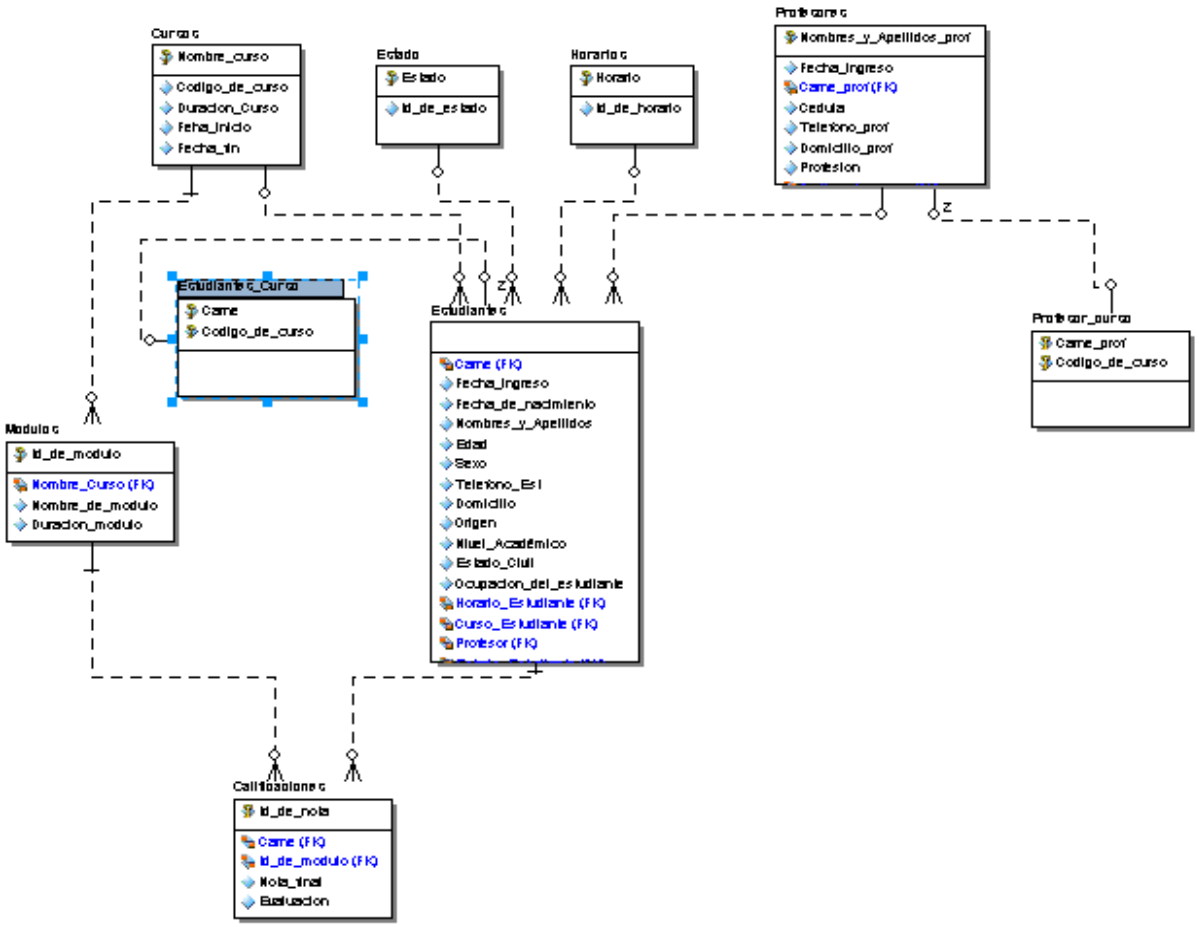


Figura 9: **Base de Datos**

- Como siguiente paso se aplicó las técnicas de minería de datos utilizando el algoritmo de Árboles de Decisión para la obtención de información.

10.2 DATAWAREHOUSE

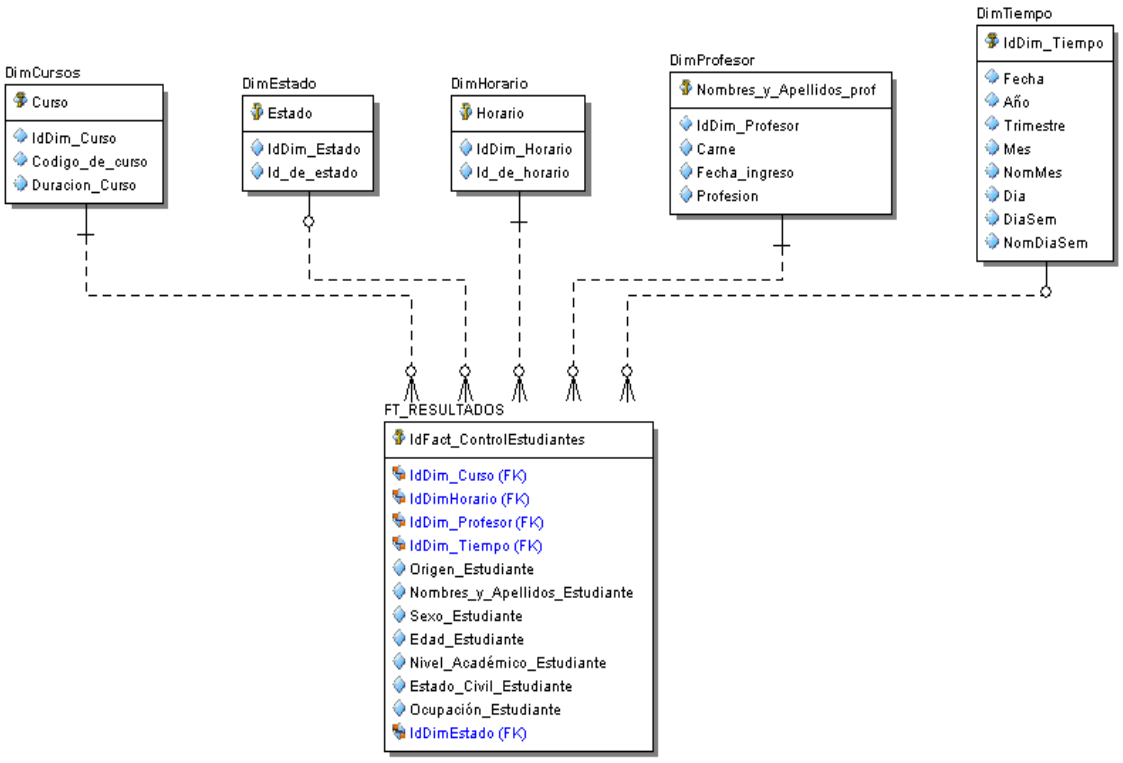


Figura 10: **Datawarahouse**

SCRIPT DE DATAWEREHOUSE

```
/*
+-------------------------------------------------------------------------
------+
|BASE DE DATOS: SISMAC_DATAWEREHOUSE_REVISADO
|
|                                                                        |
                                                |
|
|
|FECHA      : 19 /10 /2015
|
+-------------------------------------------------------------------------
------+
*/
USE [master]
GO
PRINT REPLICATE ('-', 80)
PRINT 'Eliminando la BD ...'

IF  EXISTS (SELECT name FROM sys.databases WHERE name =
N'SISMAC_DATAWEREHOUSE_REVISADO')
BEGIN
      PRINT CHAR(9) + 'Base de datos eliminada ...'
      DROP DATABASE [SISMAC_DATAWEREHOUSE_REVISADO]
END
```

```
PRINT REPLICATE ('-', 80)
PRINT 'Creando la BD ...'
CREATE DATABASE [SISMAC_DATAWEREHOUSE_REVISADO]
go

USE [SISMAC_DATAWEREHOUSE_REVISADO]
go

PRINT REPLICATE ('-', 80)
PRINT 'Tabla DimTiempo Creada ...'
PRINT REPLICATE ('-', 80)

CREATE TABLE [dbo].[DimTiempo](
	[IdDim_Tiempo] [int] NOT NULL,
	[Fecha] [date] NOT NULL,
	[Año] [smallint] NULL,
	[Trimestre] [smallint] NULL,
	[Mes] [tinyint] NULL,
	[NomMes] [varchar](50) NULL,
	[Dia] [tinyint] NULL,
	[DiaSem] [tinyint] NULL,
	[NomDiaSem] [varchar](50) NULL,
 CONSTRAINT [PK__DimTiemp] PRIMARY KEY NONCLUSTERED
(
	[IdDim_Tiempo] ASC
)WITH (PAD_INDEX  = OFF, STATISTICS_NORECOMPUTE  = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS  = ON, ALLOW_PAGE_LOCKS  = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

PRINT REPLICATE ('-', 80)
PRINT 'Tabla DimProfesor Creada ...'
PRINT REPLICATE ('-', 80)

CREATE TABLE [dbo].[DimProfesor](
	[IdDim_Profesor] [int] IDENTITY(1,1) NOT NULL,
	[Carne] [varchar](50) NOT NULL,
	[Fecha_ingreso] [date] NOT NULL,
	[Nombres_y_Apellidos_prof] [varchar](50) NOT NULL,
	[Profesion] [varchar](100) NOT NULL,
 CONSTRAINT [PK__DimProfe] PRIMARY KEY NONCLUSTERED
(
	[Nombres_y_Apellidos_prof] ASC
)WITH (PAD_INDEX  = OFF, STATISTICS_NORECOMPUTE  = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS  = ON, ALLOW_PAGE_LOCKS  = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

PRINT REPLICATE ('-', 80)
PRINT 'Tabla DimHorario Creada ...'
PRINT REPLICATE ('-', 80)

CREATE TABLE [dbo].[DimHorario](
	[IdDim_Horario] [int] IDENTITY(1,1) NOT NULL,
	[Id_de_horario] [int] NOT NULL,
	[Horario] [varchar](50) NOT NULL,
```

```
 CONSTRAINT [PK__DimHorar] PRIMARY KEY NONCLUSTERED
(
      [Horario] ASC
)WITH (PAD_INDEX  = OFF, STATISTICS_NORECOMPUTE  = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS  = ON, ALLOW_PAGE_LOCKS  = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

PRINT REPLICATE ('-', 80)
PRINT 'Tabla DimEstado Creada ...'
PRINT REPLICATE ('-', 80)

CREATE TABLE [dbo].[DimEstado](
      [IdDim_Estado] [int] IDENTITY(1,1) NOT NULL,
      [Id_de_estado] [int] NULL,
      [Estado] [varchar](50) NOT NULL,
 CONSTRAINT [PK_DimEstado] PRIMARY KEY CLUSTERED
(
      [Estado] ASC
)WITH (PAD_INDEX  = OFF, STATISTICS_NORECOMPUTE  = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS  = ON, ALLOW_PAGE_LOCKS  = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

PRINT REPLICATE ('-', 80)
PRINT 'Tabla DimCursos Creada ...'
PRINT REPLICATE ('-', 80)

CREATE TABLE [dbo].[DimCursos](
      [IdDim_Curso] [int] IDENTITY(1,1) NOT NULL,
      [Codigo_de_curso] [varchar](15) NOT NULL,
      [Curso] [varchar](50) NOT NULL,
      [Duracion_Curso] [varchar](15) NOT NULL,
 CONSTRAINT [PK__DimCurso] PRIMARY KEY NONCLUSTERED
(
      [Curso] ASC
)WITH (PAD_INDEX  = OFF, STATISTICS_NORECOMPUTE  = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS  = ON, ALLOW_PAGE_LOCKS  = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

PRINT REPLICATE ('-', 80)
PRINT 'Tabla FT_resultados Creada ...'
PRINT REPLICATE ('-', 80)

CREATE TABLE [dbo].[FT_RESULTADOS](
      [IdFact_ControlEstudiantes] [int] IDENTITY(1,1) NOT NULL,
      [IdDim_Curso] [varchar](50) NOT NULL,
      [IdDim_Profesor] [varchar](50) NOT NULL,
      [IdDim_Tiempo] [int] NULL,
      [IdDimHorario] [varchar](50) NOT NULL,
      [IdDimEstado] [varchar](50) NULL,
      [Nombres_y_Apellidos_Estudiante] [varchar](60) NULL,
      [Edad_Estudiante] [int] NULL,
      [Nivel_Académico_Estudiante] [varchar](50) NULL,
      [Estado_Civil_Estudiante] [varchar](15) NULL,
      [Ocupación_Estudiante] [varchar](50) NULL,
```

```
 CONSTRAINT [PK_FT_ControlEstudiantes] PRIMARY KEY NONCLUSTERED
(
      [IdFact_ControlEstudiantes] ASC
)WITH (PAD_INDEX  = OFF, STATISTICS_NORECOMPUTE  = OFF, IGNORE_DUP_KEY =
OFF, ALLOW_ROW_LOCKS  = ON, ALLOW_PAGE_LOCKS  = ON) ON [PRIMARY]
) ON [PRIMARY]
GO

/****** Object:  ForeignKey [FK_FT_RESULTADOS_DimCursos]  ******/
ALTER TABLE [dbo].[FT_RESULTADOS]  WITH NOCHECK ADD  CONSTRAINT
[FK_FT_RESULTADOS_DimCursos] FOREIGN KEY([IdDim_Curso])
REFERENCES [dbo].[DimCursos] ([Curso])
GO
ALTER TABLE [dbo].[FT_RESULTADOS] CHECK CONSTRAINT
[FK_FT_RESULTADOS_DimCursos]
GO
/****** Object:  ForeignKey [FK_FT_RESULTADOS_DimEstado]  ******/
ALTER TABLE [dbo].[FT_RESULTADOS]  WITH NOCHECK ADD  CONSTRAINT
[FK_FT_RESULTADOS_DimEstado] FOREIGN KEY([IdDimEstado])
REFERENCES [dbo].[DimEstado] ([Estado])
GO
ALTER TABLE [dbo].[FT_RESULTADOS] CHECK CONSTRAINT
[FK_FT_RESULTADOS_DimEstado]
GO
/****** Object:  ForeignKey [FK_FT_RESULTADOS_DimHorario]  ******/
ALTER TABLE [dbo].[FT_RESULTADOS]  WITH NOCHECK ADD  CONSTRAINT
[FK_FT_RESULTADOS_DimHorario] FOREIGN KEY([IdDimHorario])
REFERENCES [dbo].[DimHorario] ([Horario])
GO
ALTER TABLE [dbo].[FT_RESULTADOS] CHECK CONSTRAINT
[FK_FT_RESULTADOS_DimHorario]
GO
/****** Object:  ForeignKey [FK_FT_RESULTADOS_DimProfesor] ******/
ALTER TABLE [dbo].[FT_RESULTADOS]  WITH NOCHECK ADD  CONSTRAINT
[FK_FT_RESULTADOS_DimProfesor] FOREIGN KEY([IdDim_Profesor])
REFERENCES [dbo].[DimProfesor] ([Nombres_y_Apellidos_prof])
GO
ALTER TABLE [dbo].[FT_RESULTADOS] CHECK CONSTRAINT
[FK_FT_RESULTADOS_DimProfesor]
GO
/****** Object:  ForeignKey [FK_FT_RESULTADOS_FT_RESULTADOS1]  ******/
ALTER TABLE [dbo].[FT_RESULTADOS]  WITH NOCHECK ADD  CONSTRAINT
[FK_FT_RESULTADOS_FT_RESULTADOS1] FOREIGN KEY([IdDim_Tiempo])
REFERENCES [dbo].[DimTiempo] ([IdDim_Tiempo])
GO
ALTER TABLE [dbo].[FT_RESULTADOS] CHECK CONSTRAINT
[FK_FT_RESULTADOS_FT_RESULTADOS1]
GO

PRINT REPLICATE ('-', 80)
PRINT 'Base de Datos Creada ...'
PRINT REPLICATE ('-', 80)
```

10.3 INTEGRATION SERVICE

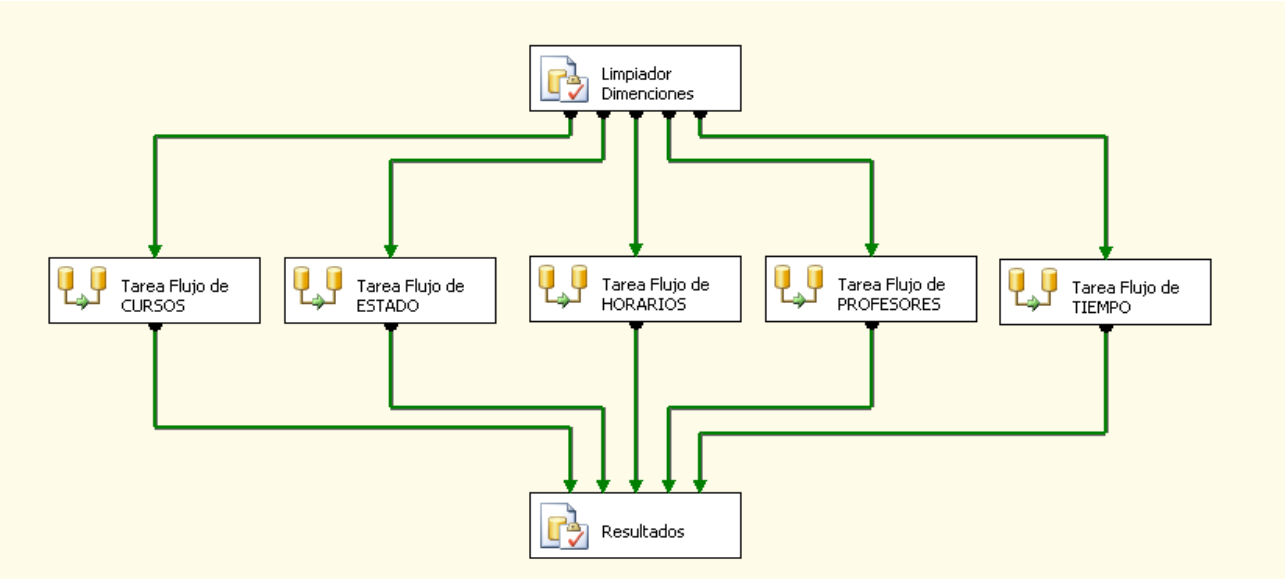
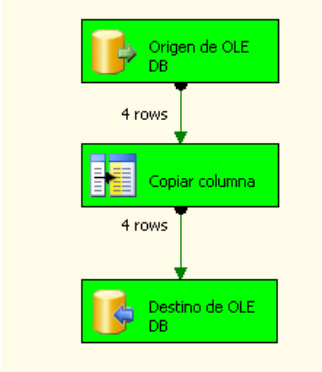
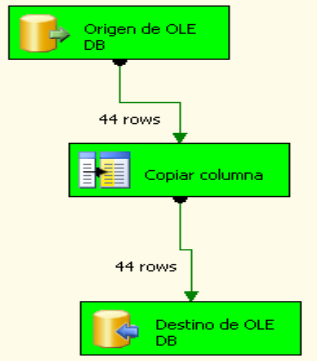


Figura 11: **Integration Service**

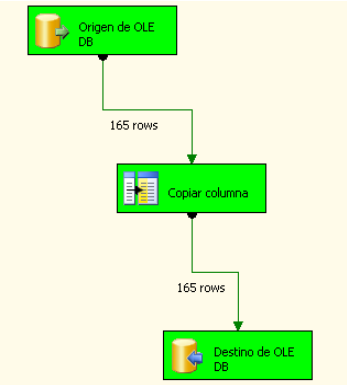
Dimensión Cursos



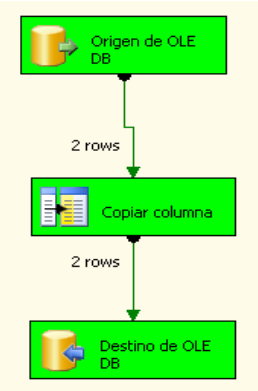
Dimensión Tiempo



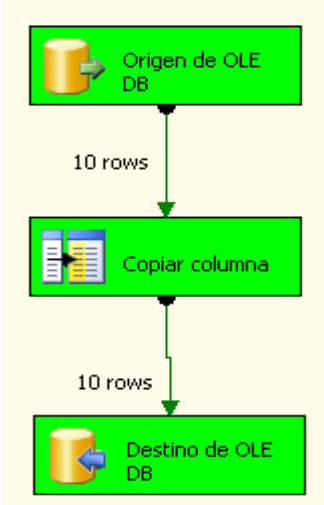
Dimensión Estudiantes



Dimensión Estado



Dimensión Horarios



Dimensión Profesor

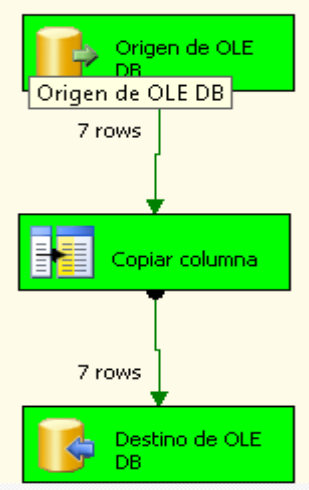
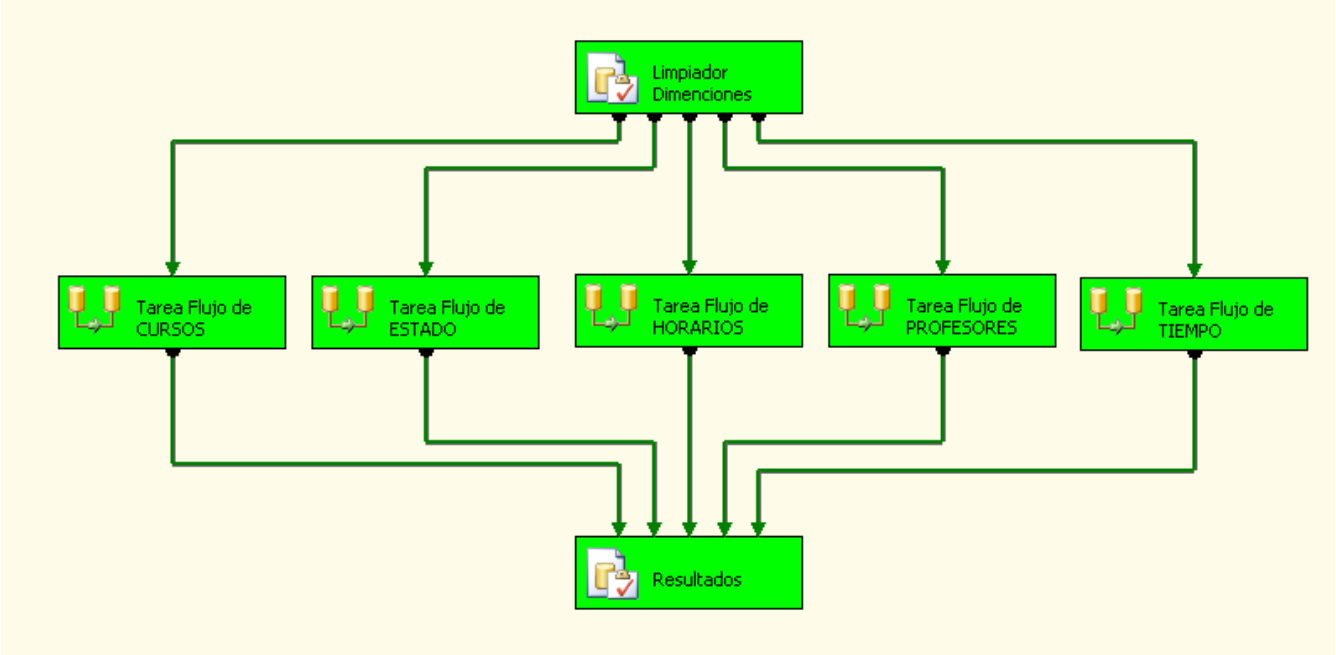


Figura 12: **Dimensiones**

Servicio de integración SQL server 2008 y Visual Estudio 2008



10.4 ANÁLISIS SERVICE

Modelo DimEstudiantes

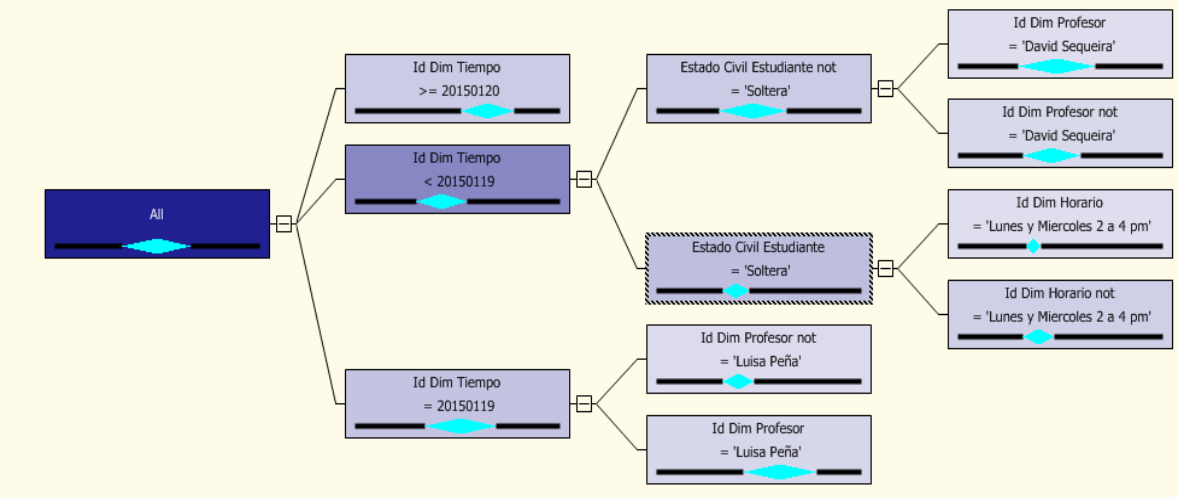
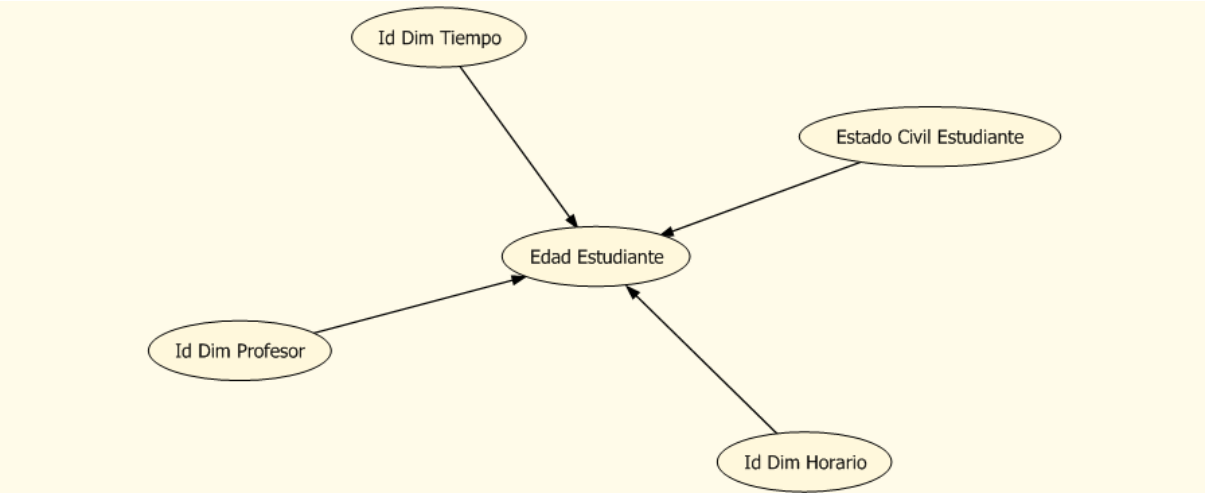
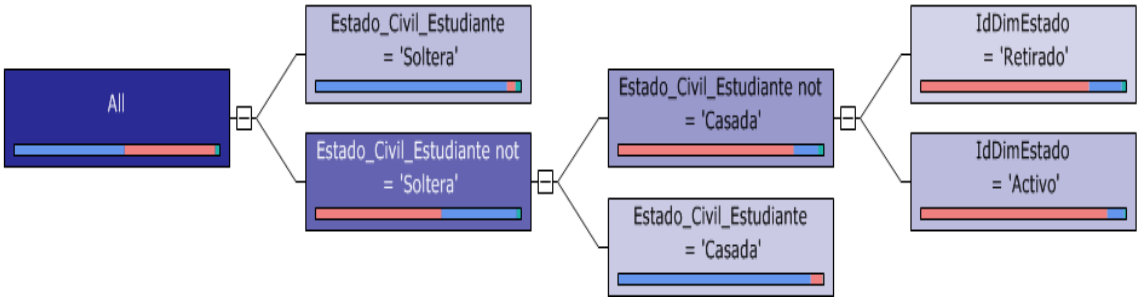


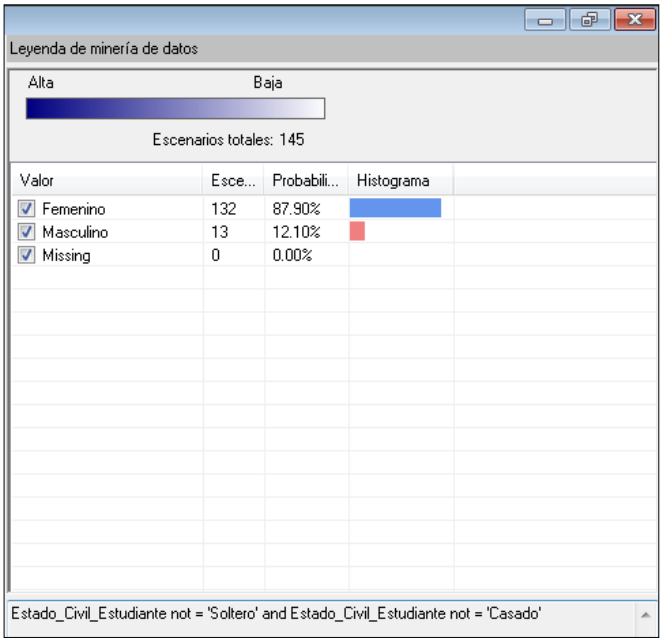
Figura 13: Árbol de Decisiones

ANALISIS SERVICE RED DE DEPENDENCIA

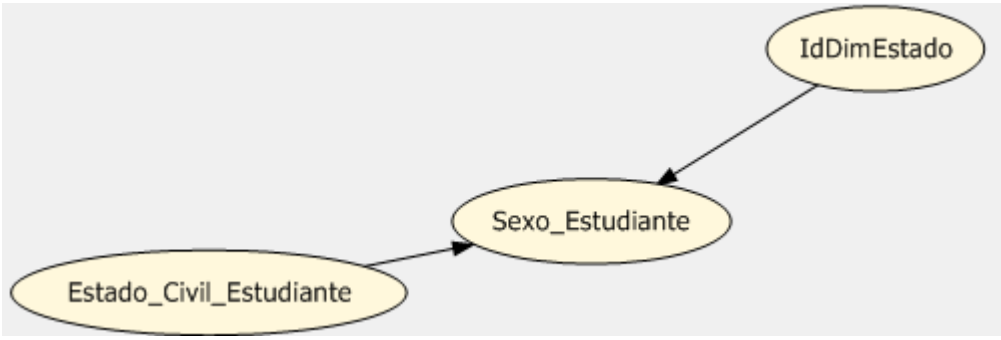


Árbol de decisión
Sexo_Estudiante

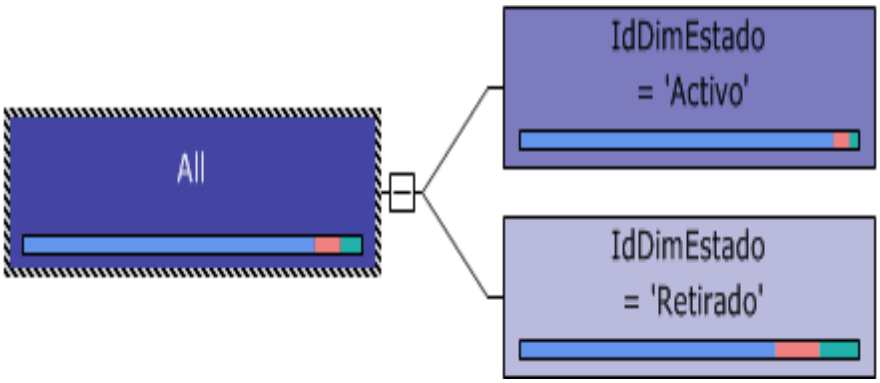


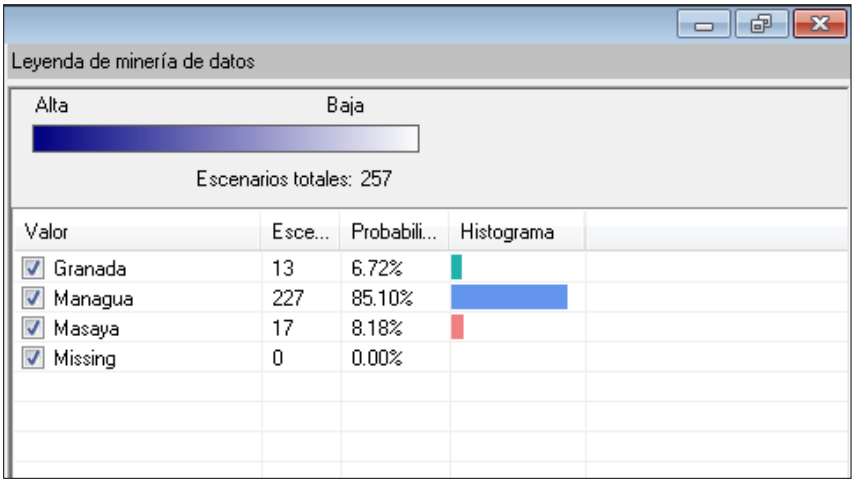


Red de dependencias

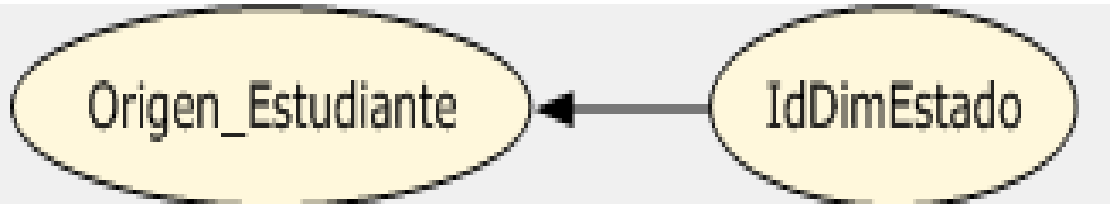


Árbol de decisión
Origen_Estudiante

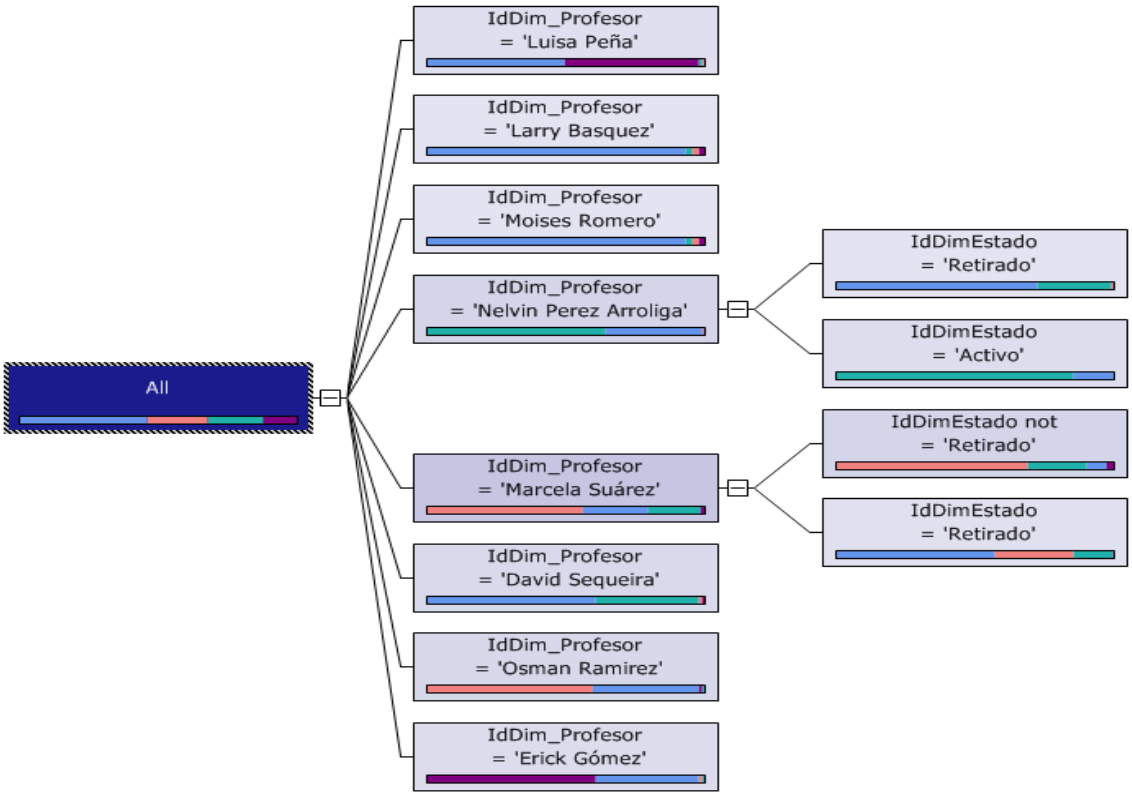




Red de dependencias



Árbol de decisión
IdDim_Curso



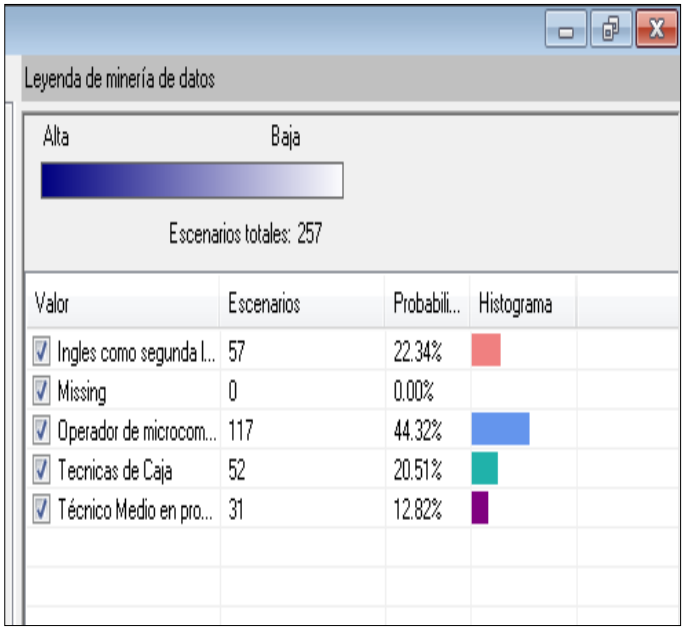


Gráfico de precisión del modelo 'Clasificar IdDimEstado'
Columna de predicción 'IdDimEstado' = Retirado
Mejora respecto al modelo predictivo 'Clasificar IdDimEstado' **105.38 %**

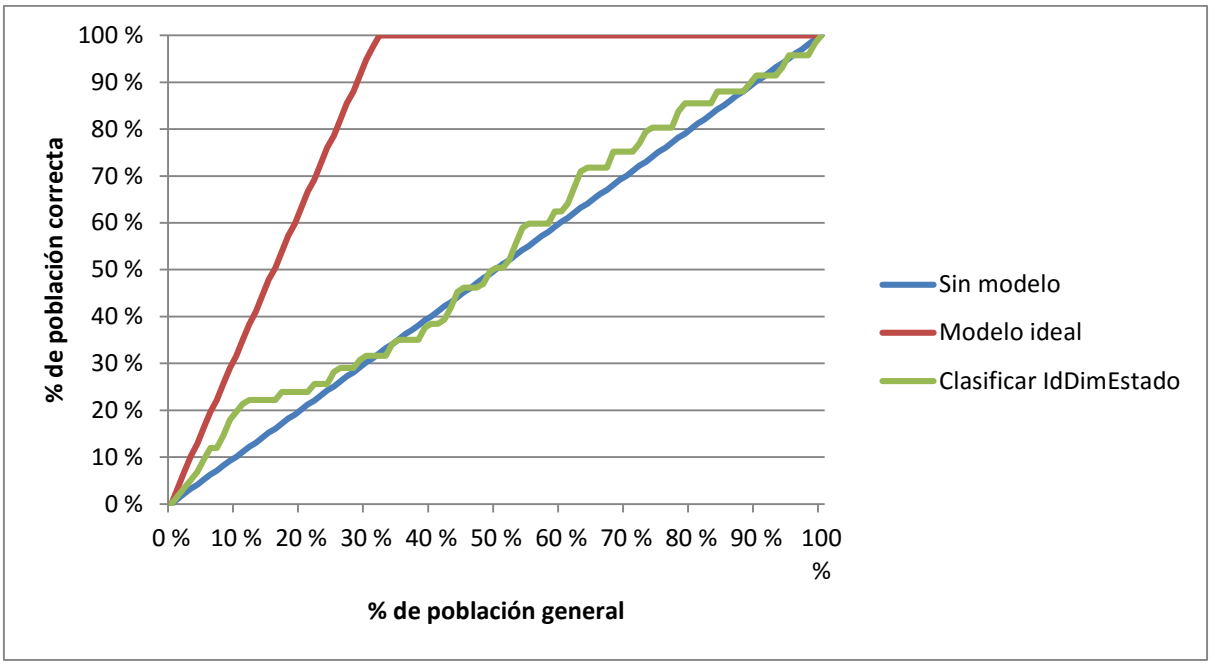
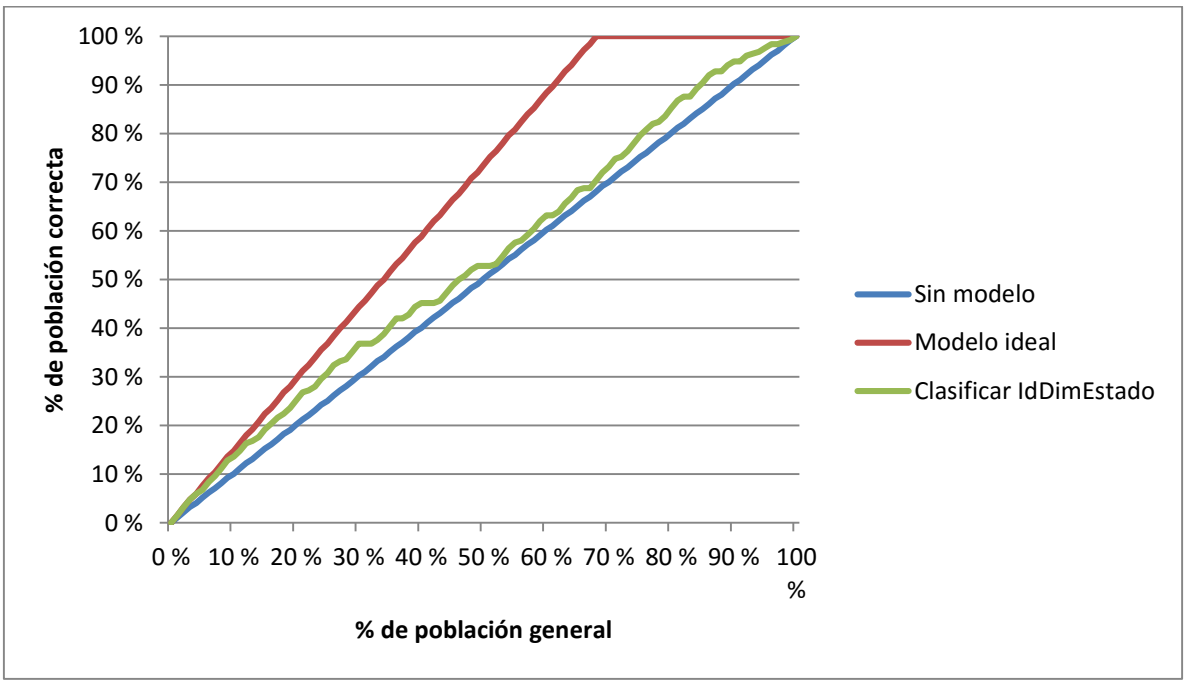


Gráfico de precisión del modelo 'Clasificar IdDimEstado'
Columna de predicción 'IdDimEstado' = Activo
Mejora respecto al modelo predictivo 'Clasificar IdDimEstado' **107.20 %**



- Luego de utilizar el algoritmo de Arboles de Decisión, se generaron los siguientes reportes dinámicos para la ayuda en la toma de decisiones relacionados en el departamento de publicidad y área de matrícula.

Reporte en Excel de estado de los estudiantes del instituto ITEC.

A	B	C	D	E
Carne_est	Nombres_y_Apellidos_est	Nombre_de_Curso	Horario	Estado
001	Marcos Suarez	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
002	milton Gacia	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0026	Amaya Acosta Gadea	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0027	Acosta Jimenez Corina	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0028	Maribel Ramos Calero	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0029	Cristoba Aguilar Gutierrez	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
003	Suyen Rivera	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0030	Elena Alarcon Saavedra	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0031	Joel Sequeira	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0032	Elisabeth Alarcon Suarez	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0033	Sebastiana Alvarez Davila	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0034	Lenin Flores Atamirano	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0035	Sebastiana Gonzales Oliveira	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0036	Juan Jose Alveres Rivera	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0037	Sofia Salazar Flores	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0038	Filomena Rojas Bautista	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0039	Arelys Davila Herbas	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
004	Eveling Zamora	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0040	Vivian Ayala Suarez	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0041	Grachis Rosales Cerna	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0042	Stephania Ayerdis Buitrago	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0043	Celia Cabrera Cerpas	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0044	German Maldonado Cabrera	Operador de microcomputadora	Domingos 1 a 5 pm	Activo
0045	Lucila Seledon Hernandez	Operador de microcomputadora	Domingos 1 a 5 pm	Activo

Tabla 2: Reporte de Estado de estudiantes

Reporte de los horarios de los Profesores del Instituto de Computación ITEC.

Nombres_y_Apellidos_prof	Horario	Nombre_de_Curso	Duracion_Curso	Nombres_y_Apellidos_est	Año	NomDiaSem
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Kevin Fernando Ruiz Machado	2015	Viernes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	German Maldonado Cabrera	2015	Jueves
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Hector Lopez Carcamo	2015	Miércoles
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Lucila Seledon Hernandez	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Esther Rodriguez Saabedra	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Yahoska Liseth Gutierrez	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Meyling Andino Hernandez	2015	Sábado
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Acosta Jimenez Corina	2015	Miércoles
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Leonel Delgado Moran	2015	Miércoles
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Emmanuel Rocha Gonzales	2015	Jueves
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Marlene Suarez Davila	2015	Jueves
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Emmanuel Rocha Gonzales	2015	Sábado
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Marbelly Ramirez Flores	2015	Sábado
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Deyling Cruz Polanco	2015	Martes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Diana Olivares Tellez	2015	Domingo
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Aracelly Argentina López	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Jannina Aleman Ortega	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Daniela Davila Chavarria	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Daryl Guissel Diaz Diaz	2015	Jueves
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Agustin Gomer Oliveira	2015	Martes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Daril Esquivel Lopez	2015	Martes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Michael Arana Fonceca	2015	Sábado
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Darling Lucia Palacios Potosme	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	Karla Manzanares Olivas	2015	Lunes
Nelvin Perez Arroliga	Domingos 1 a 5 pm	Operador de microcomputadora	1 año	German Maldonado Cabrera	2015	Martes

Tabla 3: Reporte de Horario de docentes

Reporte de los Estudiantes del Instituto de Computación ITEC.

IdDim_Estudiante	Carne_est	fecha_ingreso	Nombres_y_Apellidos_est	edad	Nivel_Academico	Estado_Civil	Ocupacion_est	Ocupacion_del_padre	Ocupacion_de_madre
1	1	15/01/2015	Marcos Suarez	35	Bachiller	casado	trabajo	contador	ama de casa
2	2	15/01/2015	milton Gacia	21	Bachiller	soltero	estudiante	trabajo	secretaria
3	26	16/01/2015	Amaya Acosta Gadea	20	Bachiller	Casada	Ama de casa	Negociante	Ama de casa
4	27	16/01/2015	Acosta Jimenez Corina	16	Bachiller	Soltera	Estudiante	Negociante	Ama de casa
5	28	16/01/2015	Maribel Ramos Calero	20	Bachiller	Casada	Comerciante	Comerciante	Ama de casa
6	29	16/01/2015	Cristoba Aguilar Gutierrez	19	Bachiller	Casado	Guarda de seguridad	Chofer	Cajera
7	3	15/01/2015	Suyen Rivera	14	Bachiller	soltera	estudiante	administrador	ama de casa
8	30	16/01/2015	Elena Alarcon Saavedra	24	Bachiller	Casada	Estudiante	Negociante	Ama de casa
9	31	16/01/2015	Joel Sequeira	22	Bachiller	Casado	Construcción	Construccion	Cajera
10	32	16/01/2015	Elisabeth Alarcon Suarez	16	Bachiller	Soltera	Comerciante	Comerciante	Cajera
11	33	16/01/2015	Sebastiana Alvarez Davila	20	Bachiller	Casada	Estudiante	Comerciante	Comerciante
12	34	16/01/2015	Lenin Flores Atamirano	30	Bachiller	Casado	Contador	Gerente	Ama de casa
13	35	16/01/2015	Sebastiana Gonzales Oliveira	17	Bachiller	Soltera	Estudiante	Negociante	Cocinera
14	36	16/01/2015	Juan Jose Alveres Rivera	17	Bachiller	Soltero	Estudiante	Policia	Ama de casa
15	37	16/01/2015	Sofia Salazar Flores	15	Bachiller	Soltera	Negociante	Negociante	Negociante
16	38	16/01/2015	Filomena Rojas Bautista	23	Bachiller	Casada	Estudiante	Comerciante	Ama de casa
17	39	16/01/2015	Arelys Davila Herbas	20	Bachiller	Soltera	Estudiante	Negociante	Ama de casa
18	4	15/01/2015	Eveling Zamora	15	Bachiller	soltera	estudiante	contador	ama de casa
19	40	16/01/2015	Vivian Ayala Suarez	20	Bachiller	Soltera	Estudiante	Contador	Ama de casa
20	41	16/01/2015	Grachis Rosales Cerna	20	Bachiller	Casada	Estudiante	Guarda de Seguridad	Ama de casa
21	42	16/01/2015	Stephania Ayerdis Buitrago	15	Bachiller	Soltera	Estudiante	Administrador	Ama de casa
22	43	16/01/2015	Celia Cabrera Cerpas	24	Bachiller	Casada	Estudiante	Contador	Ama de casa
23	44	16/01/2015	German Maldonado Cabrera	24	Bachiller	Casado	Construccion	Guarda de seguridad	Ama de casa

Tabla 4: Reporte de Estudiantes

- Evaluamos la aplicación según la norma **ISO-9126-1 "Las característica de eficiencia"**

10.4.1.1 Este modelo de Minería de Datos es **Funcional** por suministrar los servicios necesarios para cumplir con los requisitos requeridos.

10.4.1.2 Es de fácil manejo y adaptación para el usuario al momento de utilizar el producto satisfactoriamente.

10.4.1.3 Es **portable** ya que tiene la capacidad de ser transferido de un entorno a otro

10.4.1.4 **Eficiente** al momento de presentar la información concreta y concisa de manera rápida utilizando los reportes generados

EVALUACION DEL MODELO DE MINERIA DE DATOS
ALGORITMO DE ARBOLES DE DECISIÓN SEGÚN
NORMA ISO 9126-1 CARACTERISTICAS DE CALIDAD Y SUBCARACTERISTICAS
INSTITUTO TECNICO DE COMPUTACIÓN Y COMERCIO "ITEC"

TIPOS DE CARACTERISTICAS				
USUARIOS	**FUNCIONABILIDAD**	**USABILIDAD**	**PORTABILIDAD**	**EFICIENCIA**
CAJERA	78	70	50	85
AUXILIAR DE OFICINA	76	65	60	85
RESPONSABLE	80	60	75	84
TOTAL DE EVALUACIÓN	**234**	**195**	**185**	**254**

Tabla 5: **Evaluación de Modelo según la Norma ISO**

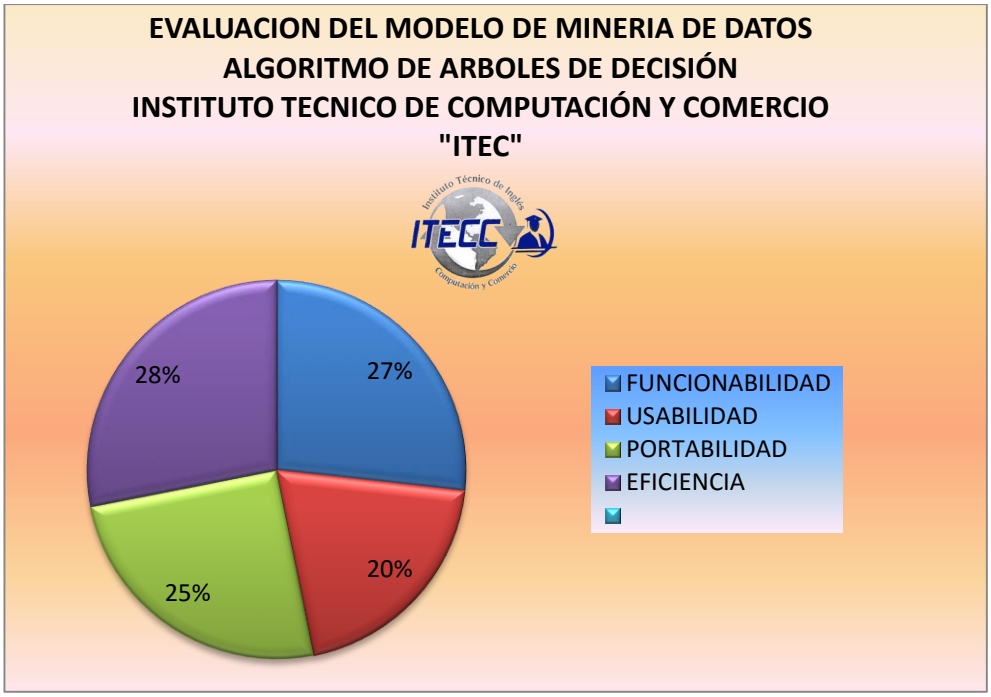


Figura 14: Evaluación de la Norma ISO

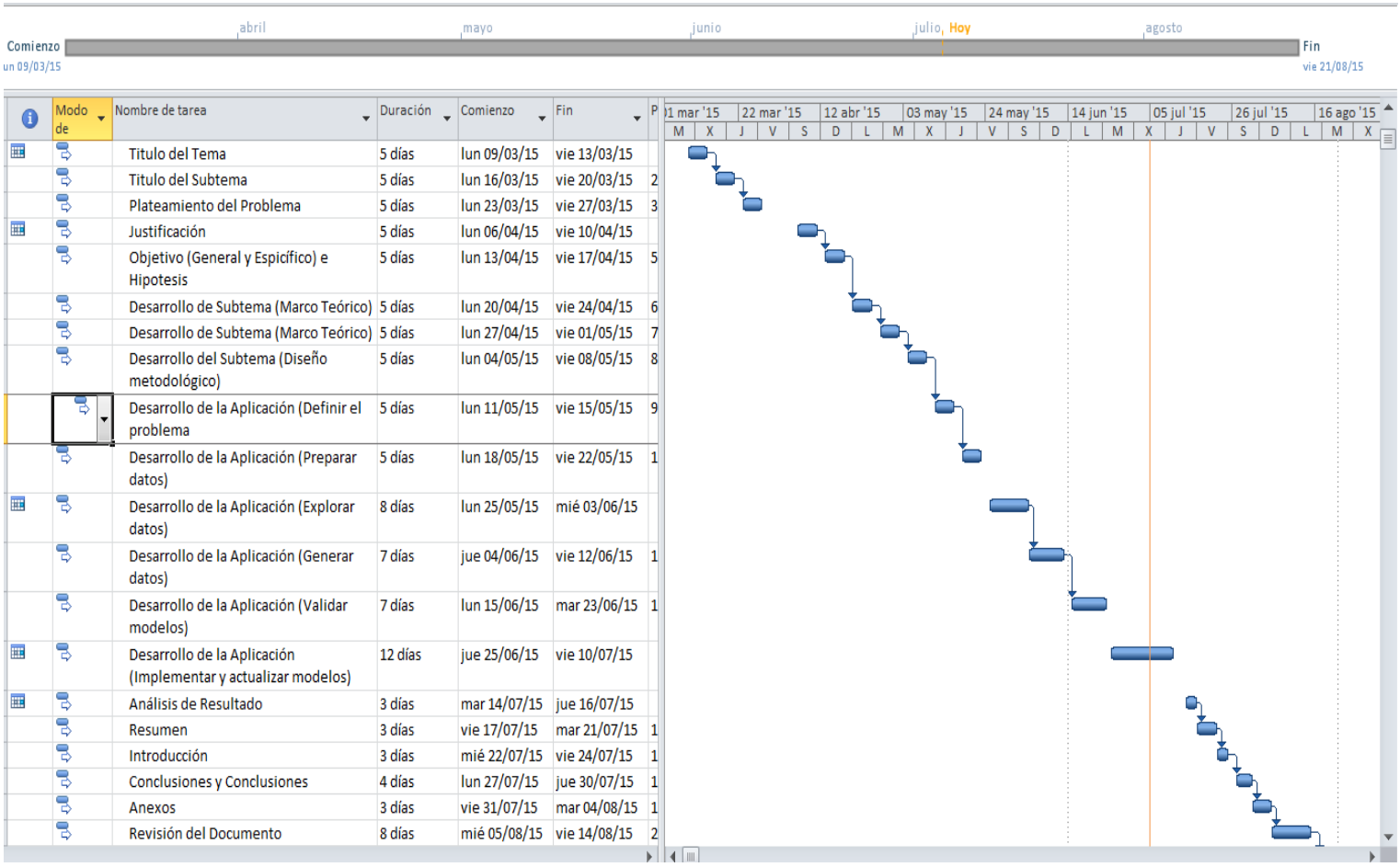
EVALUACION DEL MODELO DE MINERIA DE DATOS
ALGORITMO DE ARBOLES DE DECISIÓN
NORMA ISO 9126-1 CARACTERISTICAS DE CALIDAD Y SUBCARACTERISTICAS
INSTITUTO TECNICO DE COMPUTACIÓN Y COMERCIO "ITEC"

CARACTERISTICAS	CAJERA	AUXILIAR DE OFICINA	RESPONSABLE	TOTAL
FUNCIONABILIDAD	30	30	20	80
USABILIDAD	20	15	30	65
PORTABILIDAD	20	15	25	60
EFICIENCIA	30	25	30	85
TOTAL DE EVALUACIÓN	100	85	105	290

Tabla 6: Evaluación del modelo según Norma ISO

Figura 15: **Segunda Evaluación de la Norma ISO**

XI. CRONOGRAMA DE TRABAJO



Modo de	Nombre de tarea	Duración	Comienzo	Fin	P
	Titulo del Tema	5 días	lun 09/03/15	vie 13/03/15	
	Titulo del Subtema	5 días	lun 16/03/15	vie 20/03/15	2
	Plateamiento del Problema	5 días	lun 23/03/15	vie 27/03/15	3
	Justificación	5 días	lun 06/04/15	vie 10/04/15	
	Objetivo (General y Espicífico) e Hipotesis	5 días	lun 13/04/15	vie 17/04/15	5
	Desarrollo de Subtema (Marco Teórico)	5 días	lun 20/04/15	vie 24/04/15	6
	Desarrollo de Subtema (Marco Teórico)	5 días	lun 27/04/15	vie 01/05/15	7
	Desarrollo del Subtema (Diseño metodológico)	5 días	lun 04/05/15	vie 08/05/15	8
	Desarrollo de la Aplicación (Definir el problema	5 días	lun 11/05/15	vie 15/05/15	9
	Desarrollo de la Aplicación (Preparar datos)	5 días	lun 18/05/15	vie 22/05/15	1
	Desarrollo de la Aplicación (Explorar datos)	8 días	lun 25/05/15	mié 03/06/15	
	Desarrollo de la Aplicación (Generar datos)	7 días	jue 04/06/15	vie 12/06/15	1
	Desarrollo de la Aplicación (Validar modelos)	7 días	lun 15/06/15	mar 23/06/15	1
	Desarrollo de la Aplicación (Implementar y actualizar modelos)	12 días	jue 25/06/15	vie 10/07/15	
	Análisis de Resultado	3 días	mar 14/07/15	jue 16/07/15	
	Resumen	3 días	vie 17/07/15	mar 21/07/15	1
	Introducción	3 días	mié 22/07/15	vie 24/07/15	1
	Conclusiones y Conclusiones	4 días	lun 27/07/15	jue 30/07/15	1
	Anexos	3 días	vie 31/07/15	mar 04/08/15	1
	Revisión del Documento	8 días	mié 05/08/15	vie 14/08/15	2

Figura 16: Cronograma de Trabajo

XII. CONCLUSIONES

El Aprendizaje Automático en general, y de los métodos del Instituto Técnico de Computación y Comercio en particular, hay varios puntos claves a tener en cuenta al realizar Minería de Datos con algoritmos inteligentes. Entre ellos están:

- La información de la situación en que se encontró el Instituto técnico de computación y comercio “ITEC” en cuanto a su matrícula fue utilizada de manera eficiente para aplicar el modelo de minería de datos utilizando el algoritmo árboles de decisión.

- El trabajo de investigación demostró que es factible las técnicas de minería de datos utilizando los árboles de decisión, desarrollados en primera instancia para ser aplicados en otros campos de investigación, al contexto de los datos estadísticos para el mejoramiento en el área de matrícula.

- Con la obtención de reportes de información necesaria facilitó un alto grado de comprensión del conocimiento utilizado en la toma de decisiones futuras.

- Según la evaluación del usuario a nuestro modelo se pudo concluir que es eficiente, lo cual se confirmó que el modelo cumplió todos los requerimientos para la solución de los objetivos propuestos.

XIII. RECOMENDACIONES

- Lo principal para iniciar un modelo de inteligencia de negocio, es determinar el estado actual de la empresa, como el nivel de integración de procesos y el uso de tecnología. La información proporcionada por el instituto, sea de fiabilidad, veraz, concreta y concisa.

- Utilizar los resultados del modelo de minería de datos en aplicaciones que permitan mejorar el proceso de formación académica de los estudiantes.

- Mostrar los reportes de la información obtenida con el uso del algoritmo de Arboles de decisión en nuevas tomas de decisiones para el mejoramiento en el área de matrícula del Instituto Técnico de Computación y Comercio.

- Continuar la investigación a partir de los resultados obtenidos, siguiendo las orientaciones de la fase de Evaluación.

XIV. BIBLIOGRAFIA

E. Kendall, K. y. (2005). *Análisis y Diseño de Sistemas.* México: Guillermo Trujano Mendoza.

Gutierrez, D. R. (2014). *Modelo ISO 14598.* Prezi.

INCAP. (s.f.). *www.incap.int/sisvan/index.php*. Recuperado el 30 de Abril de 2015, de www.incap.int/sisvan/index.php: http://www.incap.int/sisvan/index.php

Kendall, K., & Kendall, J. (2011). *Análisis y Diseño de Sistemas.* Mexico: Pearson Educación.

Krall, C. (2012). *aprenderaprogramar.com/index.php.* Recuperado el 02 de Mayo de 2015, de aprenderaprogramar.com/index.php: http://www.aprenderaprogramar.com/index.php

López, C. P., & González, D. S. (2007). *Mineria de Datos.* Madrid: Paninfo.

Orallo, J. H., Ramírez, M. J., & Ferrí, C. (2004). *Introducción a la Mineria de Datos.* España: Pearson.

Piura López, J. (2010). *Metodología de la Investigación.* España: Cooperación Española.

Pressman, R. S. (2005). *Ingeniería del Software.* Madrid: Miguel Martin Romo.

Ramirez, E. (2010). *Fundamentos de Sistemas de Base de Datos.* Pearson.

Sampieri, R., & Fernandez, C. (2010). *Metodologia de la Investigación 5ta. ed.* Mexico: Mc. Graw Hill Interamericana.

Silberschatz, A. (2007). *Fundamentos de diseño de base de datos.* España: S.A. Mcgraw-Hill.

Somos estudiantes egresados de la carrera Ciencias de la computación, estamos realizando una encuesta sobre el uso de un modelo de Minería de Datos aplicando el Algoritmo Árboles de Decisión. Le agradeceríamos su colaboración con nosotros contestándonos unas preguntas. *Muchas Gracias,* por brindarnos unos minutos de su tiempo.

XV. COMPENDIOS

PROTOCOLO DE LA ENTREVISTA

10.4.1.5 Institución: Instituto Técnico de Computación y Comercio "ITEC" .

10.4.1.6 Personas a entrevistar: Docentes, Administrativos, Directivos del Instituto,

10.4.1.7 Objetivos de la Entrevista: Esta Técnica investigativa tiene como objetivo, recolectar la información veraz y concisa necesaria para darle un buen funcionamiento a nuestro modelo de Minería de Datos, aplicando nuestro algoritmo de árboles de decisión, utilizando diferentes opiniones y puntos de vistas de los trabajadores en esta institución.

10.4.1.8 Temas a tratar en esta entrevista: Las temáticas o ejes centrales bajo los cuales se realizaran las entrevistas, estarán centrados en: a) La situación actual del instituto, b) Información básica sobre los estudiantes c) La toma de decisión de desarrollar el método de minería de datos.

10.4.1.9 Referencia Técnica y Contextual del Instrumento Metodológico:

10.4.1.9.1Método: Entrevista

10.4.1.9.2Técnica: Entrevista Estructurada

10.4.1.9.3Fecha: del 10 al 20 de Mayo del 2015

10.4.1.9.4Duración: 15 minutos

10.4.1.9.5Lugar: Instituto Técnico de Computación y Comercio

10.4.1.9.6 Contexto: Ambiente propio del Instituto

10.4.1.9.7 ¿Quién lo va a entrevistar?: El grupo consultor

ENTREVISTA

1. ¿Cuándo se fundó el Instituto?
2. ¿Cómo se almacenaban los datos de los estudiantes anteriormente?
3. ¿Cuánto es el promedio de estudiantes matriculados cada año en el instituto?
4. ¿El sistema de matrícula rinde los mejores resultados actualmente?
5. ¿Cree conveniente el nuevo método de minería de datos para el sistema de matrícula?
6. ¿Cuáles son los factores que inciden en la deserción de los estudiantes en los cursos?
7. ¿Cuánto es el promedio de estudiantes que desertan del instituto en el año?
8. ¿De qué manera nos puede ayudar la minería de datos para prevenir la deserción de los estudiantes?
9. ¿Este modelo predictivo permitirá identificar a los alumnos vulnerables de retirarse al inicio de su estancia en el instituto?

Muchas gracias por su colaboración.