Asesorías y Tutorías para la Investigación Científica en la Educación Puig-Salabarría S.C.
José María Pino Suárez 400-2 esq a Lerdo de Tejada. Toluca, Estado de México. 7223898475
RFC: ATI120618V12

**Revista Dilemas Contemporáneos: Educación, Política y Valores.**
**http://www.dilemascontemporaneoseducacionpoliticayvalores.com/**
**Año: V. Número: 3 Artículo no.: 13 Período: 1ro de mayo al 31 de agosto del 2018.**

**TÍTULO:** Análisis de datos de estudiantes de ingeniería para la predicción del rendimiento académico mediante técnica de clasificación bayesiana.

**AUTORES:**

1. Máster. Andrés Rico Páez.

2. Dr. Daniel Sánchez Guzmán.

**RESUMEN:** En este trabajo se presenta un estudio para analizar datos de estudiantes de ingeniería para desarrollar un modelo predictivo del rendimiento académico por medio de la técnica de clasificación bayesiana conocida como Bayes Ingenuo con el objetivo de predecir el rendimiento académico e identificar los principales factores que inciden en éste. Para esto, se analizaron datos correspondientes a 306 estudiantes de 7 cursos distintos. A partir de los resultados obtenidos, se implementó un sistema predictor que permite la predicción automática del rendimiento académico de futuros estudiantes de dicha institución. Este estudio puede ser replicado para diferentes estudiantes y cursos con el objetivo de que las instituciones educativas puedan diseñar estrategias de identificación y prevención de reprobación.

**PALABRAS CLAVES:** modelo predictivo, algoritmo Bayes Ingenuo, rendimiento académico, atributo, minería de datos.

**TITLE:** Analysis of data of engineering students for the prediction of academic performance using the bayesian classification technique.

**AUTHORS:**

1. Máster. Andrés Rico Páez.

2. Dr. Daniel Sánchez Guzmán.

**ABSTRACT:** In this paper we present a study to analyze data of engineering students to develop a predictive model of academic performance through the bayesian classification technique known as Naïve Bayes with the aim of predicting academic performance and identifying the main factors that affect it. For this, data corresponding to 306 students from 7 different courses were analyzed. Based on the results obtained, a predictive system was implemented that allows the automatic prediction of the academic performance of future students of the mentioned institution. This study can be replicated for different students and courses with the objective that educational institutions can design strategies for identification and prevention of reprobation.

**KEY WORDS:** predictive model, Naïve Bayes algorithm, academic performance, attribute, data mining.

## INTRODUCCIÓN.

Existe un rápido crecimiento en el uso y manejo de tecnologías de la información y comunicación en un gran número de áreas. Esto ha provocado un incremento en la cantidad de información a almacenar, que generalmente, no se analiza, aunque pueda contener algún tipo de conocimiento potencialmente útil. En años recientes, el uso de técnicas de minería de datos ha sido utilizada para analizar información en diferentes sectores, principalmente los empresariales y comerciales, debido a que los patrones extraídos del conjunto de datos no son, típicamente, una parametrización

de ningún modelo preestablecido o intuido por el usuario, sino que es un modelo novedoso y original (Hernández, Ramírez y Ferri, 2004). De manera similar, existe una tendencia a utilizar este tipo de técnicas de extracción de conocimiento o de minería de datos en el área educativa (Romero y Ventura, 2010, 2012; Peña, 2014), esto debido al potencial para descubrir conocimiento útil que beneficie los procesos de enseñanza y aprendizaje; sin embargo, esta aplicación de la minería de datos es reciente, principalmente en países de Latinoamérica, tales como México (Estrada, Zamarripa, Zúñiga y Martínez, 2016).

La minería de datos aplicada a la educación o minería de datos educativos (*Educational Data Mining*, EDM) surge como un paradigma orientado al diseño, tareas, métodos y algoritmos con el propósito de descubrir conocimiento y patrones dentro de los datos, y realizar predicciones de resultados o comportamientos de los estudiantes (Ballesteros y Sánchez, 2013; Luan, 2002).

La predicción del rendimiento académico es una de las aplicaciones más populares de la EDM (Romero y Ventura, 2010, 2012; Peña, 2014), debido a que el desempeño académico es uno de los principales índices de calidad académica de las instituciones educativas (Shahiri, Husain y Rashid, 2015), de hecho, la mayoría de los trabajos de EDM que tratan acerca del fracaso escolar han sido, principalmente, en educación superior (Kotsiantis, 2009). Esto se debe a que la reprobación, en este nivel educativo, propicia la incorporación tardía de los jóvenes en el ámbito laboral y una pérdida de recursos económicos para la institución y el país (Amado, García, Brito, Sánchez y Sagaste, 2014).

Algunos de los beneficios potenciales que ofrece la predicción de rendimiento académico son proponer programas de prevención estratégicos para estudiantes vulnerables a reprobar o desertar, identificar características de los estudiantes que le permiten obtener un buen desempeño académico, entre muchos otros.

Los resultados obtenidos con este tipo de técnicas han sido prometedores y demuestran como algunos factores o características de los estudiantes pueden afectar el rendimiento académico (Márquez, Romero y Ventura, 2012); por lo que existen varias líneas de investigación abiertas sobre el uso y desarrollo de este tipo de técnicas en educación.

Una técnica bayesiana, considerada una de las técnicas de minería de datos más utilizada, es el algoritmo Bayes Ingenuo (Hernández *et al.*, 2004; Witten, Frank y Hall, 2005), la cual ha mostrado una exactitud en las predicciones semejante o superior al de otras técnicas de minería de datos (Michie, Spiegelhalter y Taylor, 1994; Kotsiantis, Pierrakeas y Pintelas, 2003). El objetivo de esta investigación es analizar datos de estudiantes de ingeniería para desarrollar un modelo predictivo del rendimiento académico basado en el algoritmo Bayes Ingenuo con el propósito de predecir su rendimiento académico e identificar los principales factores que inciden en éste.

La extracción de conocimiento a partir de datos no solo requiere de la aplicación directa de las técnicas de minería de datos, es necesario un proceso más amplio que incluye a la minería de datos como una etapa del mismo. Este proceso se describe a continuación.

## DESARROLLO.

### Descubrimiento de conocimiento en bases de datos.

El proceso completo de aplicación de técnicas de minería de datos es conocido como descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Databases*, KDD) (Hernández *et al.*, 2004), y coloca a la minería de datos como una de las etapas del mismo. El proceso KDD se muestra en la Figura 1.

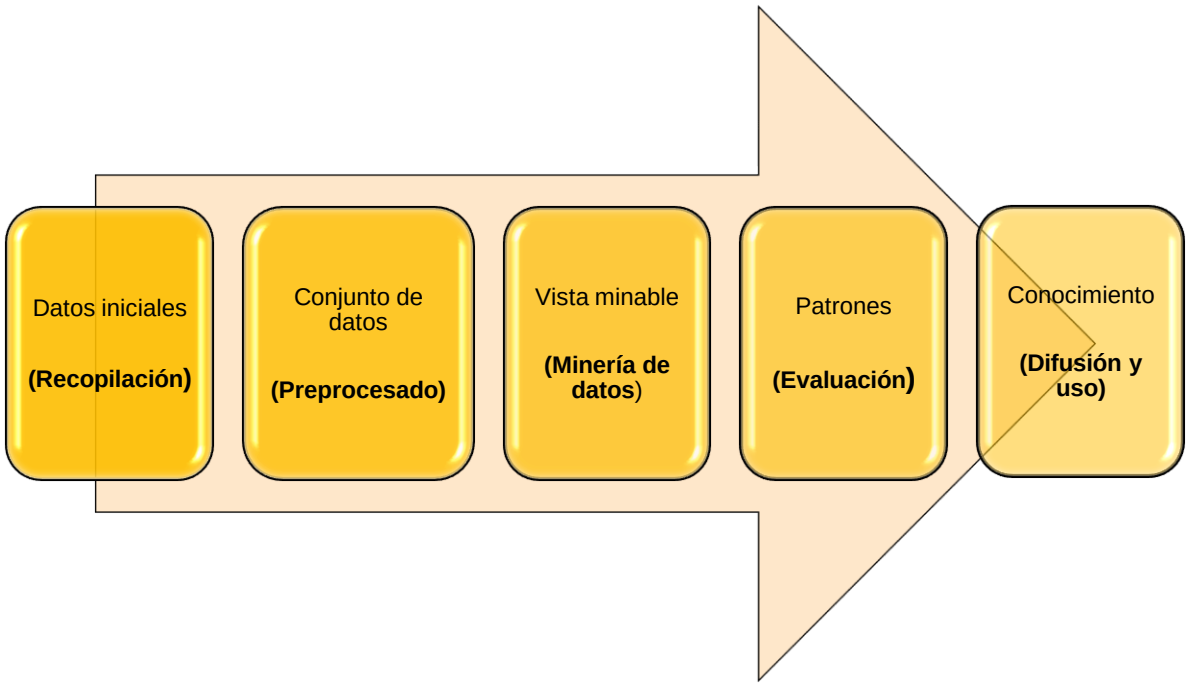


Figura 1. Proceso KDD.

La primera etapa del proceso KDD es la recopilación de datos, en la cual se determina el tipo de fuentes de información y la manera de conseguirla. La siguiente etapa es el preprocesado, en la cual los datos originales se transforman a una forma más adecuada para ser utilizados por la técnica de minería de datos en particular. Posteriormente, sigue la etapa de minería de datos en la que define el tipo de tarea a realizar y el algoritmo a implementar. A continuación, está la etapa de evaluación, en donde se determina la validez y confiabilidad de los patrones obtenidos, los cuales representan el conocimiento extraído. Finalmente, se encuentra la etapa de difusión y uso, en la que se hace partícipe a los usuarios del conocimiento obtenido. En la siguiente sección se presenta la técnica de minería utilizada en esta investigación.

**Técnica bayesiana para clasificación: Algoritmo Bayes Ingenuo.**

Dentro de las tareas de minería de datos educativa, las predictivas son las de uso más extendido (Romero y Ventura, 2010, 2012; Peña, 2014) debido a que permite detectar problemas académicos con anticipación y tomar las decisiones más adecuadas. La clasificación es una tarea predictiva de minería de datos que consiste en predecir la clase de nuevos registros, llamados datos de prueba, a partir de registros que tienen una clase conocida, llamados datos de entrenamiento. La clase se representa mediante el valor de una variable o atributo conocido como clasificador. De manera más específica, existe un conjunto de atributos $\{A_1,\ldots,A_n\}$ y una variable de clase $C_i$ perteneciente a un conjunto $\Omega_C=\{C_1,\ldots,C_k\}$. El algoritmo Bayes Ingenuo supone "ingenuamente" que todos los atributos son independientes, una vez conocido el valor de la clase. El valor de la clase a devolver, en base a esta suposición, es:

$$C= \arg\max_{C_i \in \Omega_C} P(C_i) \prod_{j=1}^{n} P(A_j|C_i) \qquad \textbf{Fórmula 1.}$$

La clasificación consta de dos partes: la primera es la construcción del modelo y la segunda es la evaluación del modelo a partir de la clasificación de nuevos registros.

Para la construcción del modelo se estiman las probabilidades *a priori* y *a posteriori*. Las probabilidades *a priori* $P(C_i)$ se estiman dividiendo el número registros de la clase $C_i$ de los datos de entrenamiento entre el total de los mismos; es decir, se considera que todos los valores de clase igualmente probables. La estimación de la probabilidades *a posteriori* $P(A_j|C_i)$ de cada atributo discreto se calculan a partir de la frecuencia de aparición en la base de datos de entrenamiento por medio del número de casos favorables entre el número de casos totales.

En este trabajo, para solucionar el caso en el que $P(A_j|C_i)=0$, se utiliza la estimación basada en la ley de sucesión de Laplace (Hernández *et al.*, 2004), la cual consiste en obtener el número de casos favorables más uno dividido entre el número de casos totales más el número de valores posibles.

Una vez construido el modelo predictivo se evalúa clasificando nuevos registros; para esto, se determinan las probabilidades de los atributos de cada nuevo registro y se aplica la fórmula 1 para determinar la clase a la que corresponde.

En la siguiente sección, se describe la metodología de análisis de datos utilizada en este trabajo, la cual está basada en el proceso KDD y en la implementación de la técnica de minería de datos Bayes Ingenuo.

**Metodología de análisis de datos utilizada en este trabajo.**

La metodología empleada se basa en los pasos típicos del proceso KDD y se muestra esquemáticamente en la Figura 2.

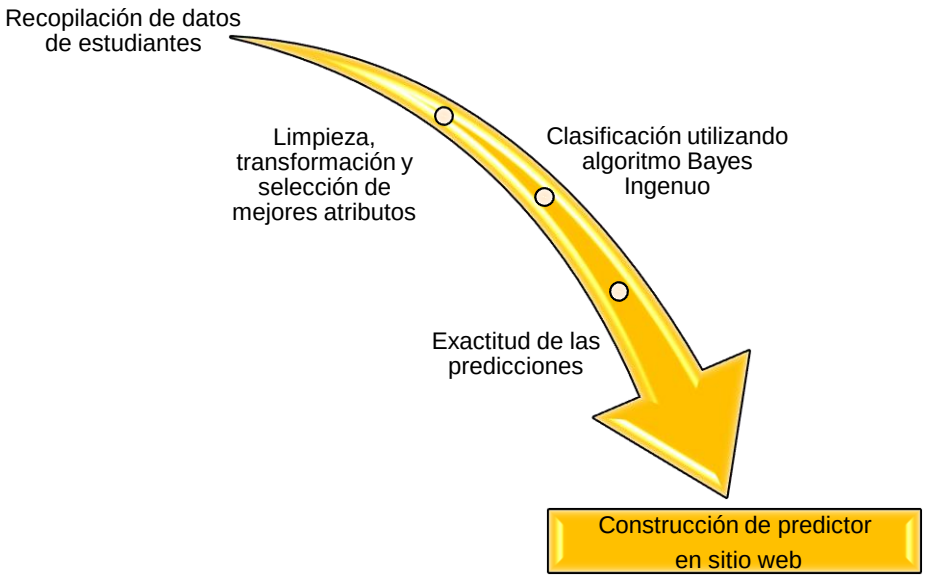


Figura 2. Metodología utilizada para la predicción del rendimiento académico.

Primeramente, se recopilan los datos de estudiantes a partir de diferentes fuentes y se integran en un conjunto de datos. Posteriormente, los datos se limpian y transforman para la etapa de minería de datos. Además, se pueden aplicar técnicas selección de atributos, tales como las que ordenan los atributos por su relevancia en la predicción de la clase (Cobo, Rocha y Álvarez, 2011), lo cual permite seleccionar los mejores atributos. La siguiente etapa es la de minería de datos, en este trabajo se utiliza la clasificación por medio del algoritmo Bayes Ingenuo. A continuación, sigue la etapa de evaluación, en donde se obtiene la exactitud de las predicciones realizadas. Por último, en la etapa de difusión y uso se propone realizar un sistema predictor en un sitio web similar a (Valero, Salvador y García, 2010) con el propósito de apoyar a los profesores para prevenir la reprobación estudiantil. A continuación, se describe un caso de estudio realizado con datos recabados de estudiantes de educación superior, específicamente, de estudiantes de los primeros semestres de una carrera de ingeniería, en donde los problemas de deserción y reprobación son frecuentes y de mayor impacto (Amado *et al.*, 2014).

**Recopilación de datos.**

La muestra de datos corresponde a 306 estudiantes de 7 cursos del primer y segundo semestre de ingeniería de una institución perteneciente al Instituto Politécnico Nacional. Es importante destacar, que con una cantidad parecida de registros, el Algoritmo Bayes Ingenuo ha obtenido una exactitud parecida o superior a otras técnicas de minería de datos en trabajos similares (Kotsiantis *et al.*, 2003; Osmanbegović y Suljić, 2012; Mueen, Zafar y Manzoor, 2016). La información de aprobación y reprobación de los estudiantes fue proporcionada por los docentes de la institución y las demás variables o atributos asociados al rendimiento académico fueron recopilados por medio de una encuesta a los estudiantes. Todos los datos fueron integrados en una tabla en formato electrónico.

**Preprocesado de datos.**

En esta etapa se transforman los datos de tal forma que puedan ser manipulados por la técnica de minería de datos a utilizar. Para esto, los atributos que no eran categóricos se les asignaron valores nominales como se muestra en la Tabla 1. De esta manera se dispone de una tabla de 306 registros de estudiantes (filas) y 21 atributos (columnas), de los cuales, el atributo "aprueba" define la etiqueta de la clase.

Tabla 1. Atributos de los estudiantes con sus posibles valores.

| Atributos | Valores posibles |
|---|---|
| Edad | Entre 18 y 19 años, Más de 20 años |
| Tipo de curso | Calculo Diferencial e Integral, Química Básica, Humanidades 1, Fundamentos de Álgebra, Física Clásica, Programación, Ecuaciones Diferenciales |
| Lugar de nacimiento | Ciudad de México, Interior de la República Mexicana |
| Escolaridad del padre | Básica, Media superior, Superior o mayor |
| Escolaridad de la madre | Básica, Media superior, Superior o mayor |
| Ingreso familiar | Menos de $6000, $6000 - $12000, Más de $12000 |
| Promedio obtenido en la media superior | 0 - 7.4, 7.5 - 8.4, 8.5 – 10 |
| Tiempo de traslado a la escuela | Menos de 60 min., 60 min. - 100 min., Más de 100 min. |
| Cantidad de cursos reprobados actualmente | 0, 1, 2 o más |
| Promedio actual | 0 - 7.4, 7.5 - 8.4, 8.5 – 10 |
| Apoyo familiar en sus estudios | Regular , Excelente |
| Nivel de inglés | Básico, Intermedio, Avanzado |
| Preferencia de estudio | Solo, En dúo, En grupo |
| Preferencia para realizar actividades en clase | Solo, En dúo, En grupo |
| Frecuencia de estudio | Continuamente, Una semana antes del examen, Un día antes del examen |
| Número de personas con quien vive | 1 – 3, 4, Más de 5 |
| Beca | SI, NO |
| Preferencia de la forma de estudio | Apuntes, Libros, Recursos de internet |
| Relación con los compañeros de clase | Regular , Excelente |
| Practica deporte | SI, NO |
| Aprobación del curso (aprueba) | SI, NO |

El conjunto de datos obtenido tiene un desbalanceo de clases considerable; es decir, el número de registros de estudiantes de una clase es mucho mayor (clase mayoritaria) que el número registros de la otra clase (clase minoritaria). En este caso, de 306 estudiantes, 256 aprobaron y 50 reprobaron. El problema de utilizar datos demasiado desbalanceados es que la técnica de minería de datos a utilizar tiende a clasificar los datos de prueba con baja sensibilidad a los elementos de la clase minoritaria. Una forma de resolver este problema es haciendo un sobre muestreo o balanceo de la distribución de clases mediante el muestreo aleatorio estratificado (Hernández *et al.*, 2004), el cual consiste en adicionar muestras aleatorias de las clases mayoritarias en las clases minoritarias. En este caso, se obtuvieron 100 muestras aleatorias sin reemplazamiento de la clase mayoritaria para adicionarlas a la clase minoritaria.

Posteriormente, se realizó un estudio de selección de atributos para determinar cuáles son los más relevantes para predecir el atributo "aprueba"; es decir, para seleccionar los mejores atributos. Como apoyo para realizar algunos de los experimentos de este estudio se utilizó el software WEKA (*Waikato Environment for Knowledge Analysis*) (Ferrari y Mariño, 2014).

En la selección de atributos se utilizó el método de ganancia de información de los atributos (Martín, Ramos, Grau, y García, 2007). Consiste en seleccionar los atributos con mayor ganancia de información con respecto a la clase a predecir. Entre más grande sea la ganancia de información de un atributo mayor será su influencia sobre la clase; en este caso, sobre el atributo "aprueba". La función de ganancia de información se encuentra implementada en el software Weka por medio de la función *InfoGainAttributeEva* (Ferrari y Mariño, 2014). Con apoyo de esta función, se calculó la ganancia de información de cada uno de los 20 atributos de los datos analizados. Estos valores se muestran de manera ordenada en la Figura 3 con el propósito de identificar de manera más sencilla los mejores atributos para predecir la aprobación de estudiantes de la muestra de datos.

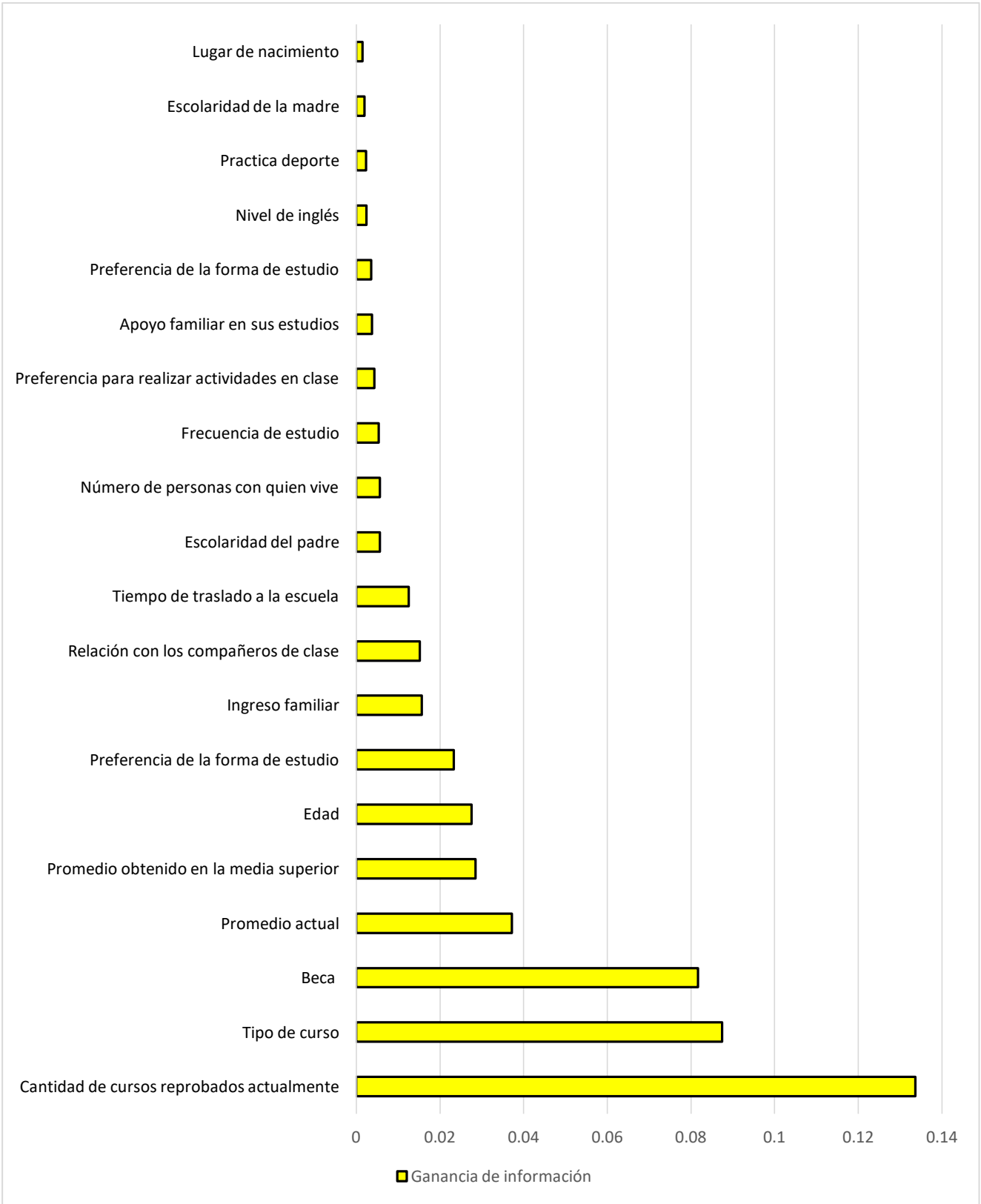


Figura 3. Ganancia de información de los atributos.

**Minería de datos y evaluación.**

Una vez realizadas las tareas de preprocesado en la sección anterior, se obtienen los datos que servirán de entrenamiento para aplicarles la técnica predictiva de minería de datos y construir el modelo. En este trabajo, la tarea predictiva empleada es la clasificación y la técnica utilizada es el Algoritmo Bayes Ingenuo debido a su uso y efectividad en la exactitud de las predicciones con cantidades de datos de entrenamiento similares a las utilizadas en este estudio (Shahiri *et al.*, 2015; Kavipriya, 2016).

En la siguiente sección se describen los experimentos realizados para la obtención de los modelos de predicción de aprobación de los estudiantes al final del semestre.

**Resultados y discusión.**

Para cada experimento de esta sección, se evalúa el modelo predictivo por medio del cálculo de la exactitud de las predicciones (porcentaje de la cantidad de registros con predicciones correctas entre el total de registros), utilizando el método conocido como validación cruzada (Hernández *et al.*, 2004). Este método consiste en dividir aleatoriamente el total de los datos de entrenamiento. En este trabajo, se dividió en dos conjuntos equitativos. Se construye un modelo con el primer conjunto y se usa para predecir los resultados en el segundo conjunto para calcular su exactitud. Después, se construye un modelo con el segundo conjunto y se usa para predecir los resultados del primer conjunto para calcular su exactitud. Finalmente, se calcula la exactitud del modelo construido, promediando las exactitudes calculadas anteriormente.

El propósito de los experimentos es seleccionar la mayor cantidad de atributos que propicien el valor más alto de exactitud. Con estos atributos, se construirá un predictor automático del rendimiento académico para futuros estudiantes.

En el primer experimento se consideran todos los 20 atributos de registros de estudiantes. Se construye el modelo utilizando el Algoritmo Bayes Ingenuo y se calcula la exactitud de las predicciones, utilizando la validación cruzada, la cual fue de 69.4581% (Figura 4).

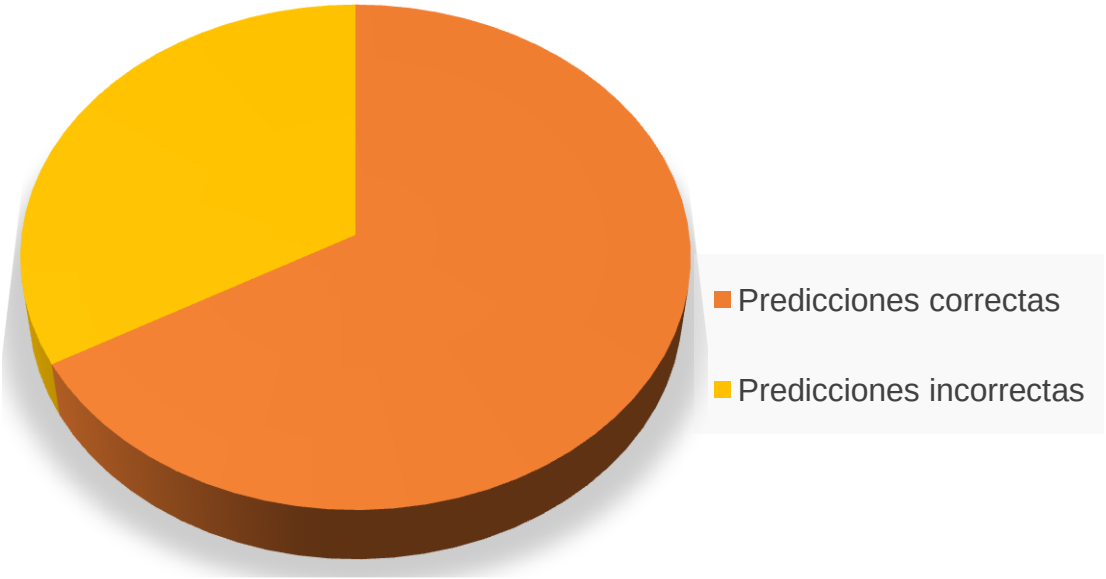


Figura 4. Exactitud de las predicciones considerando todos los atributos.

En el segundo experimento se utiliza para la evaluación solo los 15 mejores atributos del conjunto de datos de acuerdo al orden mostrado en la Figura 3.

Una cantidad similar de mejores atributos se usó en Márquez *et al.*, (2012). La exactitud de las predicciones obtenida con la validación cruzada empleando los 15 mejores atributos es de 71.4286% (Figura 5).

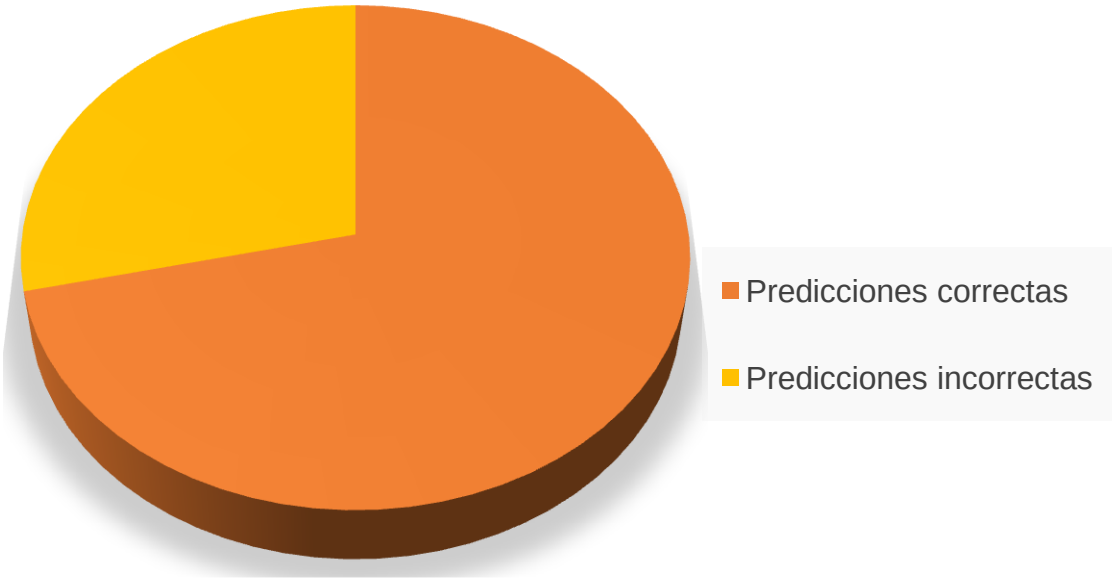


Figura 5. Exactitud de las predicciones considerando los 15 mejores atributos.

Los resultados muestran que seleccionando los 15 mejores atributos se obtiene una exactitud superior que la obtenida empleando todos los 20 atributos; no obstante, pueden existir otras cantidades de mejores atributos que propicien mejores valores de exactitud; por lo que a diferencia de Márquez *et al*., (2012), en este trabajo se aborda de forma más general el problema de selección de atributos, considerando diferentes cantidades de mejores atributos para calcular la exactitud. De esta manera, se plantea el tercer experimento, que consiste en calcular la exactitud, considerando el mejor atributo, después considerando los 3, 6, 9, 12, 15, 18 mejores atributos de acuerdo al ordenamiento mostrado en la Figura 3. Estos valores de exactitud con sus respectivas cantidades de mejores atributos se muestran en la Figura 6.

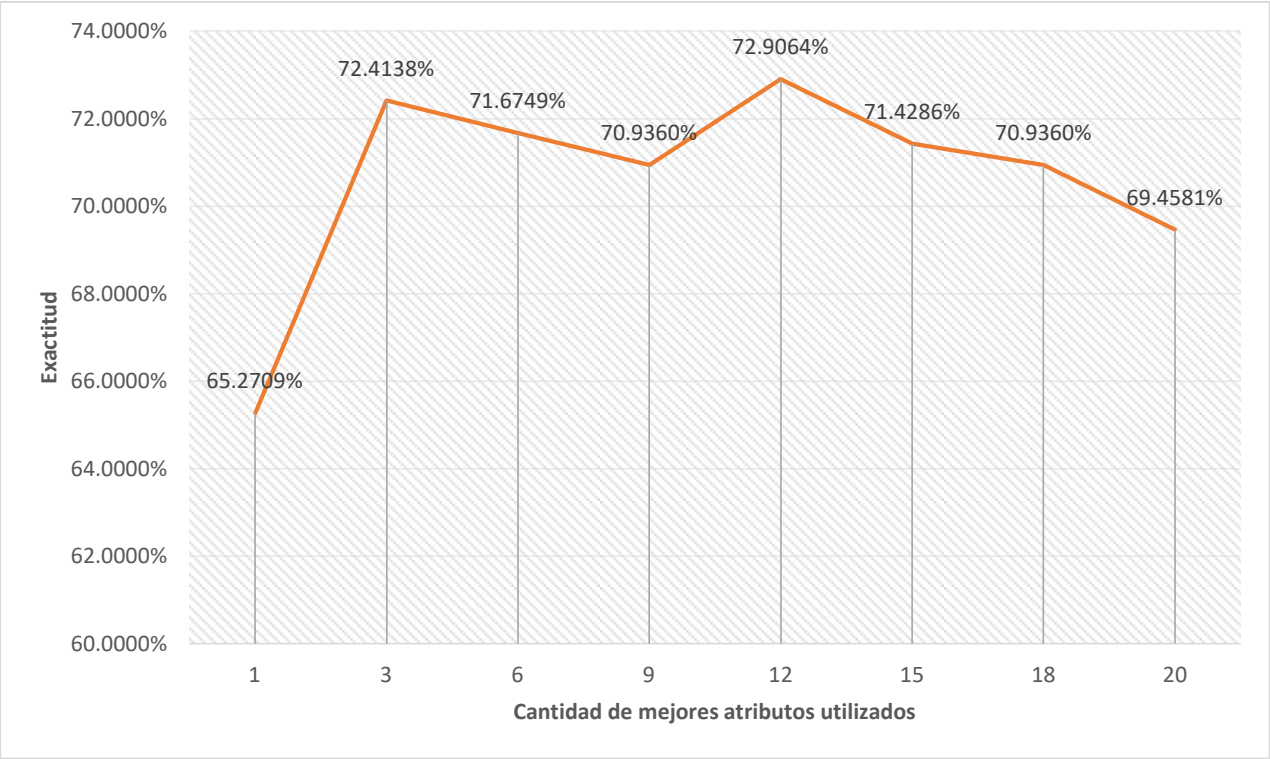


Figura 6. Exactitud vs cantidad de mejores atributos de los datos utilizados.

En esta gráfica, el valor más bajo de exactitud (65.2709 %) ocurre cuando se utiliza únicamente el mejor atributo debido a que no es suficiente la información de un único atributo para obtener predicciones adecuadas. El valor de exactitud más alto (72.9064 %) se obtuvo cuando se utilizan los mejores 12 atributos (más del 50% del total de atributos)

Utilizando los 12 mejores atributos se construyó un sistema predictor que permite la predicción automática del rendimiento académico de futuros estudiantes. Este predictor se programó utilizando el Algoritmo Bayes Ingenuo en HTML5 (*HyperText Markup Language*, versión 5) y PHP (*Hypertext Pre-Processor*) con el objetivo de ser publicado en un sitio web. La interfaz gráfica del predictor se muestra en la Figura 7. Los datos de entrada del predictor son los 12 atributos seleccionados en el estudio anterior pertenecientes al estudiante a predecir su aprobación o reprobación del curso especificado en el atributo “Tipo de curso”. El predictor proporciona como salida si el alumno aprueba o reprueba el curso y la probabilidad de reprobación.

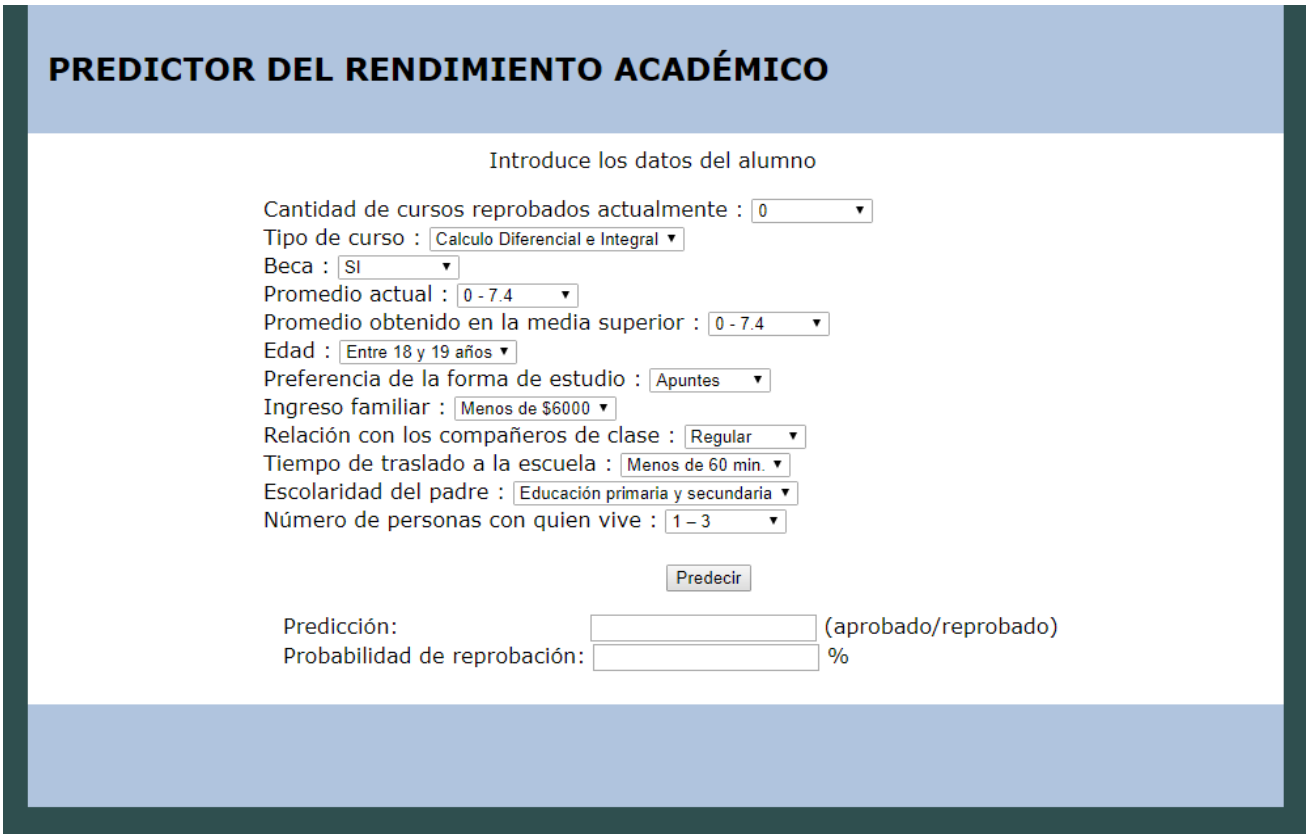


Figura 7. Interfaz gráfica del sistema predictor resultado del estudio realizado.

Para mostrar la utilidad del sistema predictor, se aplicó a 20 estudiantes del curso de Ecuaciones Diferenciales de la misma institución de educación superior de donde se recopilaron los datos de entrenamiento. Se obtuvieron las predicciones de aprobación y las probabilidades de reprobación de cada estudiante. Posteriormente, se recopilaron los resultados reales de aprobación obtenidos por los estudiantes al final del curso. Estos valores se presentan en la Figura 8.

| Cantidad de cursos reprobados actualmente | Beca | Promedio actual | Promedio obtenido en la media superior | Edad | Preferencia de la forma de estudio | Ingreso familiar | Relación con los compañeros de clase | Tiempo de traslado a la escuela | Escolaridad del padre | Número de personas con quien vive | Predicción de aprobación | Probabilidad de reprobación (%) | Aprobación real |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | SI | 7.5 - 8.4 | 7.5 - 8.4 | Entre 18 y 19 años | Apuntes | Menos de $6000 | Excelente | Menos de 60 min. | Básica | 1 – 3 | SI | 0.9 | SI |
| 1 | NO | 0 - 7.4 | 7.5 - 8.4 | Más de 20 años | Recursos de internet | Menos de $6000 | Excelente | 60 min. - 100 min. | Básica | 1 – 3 | NO | 78.8 | NO |
| 2 o más | NO | 8.5 – 10 | 7.5 - 8.4 | Más de 20 años | Apuntes | $6000-$12000 | Regular | Menos de 60 min. | Básica | Más de 5 | SI | 33 | SI |
| 1 | NO | 0 - 7.4 | 0 - 7.4 | Entre 18 y 19 años | Recursos de internet | Más de $12000 | Excelente | 60 min. - 100 min. | Superior o mayor | 1 – 3 | NO | 89.9 | NO |
| 0 | SI | 7.5 - 8.4 | 8.5 - 10 | Más de 20 años | Libros | Más de $12000 | Excelente | Menos de 60 min. | Media superior | 1 – 3 | SI | 7 | SI |
| 2 o más | NO | 0 - 7.4 | 0 - 7.4 | Más de 20 años | Recursos de internet | Menos de $6000 | Regular | 60 min. - 100 min. | Básica | 4 | NO | 95.9 | NO |
| 2 o más | NO | 0 - 7.4 | 0 - 7.4 | Entre 18 y 19 años | Libros | Más de $12000 | Excelente | 60 min. - 100 min. | Superior o mayor | 1 – 3 | NO | 95.8 | SI |
| 1 | NO | 0 - 7.4 | 7.5 - 8.4 | Más de 20 años | Libros | Más de $12000 | Regular | Más de 100 min. | Básica | 4 | NO | 82.2 | NO |
| 0 | NO | 0 - 7.4 | 7.5 - 8.4 | Entre 18 y 19 años | Libros | $6000-$12000 | Regular | Más de 100 min. | Superior o mayor | 4 | SI | 24.2 | SI |
| 2 o más | NO | 0 - 7.4 | 0 - 7.4 | Más de 20 años | Recursos de internet | Más de $12000 | Regular | Menos de 60 min. | Superior o mayor | 1 – 3 | NO | 98.4 | NO |
| 1 | NO | 0 - 7.4 | 7.5 - 8.4 | Entre 18 y 19 años | Apuntes | $6000-$12000 | Excelente | 60 min. - 100 min. | Media superior | 4 | SI | 39.8 | SI |
| 2 o más | NO | 0 - 7.4 | 0 - 7.4 | Más de 20 años | Libros | Menos de $6000 | Regular | Menos de 60 min. | Media superior | 4 | NO | 95.9 | NO |
| 0 | SI | 0 - 7.4 | 8.5 - 10 | Entre 18 y 19 años | Recursos de internet | Menos de $6000 | Excelente | Menos de 60 min. | Media superior | 1 – 3 | SI | 3.3 | SI |
| 0 | SI | 0 - 7.4 | 7.5 - 8.4 | Entre 18 y 19 años | Recursos de internet | Menos de $6000 | Excelente | 60 min. - 100 min. | Media superior | Más de 5 | SI | 5.7 | NO |
| 0 | SI | 8.5 – 10 | 8.5 - 10 | Entre 18 y 19 años | Apuntes | Menos de $6000 | Excelente | Menos de 60 min. | Básica | Más de 5 | SI | 0.1 | SI |
| 2 o más | NO | 0 - 7.4 | 0 - 7.4 | Entre 18 y 19 años | Recursos de internet | Menos de $6000 | Regular | Más de 100 min. | Media superior | 1 – 3 | NO | 92.1 | NO |
| 1 | NO | 0 - 7.4 | 0 - 7.4 | Más de 20 años | Libros | $6000-$12000 | Regular | 60 min. - 100 min. | Básica | 1 – 3 | NO | 91.5 | NO |
| 0 | NO | 8.5 – 10 | 0 - 7.4 | Más de 20 años | Recursos de internet | $6000-$12000 | Excelente | Menos de 60 min. | Superior o mayor | 1 – 3 | SI | 18.6 | NO |
| 0 | SI | 7.5 - 8.4 | 0 - 7.4 | Entre 18 y 19 años | Apuntes | $6000-$12000 | Excelente | Menos de 60 min. | Superior o mayor | Más de 5 | SI | 1.5 | SI |
| 2 o más | NO | 0 - 7.4 | 0 - 7.4 | Más de 20 años | Apuntes | $6000 - $12000 | Regular | Menos de 60 min. | Media superior | Más de 5 | NO | 89.2 | SI |

Figura 8. Predicciones y probabilidades de reprobación obtenidas con el sistema predictor.

En la Figura 8 se observan los valores de los atributos de cada estudiante del grupo de prueba. Se identificaron 6 estudiantes con alta probabilidad de reprobación (mayor al 90%). Además, se compararon las predicciones de aprobación con la aprobación real de los estudiantes al final del curso y se observa que 16 de las 20 predicciones son acertadas; es decir, se tuvo una exactitud del 80 %.

Se debe notar que el sistema predictor realizado no requiere que los usuarios tengan conocimientos profundos de minería de datos para utilizarlo como es el caso de otras herramientas informáticas que existen en la actualidad, tales como las usadas en (Jaramillo y Paz, 2015; Pacheco y Fernández, 2015). De esta forma, se espera que este predictor sea un apoyo a los profesores de la institución donde se recopilaron los datos para identificar y prevenir la reprobación de estudiantes.

De manera más general, la metodología presentada en este trabajo puede ser replicada para la construcción de predictores automáticos del rendimiento académico diseñados para datos de estudiantes de instituciones educativas específicas.

## CONCLUSIONES.

En este trabajo se presentó una metodología de análisis de datos basada en el proceso clásico de KDD, para implementar el algoritmo predictivo Bayes Ingenuo en estudiantes de ingeniería con el propósito de predecir su rendimiento académico e identificar los principales factores que inciden en éste. En este estudio, participaron 306 estudiantes de una carrera ingeniería como datos de entrenamiento y 20 estudiantes de la misma institución como datos de prueba.

Los atributos de los estudiantes se ordenaron en base a su influencia en el rendimiento académico. Los atributos con mayor relevancia fueron la cantidad de cursos reprobados, el tipo de curso, si tiene beca, el promedio actual y el promedio en media superior. Estos atributos están relacionados de manera directa con las calificaciones obtenidas por los estudiantes en los cursos anteriores.

Se realizaron experimentos para encontrar los mejores atributos que consigan una mayor exactitud en las predicciones. En los datos analizados se obtuvo este valore con los 12 mejores atributos. A partir de estos resultados, se implementó un sistema predictor que permite la predicción automática del rendimiento académico de futuros estudiantes de dicha institución. Se aplicó a 20

estudiantes de la misma institución en donde se identificaron aquellos estudiantes con mayor probabilidad de reprobación y se compararon las predicciones con los resultados reales de aprobación obtenidos por los estudiantes al final del curso, obteniendo una exactitud del 80%.

Este predictor fue realizado en lenguajes de programación adecuados para su posterior publicación en un sitio web. De esta manera, se espera que sea un apoyo a profesores de la institución educativa donde fueron obtenidos los datos para determinar el factor de riesgo de reprobación de manera oportuna, y así dar un seguimiento a los estudiantes que son vulnerables a reprobar.

El método utilizado puede ser emulado para analizar datos, identificar factores relevantes del rendimiento académico y construir predictores para otras instituciones educativas para diferentes estudiantes y diferentes tipos de cursos. Este tipo análisis ofrece la posibilidad a los profesores de identificar desde el inicio de sus cursos lo principales atributos del rendimiento académico y las probabilidades de reprobación de los estudiantes. Esto le permite diseñar estrategias de prevención y disminuir las estrategias de recuperación que impliquen que el estudiante repruebe alguna evaluación parcial para realizar algún tipo de intervención.

**REFERENCIAS BIBLIOGRÁFICAS:**

1. Amado, M. G., García, A., Brito, R. A., Sánchez, B. I. y Sagaste C. A. (2014). Causas de reprobación en ingeniería desde la perspectiva del académico y administradores. Ciencia y Tecnología, 14, 233-250. Recuperado de: http://www.palermo.edu/ingenieria/pdf2014/14/CyT_14_15.pdf
2. Ballesteros, A. y Sánchez, D. (2013). Minería de datos educativa: Una herramienta para la investigación de patrones de aprendizaje sobre un contexto educativo. Revista Latinoamericana de Física Educativa, 7(4), 662-668. Recuperado de: http://www.lajpe.org/dec13/22-LAJPE_814_bis_Alejandro_Ballesteros.pdf

3. Cobo, A., Rocha, R. y Álvarez, Y. (2011). Selección de atributos predictivos del rendimiento académico de estudiantes en un modelo de B-Learning. Revista Electrónica de Tecnología Educativa, 37, 1-13. doi: 10.21556/edutec.2011.37.390

4. Estrada, R. I., Zamarripa, R. A., Zúñiga, P. G. y Martínez I. (2016). Aportaciones desde la minería de datos al proceso de captación de matrícula en instituciones de educación superior particulares. Revista Electrónica Educare, 20(3), 1-21. doi: 10.15359/ree.20-3.11

5. Ferrari, S. y Mariño, S. (2014). Experiencia de personalización de la herramienta WEKA. Revista de Informática Educativa y Medios Audiovisuales, 11(8), 1-7.

6. Jaramillo, A. y Paz H. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje. Revista Tecnológica ESPOL, 28(1), 64-90. Recuperado de: http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/351/229

7. Hernández, J., Ramírez, M. y Ferri, C. (2004). *Introducción a la minería de datos.* Madrid, España: Pearson.

8. Kavipriya, P., (2016). A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 6(12), 101-105. doi: 10.1016/j.procs.2015.12.157

9. Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. En V. Palade, R. J. Howlett y L. Jain (Eds.), Lecture Notes in Computer Science: Vol. 2774. Knowledge-Based Intelligent Information and Engineering Systems (pp. 267–274). Heidelberg, Alemania: Springer-Verlag. doi: 10.1007/978-3-540-45226-3_37

10. Kotsiantis, S. B. (2009). Educational data mining: a case study for predicting dropout-prone students. International Journal of Knowledge Engineering and Soft Data Paradigms, 1(2), 101-111. doi: 10.1504/IJKESDP.2009.022718

11. Luan, J. (2002). Data Mining and Its Applications in Higher Education. New Directions for Institutional Research, 2002(113), 17-36. doi: 10.1002/ir.35

12. Márquez, C., Romero, C. y Ventura, S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *IEEE-RITA, 7*(3), 109-117. Recuperado de: http://rita.det.uvigo.es/201208/uploads/IEEE-RITA.2012.V7.N3.A1.pdf

13. Martín, R., Ramos, R. M., Grau, R. y García, M. M. (2007). Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños. Revista Gaceta Médica Espirituana, 9(1), 1-7.

14. Michie, D., Spiegelhalter D. y Taylor, C. (1994). Machine learning, neural and statistical classification. Nueva Jersey, EUA: Prentice Hall.

15. Mueen, A., Zafar, B. y Manzoor U. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. International Journal of Modern Education and Computer Science, 11, 36-42. doi: 10.5815/ijmecs.2016.11.05

16. Osmanbegović, E. y Suljić, M. (2012). Data mining approach for predicting student performance. Journal of Economics and Business, *10*(1), 3-12. Recuperado de: https://www.researchgate.net/publication/242341193_DATA_MINING_APPROACH_FOR_PREDICTING_STUDENT_PERFORMANCE

17. Pacheco, A. y Fernández, Y. (2015). Aplicación de técnicas de descubrimiento de conocimientos en el proceso de caracterización estudiantil. Ciencias de la Información, 46(3), 25-30. Recuperado de: http://www.redalyc.org/articulo.oa?id=181443340004

18. Peña, A. (2014). Review: Educational data mining: A survey and a data mining based analysis of recent works. Expert Systems with Applications, 41(4),1432-1462. doi: 10.1016/j.eswa.2013.08.042

19. Romero, C. y Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618. doi: 10.1109/TSMCC.2010.2053532

20. Romero, C. y Ventura, S. (2012). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27. doi: 10.1002/widm.1075

21. Shahiri, A., Husain, W. y Rashid, N. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414-422. doi: 10.1016/j.procs.2015.12.157

22. Valero, S., Salvador, A., y García, M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. En M. E. Prieto, J. M. Dodero y D. O. Villegas (Eds.), Lecture Notes in Computer Science: Vol. Kaambal. Recursos digitales para la educación y la cultura. (pp. 33-39). Mérida, México. Recuperado de: http://www.utim.edu.mx/~svalero/docs/e1.pdf

23. Witten, I., Frank, E. y Hall, M. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Massachusetts, EUA: Morgan Kaufmann Publishers.

**DATOS DE LOS AUTORES:**

1. **Andrés Rico Páez.** Maestro en Ciencias en la Especialidad de Ingeniería Eléctrica con opción en Comunicaciones por el Centro de Investigación y Estudios Avanzados, Unidad Zacatenco del Instituto Politécnico Nacional. Estudiante de Doctorado en Tecnología Avanzada en el Centro de Investigación de Ciencia Aplicada y Tecnología Avanzada, Unidad Legaria del Instituto Politécnico Nacional. Correo electrónico: aricop.ipn@gmail.com

2. **Daniel Sánchez Guzmán.** Doctor en Tecnología Avanzada y Profesor Titular en el Centro de Investigación de Ciencia Aplicada y Tecnología Avanzada, Unidad Legaría del Instituto Politécnico Nacional. Correo electrónico: dsanchez@ipn.mx

**RECIBIDO:** 27 de enero del 2018. **APROBADO:** 19 de febrero del 2018.