# OBTENCIÓN DE PATRONES Y REGLAS EN EL PROCESO ACADÉMICO DE LA UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS

***Ernesto González Díaz[1], Zady Pérez Hernández[2], Ivet Espinosa Conde [3], Susel Alvarez Reyes[4]***

*[1] CEIS. Master en Ciencias. egonzalez@ceis.cujae.edu.cu*

*[2] CEIS, Ingeniero en Informática, iespinosa@ceis.cujae.edu.cu*

*[3] CEIS, Ingeniero en Informática, zperez@ceis.cujae.edu.cu*

*4 UCI, Ingeniero en Informática, sucellealvarez@gmail.com*

## RESUMEN

A partir de la aplicación de un grupo de técnicas de Minería de Datos como el clustering, los árboles de decisión y algoritmos de aprendizaje inductivo; se pretende clasificar a los estudiantes de acuerdo a su rendimiento académico, para posteriormente encontrar patrones ocultos y reglas que los caractericen; basado en las relaciones que se establecen entre el centro de procedencia de los estudiantes, nivel de escolaridad de los padres y provincia de origen con sus resultados académicos en el primer curso en la universidad. Estos resultados pueden mejorar el proceso de formación académica y elevar la calidad de la educación en la Universidad de las Ciencias Informáticas (UCI).

Palabras claves: Calidad del proceso docente, Descubrimiento de Conocimientos en Bases de Datos, Minería de Datos.

## *ABSTRACT*

*This investigation intends to classify the students of the University of Informatics Sciences according to their academic behaviour using a set of Data Mining techniques like clustering, decision trees and inductive learning algorithms. The main goal of this work is to find hidden patterns and rules that define this behaviour, based on the relationship established between the scholarship level of the student's parents, and their academic origins with their grades in the first year of their career. These results can help to improve the quality of the academic process in the UCI.*

*Key words: Quality of the academic process, Knowledge Discovery in Databases, Data Mining*

# INTRODUCCIÓN

La Universidad de las Ciencias Informáticas (UCI) cuenta desde el curso escolar 2006-2007 con una matrícula de alrededor de 10 000 estudiantes procedentes de todas las provincias y municipios del país, con los más diversos orígenes sociales y académicos; sin que, hasta el momento, se hayan realizado estudios que evalúen la influencia de estos factores en su formación posterior. Por lo que estos factores no son tomados en cuenta a la hora de realizar el proceso de captación de los estudiantes de nuevo ingreso a la universidad, ni de brindarles a los ya matriculados el seguimiento necesario, lo que puede conducirlos en condiciones extremas a causar baja del centro. Mientras que en otros casos se dejan de identificar a los alumnos con mayor potencial, que pudieran formar parte de proyectos o grupos de investigación, o simplemente armar al claustro de profesores con la información conveniente para que puedan brindarle atención diferenciada a sus estudiantes en aras de fomentar el pleno desarrollo de sus capacidades y dándole así cumplimiento al objetivo primordial de la Universidad, que es el de formar profesionales de la informática cada vez mejor preparados.

Toda la información personal y docente de los estudiantes, desde hace cinco años se encuentra digitalizada y se mantiene en históricos que no brindan mayor utilidad que la de los reportes tradicionales.

Es por esto que en la Universidad se hace necesario contar con métodos eficientes y automáticos para explorar las grandes Bases de Datos, procesando de forma rápida y fiable la información para encontrar patrones de conocimiento apropiados para resolver un problema.

Es por esto que el objetivo fundamental de este trabajo está orientado a determinar el vínculo que existe entre el origen y procedencia social de los estudiantes de la UCI con sus resultados académicos mediante la aplicación de

técnicas de agrupación y reglas de asociación de Minería de Datos.

# 1. La Minería de Datos y el Descubrimiento de Conocimiento en Bases de Datos.

La Minería de Datos (DM) por las siglas en inglés Data Mining es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [1]. Las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios la toma de decisiones.

Existen términos que se utilizan frecuentemente como sinónimos de la minería de datos. Uno de ellos se conoce como "análisis (inteligente) de datos" [2], que suele hacer un mayor hincapié en las técnicas de análisis estadístico. Otro término muy utilizado, y el mas relacionado con la minería de datos, es la extracción o "descubrirniento de conocimiento en bases de datos" (Knowledge Discovery in Databases o KDD, según sus siglas en inglés). [3]

Aunque algunos autores usan los términos Minería de Datos y KDD indistintamente, como sinónimos, existen claras diferencias entre los dos. Así la mayoría de los autores coinciden en referirse al KDD como un proceso que consta de un conjunto de fases, una de las cuales es la minería de datos. [2] De acuerdo con esto, el proceso de minería de datos consiste únicamente en la aplicación de un algoritmo para extraer patrones de datos y se llamará KDD al proceso completo que incluye pre-procesamiento, minería y post-procesamiento de los datos.

El KDD según [4] es la extracción automatizada de conocimiento o patrones interesantes, no triviales, implícitos, previamente desconocidos, potencialmente útiles y predictivos de la información de grandes Bases de Datos.

La figura 1 muestra las fases del proceso de KDD, una de las cuales es la Minería de Datos

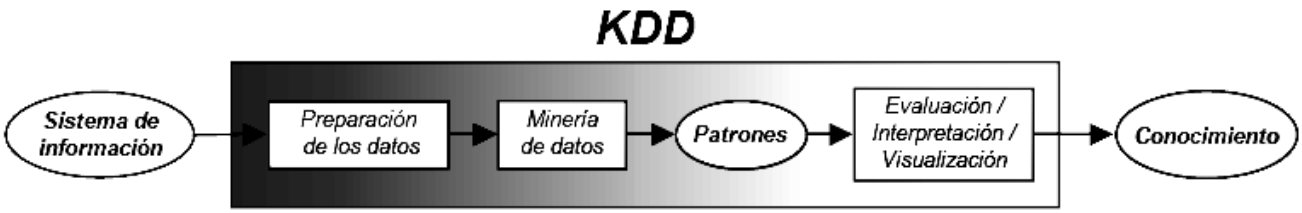


**Figura 1**: Fases del proceso KDD

Las investigaciones en temas de KDD incluyen análisis estadístico, técnicas de representación del conocimiento y visualización de datos, entre otras. Algunas de las tareas más frecuentes en procesos de KDD son la clasificación y clustering, el reconocimiento de patrones, las predicciones y la detección de dependencias o relaciones entre los datos.

## 1.1 Proyectos en Minería de Datos

Los pasos a seguir para la realización de un proyecto de minería de datos son siempre los mismos, independientemente de la técnica específica de extracción de conocimiento usada.

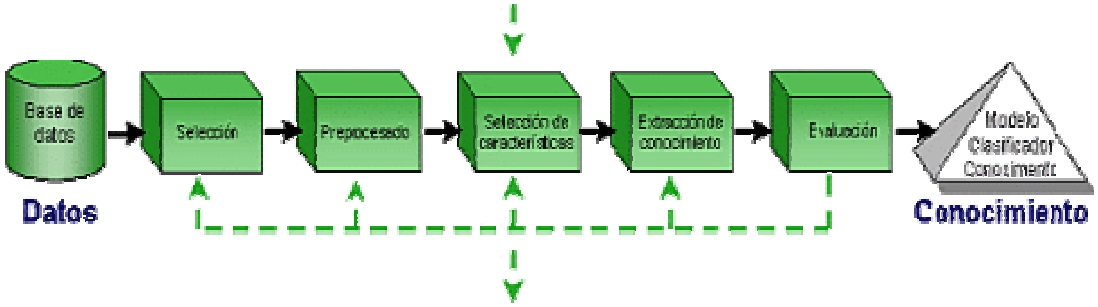


**Figura 2:** Fases dentro de un proceso de Minería de Datos

El proceso de minería de datos pasa por las siguientes fases:

1. Comprensión del negocio y del problema que se quiere resolver.
2. Filtrado de datos:
   El formato de los datos contenidos en la fuente de datos nunca es el correcto, y la mayoría de las veces no es posible ni siquiera utilizar algún algoritmo de minería sobre los datos iniciales sin que requieran alguna transformación. En este paso se filtran los datos con el objetivo de eliminar valores incorrectos, no válidos o desconocidos; según las necesidades y el algoritmo a utilizar. Además se obtienen muestras de los datos en

busca de mayor velocidad y eficiencia de los algoritmos, o se reducen el número de valores posibles para los atributos de análisis.

3. Selección de variables:
   Después de realizar la limpieza de los datos, en la mayoría de los casos se tiene una gran cantidad de variables o atributos. La selección de características reduce el tamaño de los datos, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería; seleccionando las variables más influyentes en el problema.
   Los métodos para la selección de los atributos que más influencia tienen en el problema son básicamente dos:
   - Aquellos basados en la elección de los mejores atributos del problema.
   - Aquellos que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.
4. Extracción de Conocimiento

La extracción del conocimiento es la esencia de la Minería de Datos donde mediante una técnica, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Los modelos que se generan son expresados de diversas formas:

- reglas
- árboles
- redes neuronales

También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre-procesado diferente de los datos.

5. Interpretación y Evaluación

Una vez obtenido el modelo, se procede a su validación; donde se comprueba que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos para buscar el que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

## 2. Herramientas para la minería de Datos. SQL Server 2005.

Microsoft SQL Server 2005 incorpora la herramienta SQL Analysis Server estableciendo nuevas facilidades para realizar Minería de Datos, entre las que se cuentan:

- El procesamiento de los modelos de una misma estructura de minería ocurre en paralelo, en una sola lectura de los datos.
- Proporciona más de 12 visores de resultados para los algoritmos que ayudarán a comprender mejor los patrones encontrados en el proceso de minería.
- Proporciona gráficos de elevación, de beneficios y una matriz de clasificación que permite establecer una comparación de lo real con lo previsto; para contrastar y comparar la calidad de los modelos.
- Posee un lenguaje para la creación de consultas de minería (DMX) similar al SQL que facilita la tarea de creación de aplicaciones de minería de datos.
- Posee una interfaz gráfica para generar las consultas DMX.
- Cuenta con los algoritmos de minería más avanzados: Naive Bayes, Clustering, Clústeres de Secuencia, Árboles de Decisión, Redes Neuronales, Series Temporales, Reglas de Asociación, Regresión Logística, y Regresión Lineal y minería de textos.
- Marco de desarrollo para agregar nuevos algoritmos y también para construir visores propios para los modelos generados. [5] [6] [7] [8] [9] [10].

## 3. Metodologías de desarrollo para proyectos de Minería de Datos. CRISP-DM.

La metodología CRISP-DM [11] consiste en un conjunto de tareas descritas en cuatro niveles de abstracción: fase, tarea genérica, tarea especializada, e instancia de proceso, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos.

**Fase**: Se le denomina fase al asunto o paso dentro del proceso.CRISP-DM consta de 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelación, evaluación y explotación.

**Tarea genérica**: Cada fase esta formada por tareas genéricas, o sea, la tarea genérica es la descripción de las actividades que se realizan dentro de cada fase. Por ejemplo, la tarea Limpiar los datos es una tarea genérica.

**Tarea especializada**: La tarea especializada describe cómo se pueden llevar a cabo las tareas genéricas en situaciones específicas. Por ejemplo, la tarea Limpiar los datos tiene tareas especializadas, como limpiar valores numéricos, y limpiar valores categóricos.
**Instancias de proceso**: Las instancias de proceso son las acciones y resultados de las actividades realizadas dentro de cada fase del proyecto.
Las fases del proyecto de Minería de acuerdo a lo establecido por la metodología CRISP-DM interactúan entre ellas de forma iterativa durante el desarrollo del proyecto. La secuencia de las fases no siempre es ordenada, o en ocasiones si se determina al realizar la evaluación que los objetivos del negocio no se cumplieron se debe regresar y buscar las causas del problema para redefinirlo.

# 4. Resultados del Caso de Estudio.

El caso de estudio seleccionado para realizar el proyecto de KDD se refiere a la predicción de las notas de las asignaturas del primer año de los estudiantes de la UCI basado en las relaciones que se establecen entre el nivel de escolaridad de los padres, tipo de centro de procedencia, provincia y resultados académicos.
Los datos seleccionados para realizar el proyecto de Minería de Datos corresponden a la información personal y calificaciones en las asignaturas del primer curso escolar de los estudiantes de la UCI que actualmente cursan el segundo, tercero, cuarto y quinto año. Se utiliza la información del primer curso escolar ya que los estudiantes de primer año reciben mayor influencia de las variables a analizar como entradas para las predicciones.
Se utiliza una muestra aleatoria representativa del 90% de los datos para realizar el proyecto de KDD.
El proyecto fue desarrollado por siguiendo los lineamientos de la metodología CRISP-DM.

**1- Comprensión del negocio.**

La UCI dispone de un Sistema Automatizado para la gestión académica de los estudiantes (AKADEMOS). En el mismo se almacena información personal y resultados académicos de los estudiantes en las diferentes asignaturas. El sistema brinda la utilidad de los reportes tradicionales que permiten obtener información de los estudiantes que han matriculado en la universidad.
AKADEMOS es un sistema informático en el cual todos los involucrados (directivos, personal de secretaría, profesores y estudiantes) tienen un papel activo en el proceso de gestión académica. A partir de la información que brinda este sistema y con los test evaluativos que se realizan a los estudiantes antes de matricular en la universidad, en la UCI; específicamente en el Centro de Investigaciones por la Calidad de la Educación (CICE), se está desarrollando el proyecto “Perfeccionamiento del proceso de selección para nuevos ingresos al curso regular de la Universidad de Ciencias Informáticas”.
Después de realizar entrevistas a usuarios, personal de la Dirección de Informatización de la UCI y de la Dirección del Centro de Investigación por la Calidad de la Educación (CICE) en la Universidad de las Ciencias Informáticas; se definió el siguiente caso de estudio a realizar en el proyecto de KDD:
Predecir las notas de las asignaturas del primer año de los estudiantes de la UCI basado en las relaciones que se establecen entre el nivel de escolaridad de los padres, tipo de centro de procedencia, provincia y resultados académicos.

2- **Comprensión de los datos.**

Los datos utilizados pertenecen al período del 2001 hasta el 2006, específicamente a la información personal y académica de los estudiantes que eran matrícula de la UCI en esta etapa; tomando de estos la información histórica en su primer año en la universidad. La Base de Datos se encontraba en un servidor SQL Server 2000, por lo que fue necesario importarla para un servidor SQL Server 2005, en orden de poder utilizar las facilidades que brinda esta herramienta para la Minería de Datos.
Para decidir que datos utilizar se realizó un estudio conjunto entre especialistas y desarrolladores; donde se analizó el contenido y la complejidad de la Base de Datos, de las tablas implicadas y sus relaciones; así como el tipo de datos de los atributos, sus posibles valores, significado en el negocio y relevancia dentro del mismo; además se comprobaron los atributos de entradas libres y si existían llaves repetidas.
Sólo se tomaron en cuenta los resultados académicos del primer curso escolar; pues sobre estas existe mayor influencia de las variables centro de procedencia, provincia y nivel de escolaridad de los padres. Debido a los límites de la investigación no se seleccionaron todas las asignaturas de primer año, sino aquellas que se consideraron más relevantes, Matemática Discreta, Introducción a la Programación, Programación I, Matemática I y Algebra Lineal.
Los atributos más importantes para el proyecto de Minería fueron analizados en el diseñador de vistas de origen de datos de Business Intelligence Development Studio y el editor de consultas del Management Studio. Con estas herramientas se estudiaron los atributos, sus valores y el comportamiento de los mismos.

Se realizó una búsqueda de los posibles valores de los atributos, a partir de la fuente de datos con las herramientas de Microsoft Office Web Components, con el objetivo de encontrar valores incorrectos que pudieran traer problemas en las predicciones, además para analizar cuales atributos podrían requerir discretización.

**3- Preparación de los datos.**

Toda la información necesaria para realizar la investigación se encuentra en la Base de Datos AKADEMOS por lo que no fue necesario integrar varios orígenes de datos. Los atributos seleccionados para realizar el proyecto de Minería correspondiente a los datos personales de los estudiantes se encontraban en varias vistas dentro de la Base de Datos.

Con el objetivo de asociar en una sola tabla los datos personales de los estudiantes; en el Integration Services utilizando el componente Union Alll se obtuvo la tabla *Datos Históricos* a partir de las 4 vistas Hoja de matricula_108_e, Hoja de matricula_110_e, Hoja de matricula_112_e y Hoja de matricula_114_e donde se encuentra la información de los estudiantes matriculados en la Universidad en el período comprendido entre los años 2001 al 2006; como se observa en la figura 3.

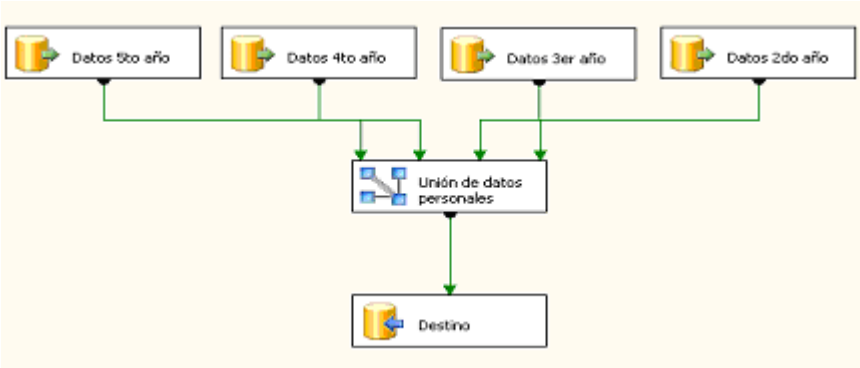


**Figura 3:** Unión de los datos personales de los estudiantes.

A partir de la tabla que contiene las asignaturas pivoteadas y de la tabla donde se encuentran los datos personales de los estudiantes; se obtiene una nueva vista (*Notas Datos)* donde se asocia la información perteneciente a los mismos objetos.

Referente a los casos sobre los que se trabaja, los mismos fueron seleccionados de la tabla que contiene toda la información personal y académica de los estudiantes (*Notas_Datos*), utilizando el componente Percentage Sampling del SQL Server Integration Services (SSIS), en el proyecto se seleccionó el 90% de los datos.

**4- Modelación**

Para la realización de este paso se utilizaron las técnicas de Minería de Datos del SQL Server 2005, utilizando la herramienta SQL Server Business Intelligence Development Studio, específicamente SQL Server Analysis Services (SSAS).

A continuación se muestran las técnicas y visores a utilizar por cada objetivo de la Minería.

| **Objetivo de Minería** | **Técnica** |
|---|---|
| 1. Realizar una segmentación adecuada de los estudiantes, tomando como columnas de entrada la provincia, nivel de escolaridad de los padres, centro de procedencia y las notas de las asignaturas del primer año de la carrera.<br><br>2. Analizar los clústeres obtenidos de acuerdo a las notas que predominan en cada grupo; como paso analítico para el próximo objetivo. | Algoritmo de clustering de Microsoft<br>• Visor de clústeres de Microsoft<br>• Diagrama del clúster<br>• Perfiles del clúster<br>• Características del clúster<br>• Distinción del clúster<br>. |
| 3. Obtener reglas que permitan descubrir la influencia que tiene la provincia, nivel de escolaridad de los padres y centro de procedencia de los estudiantes en sus | Algoritmo de Árboles de Decisión de Microsoft<br>• Visor de árboles de decisión de Microsoft. |

| | |
|---|---|
| resultados académicos; y permitan predecir la nota final en cada asignatura analizada. | • Red de dependencia<br>• Gráfico de elevación<br>• Matriz de Clasificación |

**Tabla 1**: Técnicas y visores a aplicar por objetivos de la minería.

### Diseño de pruebas.

El diseño de las pruebas sobre los datos se realizó utilizando la herramienta SQL Server Integration Services, empleando la técnica de validación cruzada.

SQL Server Integration Services tiene componentes que permiten obtener muestras aleatorias representativas según un porciento de los datos o según determinada cantidad de filas, estos componentes son el Percentage Sampling y Row Sampling y proporcionan, además otros componentes para unir varias muestras desde diversos orígenes o fuentes de datos, realizar consultas SQL y guardar los resultados obtenidos en diversos destinos.

Utilizando estos componentes se realizaron los diseños de casos de prueba según la técnica de Validación Cruzada; la cual consiste en dividir los datos en 10 grupos o muestras y realizar 10 corridas o iteraciones donde en cada una se combinan 9 muestras para obtener una muestra de experimento y se deja una como muestra de prueba. De esta forma todas las muestras son utilizadas como experimento y como prueba. Al final se selecciona el experimento sobre el cual se realicen mejores predicciones, o sea donde el error sea menor.
En la siguiente figura se muestra el flujo de control del paquete de pruebas del Integration Services, utilizando validación cruzada.

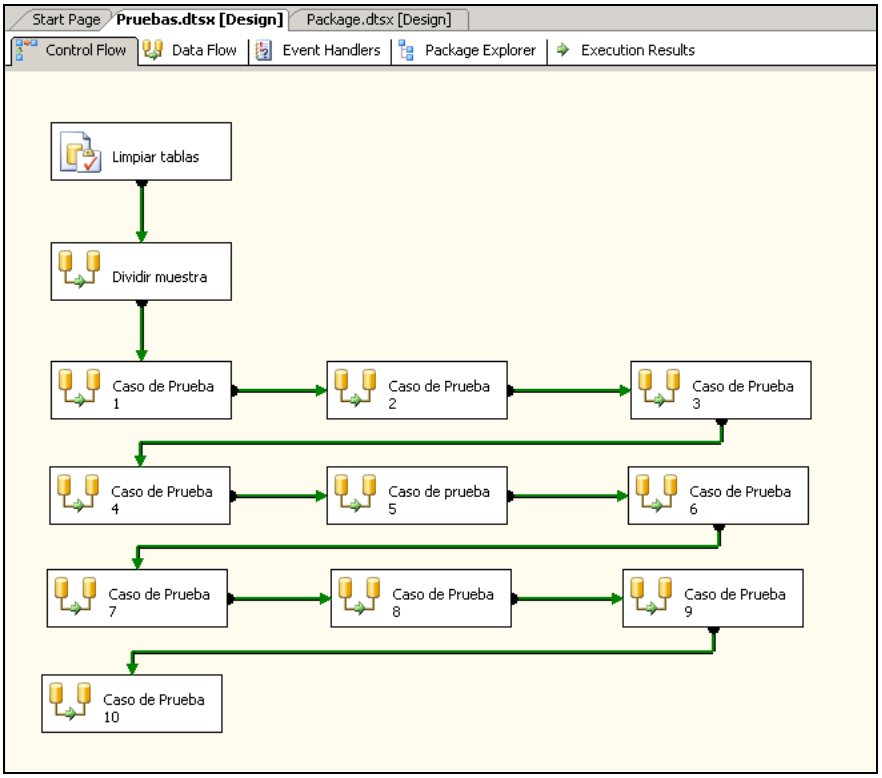


**Figura 5**: Flujo de Control del paquete de pruebas utilizando validación cruzada

### Valoración del modelo Árboles Predicción Nota en el Experimento #3

A partir de los resultados obtenidos por los modelos que se explican en la fase de Evaluación; la predicción logró resolver con éxito los siguientes por cientos de los casos de entrada según las asignaturas y notas:

- En la asignatura Algebra Lineal se resuelven con éxito el 26% de los casos donde la nota es 5 con probabilidad de 0.76; el 40% donde la nota es 4 con probabilidad entre 0.74 y 0.96; el 28% para la nota de 3 con probabilidad mayor que 0.75; y el 4% para la nota 2 con probabilidad de 0.50.
- En la asignatura Introducción a la Programación se resuelven con éxito el 35% de los casos donde la nota es 5 con probabilidad de 0.78; el 22% donde la nota es 4 con probabilidad de 0.60; el 37% para la nota de 3 con probabilidad mayor que 0.72 y el 13% para la nota 2 con probabilidad entre 0.52 y 0.82.
- En la asignatura Matemática Discreta se resuelven con éxito el 21% de los casos donde la nota es 5 con probabilidad mayor que 0.72; el 38% donde la nota es 4 con probabilidad entre 0.65 y 0.73; el 37% para la nota de 3 con probabilidad entre 0.85 y 0.95 y el 4% para la nota 2 con probabilidad de 0.31.
- En la asignatura Matemática I se resuelven con éxito el 18% de los casos donde la nota es 5 con probabilidad de 0.67; el 31% donde la nota es 4 con probabilidad de 0.53; el 46% para la nota de 3 con probabilidad entre 0.85 y 0.93 y el 14% para la nota 2 con probabilidad de 0.65.
- En la asignatura Programación I se resuelven con éxito el 21% de los casos donde la nota es 5 con probabilidad entre 0.68 y 0.93; el 24% donde la nota es 4 con probabilidad entre 0.72 y 0.87; el 50% para la nota de 3 con probabilidad de 0.87 y el 9% para la nota 2 con probabilidad mayor que 0.43.

La predicción es altamente efectiva, las probabilidades son altas en la mayoría de los casos.

**5- Evaluación**

En esta fase se evalúa el modelo escogido, no desde el punto de vista general, sino del cumplimiento de los objetivos del negocio. Se debe revisar el proceso teniendo en cuenta los resultados obtenidos, para repetir alguna fase en caso que se hayan cometido errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase y de la precisión del mismo, se procede al despliegue de éste en caso de requerirse.
Se mostrarán a continuación algunas de las reglas obtenidas, a partir de los modelos de árboles de decisión generados para cada asignatura por nota.

| **Algebra Lineal** | | |
|---|---|---|
| Nota | Reglas | Prob |
| 5 | ➢ TC_De Procedencia <> 'DEPORTE'<br>➢ TC_De Procedencia = 'DEPORTE' | 0.76<br>0.52 |
| 4 | ➢ Provincia = Pinar del Río<br>➢ Provincia <> Pinar del Río<br>➢ Provincia <> Pinar del Río y NE del Padre <> Técnico medio<br>➢ Provincia <> Pinar del Río y NE del Padre = Técnico medio | 0.96<br>0.77<br>0.74<br>0.89 |
| 3 | ➢ NE_Del Padre = 'Preuniversitario'<br>➢ NE_Del Padre <> 'Preuniversitario'<br>➢ NE_Del Padre <> 'Preuniversitario' y NE_De La Madre <> 'Secundaria'<br>➢ NE_Del Padre <> 'Preuniversitario' y NE_De La Madre = 'Secundaria' | 0.93<br>0.75<br>0.73<br>0.79 |
| 2 | ➢ No tiene influencia ninguno de los factores analizados sobre la nota | 0.50 |

**Tabla 2**: Reglas obtenidas para la asignatura Algebra Lineal

| **Programación I** | | |
|---|---|---|
| Nota | Reglas | Prob |
| 5 | ➢ TC_De Procedencia = 'DEPORTE'<br>➢ TC_De Procedencia <> 'DEPORTE'<br>➢ TC_De Procedencia <> 'DEPORTE' y NE_Del Padre <> 'Ninguno Terminado' y Provincia = 'Holguín'<br>➢ TC_De Procedencia <> 'DEPORTE' y NE_Del Padre <> 'Ninguno Terminado' y Provincia <> 'Holguín' | 0.93<br>0.68<br>0.69<br>0.68 |
| 4 | ➢ T C_De Procedencia = 'IPUEC'<br>➢ T C_De Procedencia <> 'IPUEC'<br>➢ T C_De Procedencia <> ‘DEPORTE'<br>➢ T C_De Procedencia = ‘DEPORTE' | 0.87<br>0.72<br>0.72<br>0.48 |
| 3 | ➢ No tiene influencia ninguno de los factores analizados sobre la nota | 0.84 |
| 2 | ➢ T C_De Procedencia = IPUEC<br>➢ T C_De Procedencia <> IPUEC | 0.53 |

**Tabla 3**: Reglas obtenidas para la asignatura Programación I

| **Introducción a la Programación** | | |
|---|---|---|
| Nota | Reglas | Prob |
| 5 | ➢ No tiene influencia ninguno de los factores analizados sobre la nota | 0.78 |
| 4 | ➢ No tiene influencia ninguno de los factores analizados sobre la nota | 0.60 |
| 3 | ➢ Provincia = 'Ciego de Avila'<br>➢ Provincia <> 'Ciego de Avila' | 0.72<br>0.78 |
| 2 | ➢ TC_De Procedencia <> IPUEC<br>➢ TC_De Procedencia = IPUEC y NE_De la madre <> Universitario<br>➢ TC_De Procedencia = IPUEC and NE_De la madre = Universitario | 0.52<br>0.82<br>0.52 |

**Tabla 4**: Reglas obtenidas para la asignatura Introducción a la programación.

Al analizar los resultados obtenidos se comprobó que las variables que más influyen sobre los resultados académicos de los estudiantes en su primer curso en la Universidad; es el tipo de centro de procedencia y la provincia de origen

**Resumen de evaluación de los resultados**

A continuación se muestra una tabla con el por ciento estimado de cumplimiento del objetivo del negocio basado en los criterios de éxito.

| **Criterios de éxito del negocio** | **Cumplimiento estimado** |
|---|---|
| Obtener un modelo de conocimiento y comprobar que las conclusiones obtenidas son válidas o útiles | 100% |
| Desarrollar el caso de estudio utilizando las herramientas de SQL Server 2005 para minería de datos | 100% |
| Realizar un proyecto de KDD guiado por la metodología CRISP-DM y la documentación de cada una da las fases | 100% |
| Interpretar los resultados de la relación que existe entre la procedencia social o académica de los estudiantes y sus resultados académicos actuales | 100% |

**Tabla 5**: Estimado de cumplimiento de los criterios de éxito del negocio.

Se estima que fue cumplido el objetivo del negocio correspondiente al descubrimiento de patrones ocultos en los datos; que permitan predecir los resultados académicos de los estudiantes de la UCI, basado en las relaciones que se establecen entre Centro de Procedencia – Provincia – Nivel de escolaridad de los padres, con las Notas de las asignaturas recibidas en el primer año de la carrera.

### 6- Despliegue

Los modelos y reglas obtenidas podrán ser utilizados por el Centro de Investigaciones por la Calidad de la Educación (CICE), por la Dirección de Formación Académica y en otras investigaciones sobre los resultados académicos de los estudiantes de la UCI. Con las relaciones y patrones encontrados se podrán trazar estrategias que permitan elevar la formación docente de los nuevos ingresos a la Universidad, de acuerdo a las características propias de los estudiantes.

# CONCLUSIONES

Con la realización del presente trabajo se desarrolló un proyecto de minería de datos guiado por la metodología CRISP-DM, para determinar la relación que existente entre la procedencia del origen social y los resultados académicos en los estudiantes de la UCI. Se construyeron, entrenaron y evaluaron los modelos de Clustering o agrupamiento y de Árboles de Decisión para obtener las reglas y patrones ocultos en los datos. Se obtuvieron modelos de predicción precisos que logran reglas con alto valor de certeza y que permiten caracterizar los datos analizados y diseños de prueba eficientes para proceder con posteriores análisis.

# RECOMENDACIONES

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

- Continuar la investigación a partir de los resultados obtenidos, siguiendo las orientaciones de la fase de Evaluación, guiado por la metodología CRISP-DM.
- Fomentar el desarrollo de proyectos de Descubrimiento de Conocimiento en Bases de Datos en la Universidad de las Ciencias Informáticas

# REFERENCIAS


[1] Ian H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 2000.

[2] Berthold, M.; Hand, D.J. (eds.) *Intelligent Data Analysis. An Introduction,* Springer, 2ndEdition, 2003.

[3] Orallo Hernández, J.:Quintana Ramírez, Ma. J..:Ramírez Ferri, C.:Introducción a la Minería de Datos. Prentice Hall, 2004

[4] Fayyad, U. M., Piatetsky-Shapiro, G., Smith, P., Uthurusamy R.: Advances in Knowledge Discovery and Data-Mining, AAAI Press / The MIT Press, 1996.

[5] Crivat, B.: SQL Server Data Mining Programmability. . URL: http://msdn.microsoft.com/sql/bi/dmining/default.aspx?pull=/library/en-us/dnsql90/html/sqldmprgrm.asp. Fecha de Acceso: Dic 12, 2006.

[6] Iyer, Raman and Crivat, Bogdan SQL Server Data Mining: Plug-In Algorithms. . Fecha de Acceso: Dic 13, 2006 URL: http://msdn.microsoft.com/sql/bi/dmining/default.aspx?pull=/library/en-us/dnsql90/html/ssdmpia.asp.

[7] MacLennan, J.: Unearth the New Data Mining Features of Analysis Services 2005.; development lead for the Data Mining engine in the SQL Server 2005. MSDN Magazine, September 2004. URL: http://msdn.microsoft.com/msdnmag/issues/04/09/AnalysisServices2005/. Fecha de Acceso: Dic 13, 2006.

[8] Netz, A.; SQL Server 2000: Data Mining Helps Customers Make Better Business Decisions. Interviewed Netz, Amir; Microsoft SQL Server Development Manager. URL: http://www.microsoft.com/presspass/features/2000/04-24sql.mspx. Fecha de Acceso: Dic 15, 2006.

[9] Tang, L. and Bradley, P...AMO Lets You Dig Deeper into Your Data from Your Own Applications, MSDN Magazine, June 2005. URL: http://msdn.microsoft.com/sql/bi/dmining/default.aspx?pull=/msdnmag/issues/05/06/am_o/toc.asp. Fecha de Acceso: Dic 15, 2006.

[10]. Tang, Z., MacLennan J.: Data Mining with SQL Server, ISBN-10: 0-471-46261-6.

[11] Chapman, P.: Clinton, J.: Kerber, R.: Khabaza, T.: Reinartz, T.: Shearer, C.: Wirth, R.: CRISP-DM 1.0 Step-by-step data mining guide, 1999.


# AUTORES

En esta última sección se incluirá una breve reseña del autor o autores que debe incluir:

**Nombre y Apellidos,** título Universitario, Grado científico, categoría docente y/o de investigador, centro de trabajo, dirección postal, número de teléfono y FAX, correo electrónico y una nota breve sobre su labor actual.