UNIVERSIDAD TÉCNICA DE AMBATO

FACULTAD DE CIENCIAS HUMANAS Y DE LA EDUCACIÓN

CARRERA DEDOCENCIA EN INFORMATICA

MODALIDAD PRESENCIAL

Proyecto de investigación previo a la obtención del Título de Licenciado en Ciencias de la Educación, Mención:

Informática y Computación

TEMA:

“APLICACIÓN DE LA TÉCNICA DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA**.”**

AUTOR: Victor Xavier Vicente Guerrero.

TUTOR: Ing. Blanca Rocio Cuji Chacha. Mg.

Ambato-Ecuador

2019 - 2020

A. PÁGINAS PRELIMINARES

Tema

"APLICACIÓN DE LA TÉCNICA DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA"

Aprobación.

CERTIFICA:

Yo, Ing. Blanca Cuji, Mg. CI. 1803127594, en calidad de Tutor del trabajo de Graduación o Titulación, sobre el tema "APLICACIÓN DE LA TÉCNICA DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA", desarrollado por el Sr. Vicente Guerrero Victor Xavier, estudiante de Licenciatura en Ciencias Humanas y de la Educación, mención Informática y Computación, considero que dicho informe investigativo reúne los requisitos técnicos, científicos y reglamentarios, por lo que autorizo la presentación del mismo ante el organismo pertinente, para ser sometido a la evaluación de la comisión calificadora designada por el H. Consejo Directivo.

Ing. Blanca Cuji, Mg.
C.I.: 1803127594

Autoría

Los criterios emitidos en el trabajo de investigación: "APLICACIÓN DE LA TÉCNICA DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA", los contenidos, ideas, análisis, conclusiones y propuesta son de exclusiva responsabilidad del autor de este trabajo de grado.

Vicente Guerrero Victor Xavier
C.I.: 110509786-7
AUTOR

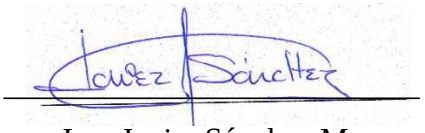
Aprobación del Tribunal de grado

AL CONSEJO DIRECTIVO DE LA FACULTAD DE CIENCIAS HUMANAS Y DE LA EDUCACIÓN:

La comisión de Estudio y Calificación del Informe del Trabajo de Graduación o Titulación, sobre el Tema: "APLICACIÓN DE LA TÉCNICA DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA". Presentado por El Sr. Vicente Guerrero Victor Xavier, egresado de la Carrera de Docencia en Informática, promoción 2020, una vez revisada y calificada la investigación, se APRUEBA debido a que cumple con los principios básicos técnicos y científicos de investigación y reglamentarios. Por lo tanto, se autoriza la presentación ante los organismos pertinentes.

LA COMISIÓN

Ing. Wilma Gavilanes Mg.



Ing. Javier Sánchez Mg.

Dedicatoria

A Dios por su infinita bondad que me ha permitido alcanzar los objetivos propuestos y me ha rodeado de excelentes personas que día a día me motivan a ser mejor.

A mi familia quienes han estado a mi lado, siempre creyendo en mí y dándome la fortaleza y apoyo emocional para seguir adelante en la búsqueda constante del crecimiento personal como profesional.

Victor Xavier Vicente Guerrero

Agradecimiento

En primer lugar, a Dios que es el pilar fundamental de mi vida, que ha iluminado mi camino en todo momento y me ha permitido alcanzar un objetivo más en mi vida.

A la Universidad Técnica De Ambato, Carrera de Docencia en Informática, por haberme abierto las puertas y guiado siempre en el proceso de formación profesional y a quienes de alguna manera han contribuido para poder llevar a cabo esta meta propuesta.

A mi familia, que con su ejemplo ha sabido inculcarme el deseo de constante superación, pero sobre todo por ese apoyo incondicional para permitirme alcanzar mis metas.

A mí tutor de tesis Ing. Mg. Blanca Cuji, por su constante apoyo durante todo el proceso de desarrollo del proyecto, por su calidad humana y profesional demostrada a cada momento.

Victor Xavier Vicente Guerrero

Índice general de contenidos

Índice de tablas.

Índice de figuras.

Resumen ejecutivo

Tema: Aplicación de la técnica de minería de datos para la predicción de la deserción estudiantil universitaria.
Autor: Victor Xavier Vicente Guerrero.
Tutor: Ing. Blanca Rocio Cuji Chacha. Mg.

Dado que la deserción es un fenómeno que afecta a estudiantes en la mayoría de instituciones educativas, se hace necesario aplicar nuevas estrategias de predicción de deserción, ante dicha necesidad, este estudio propone el uso de la minería de datos, ya que es una técnica que cada día adquiere mayor importancia en la predicción en variados ámbitos de la sociedad, mas en el ámbito educativo aún no se ha explorado, los beneficios que se podrían obtener de aplicar esta técnica, con el fin de predecir y evitar en lo posible la deserción estudiantil. Para la creación del modelo, se utiliza la metodología Knowledge Discovery in Databases (KDD) que consta de cinco etapas, en las cuales el primer paso es el análisis de los datos, se adecua los mismos y se les da el formato adecuado, mediante el uso del programa R y la función Loggit (función matemática de regresión logística), se aplica los procesos necesarios para la creación de un modelo de regresión logística y se hace las pruebas necesarias para verificar el correcto funcionamiento del modelo, encontrando a las variables relacionadas con las calificaciones de los niveles segundo, cuarto y quinto como las más influyentes del modelo. El estudio se enfoca en la línea de investigación de ciencias e ingeniería, por lo que puede tomarse como contribución para futuras investigaciones en el mismo campo, dada su relevancia por permitir la predicción y prevención de abandono estudiantil y por la implementación de nuevas técnicas en este proceso, lo que lo convierte en novedoso.

Palabras clave: Minería de datos, regresión, deserción, predicción, modelo.

Abstract

Theme: Application of the data mining technique for the prediction of university student desertion.

Author: Victor Xavier Vicente Guerrero.

Tutor: Ing. Blanca Rocio Cuji Chacha. Mg.

Since desertion is a phenomenon that affects students in most educational institutions, it is necessary to apply new desertion prediction strategies, in the face of this need, this study proposes the use of data mining, as it is a technique that increasingly becomes more important in predicting in various areas of society, but in the educational field has not yet been explored, the benefits that could be obtained from applying this technique, in order to predict and avoid as far as possible the desertion Student. For the creation of the model, Knowledge Discovery in Databases (KDD) methodology is used, which consists of five stages, in which the first step is the analysis of the data, they are adapted and given the appropriate format, through the use of the R program and the Loggit function (mathematical function of logistic regression), the necessary processes for the creation of a logistic regression model is applied and the necessary tests to verify the correct functioning of the model are done, finding the variables related to the qualifications of the second, fourth and fifth levels as the most influential of the model. The study focuses on the research line of science and engineering, so it can be taken as a contribution for future research in the same field, given its relevance for allowing the prediction and prevention of student dropout and the implementation of new techniques in this process, making it novel.

Keywords: Data mining, regression, dropout, prediction, model.

CAPITULO I
MARCO TEÓRICO

1.1.Antecedentes Investigativos

Revisadas diferentes bases de datos como: Scielo, Scopus y el repositorio de la Universidad Técnica de Ambato, se encontraron las siguientes investigaciones, que tienen referencia al tema planteado.

El abandono escolar, viene dado por varios factores, según Bonaldo y Pereira (2016), en su artículo "Dropout: Demographic Profile of Brazilian University Students ", las variables de mayor influencia en el abandono escolar son: género, edad, nivel de educación de la familia, desempeño académico del estudiante, estado civil, tipo de beca o financiación para estudios.

Martelo, Herrera, y Villabona (2017), por su parte en el artículo "Estrategias para disminuir la deserción universitaria mediante series de tiempo", analizan estrategias para prevenir la deserción estudiantil, mediante series de tiempo se analiza las variables relacionadas con el rendimiento académico, la variable principal es la deserción, sus resultados mostraron que aplicar series de tiempo, permiten determinar las variables más adecuadas a utilizar, para crear un modelo de predicción, además, se establece que las estrategias de creación de modelos flexibles ayudan de manera significativa a reducir el abandono en instituciones de educación superior, destacando una vez más, la importancia de la variable relacionada con el rendimiento académico.

Para Sivakumar, Venkataraman, y Selvaraj (2016), en su artículo "Predictive modeling of student dropout indicators in educational data mining using improved decision tree", quien busca identificar atributos relevantes de estudiantes de pregrado de la universidad de la India y desarrollar un algoritmo de árbol de decisión mejorado de predicción de abandono escolar, toma el algoritmo ID3 y árbol de decisión, analizando las variables residencia, tipo de familia, ingreso anual familiar, educación del padre, educación de la madre, ocupación del padre, ocupación de la madre, ubicación universitaria del estudiante, grado del estudiante, curso admitido, tipo de admisión, satisfacción con el curso, programa de estudios, variable que determina si los padres

cumplen con los gastos universitarios, experiencias familiares de estrés, características universitarias, las razones más altas de abandono están relacionadas con la familia y el entorno de estudio, mientras que pocos abandonan por problemas económicos o de salud, el modelo permite predecir el abandono gracias a la aplicación de minería de datos.

Por otro lado Kerby (2015), en su estudio "Toward a new predictive model of student retention in higher education: An application of classical sociological theory", sobre los factores de deserción estudiantil, encuentra que algunas razones de deserción son las dificultades económicas, las malas relaciones entre estudiantes y maestros y el gran tamaño de las clases, de igual manera influye el número de estudiantes. Mientras que en la investigación "Generating descriptive model for student dropout: a review of clustering approach", como generar un modelo descriptivo para la predicción de deserción estudiantil, mediante el uso de k-means, se encuentra que, un historial pasado de bajas calificaciones, influye directamente, en el abandono escolar, afectando negativamente no solo a los estudiantes sino a los padres y la universidad, por lo que se determina las calificaciones como un elemento imprescindible en la predicción de deserción (Iam-On y Boongoen, 2017),.

Para la selección de variables, Abbas, Sarker, Mahmood, Hasan, y Palaniappan (2015), en el artículo "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques", propone un marco para predecir el rendimiento académico, aplicando árboles de decisión, redes bayesianas y clasificación basada en reglas, analizan las variables grado de los estudiantes, carrera, género, ingreso familiar, modalidad de ingreso a la universidad y certificado de educación de Malasia encuentran que los algoritmos de clasificación pueden lograr la mayor precisión de predicción si utilizan más datos y si el conjunto de datos preparado no contiene información ruidosa o incompleta. Mientras que para, Ricard y Pelletier (2016), en el artículo "Dropping out of high school: The role of parent and teacher self-determination support, reciprocal friendships and academic motivation", las amistades de los estudiantes, apoyo de los padres para las necesidades psicológicas básicas y apoyo docente son variables predictoras significativas, resultando de menos importancia en el estudio variables como la motivación académica.

Goswami y Chakrabarti (2015), investigan en su artículo "Feature Selection: A Practitioner View", la eficiencia de los algoritmos Naive Bayes, chi-cuadrado y fs-chiclust, encontraron que Naive Bayes combinada con FS-CHICLUST produce contenidos más acertados y se ejecuta en menor tiempo. Según las investigaciones de Strecht, Cruz, Soares, Mendes-Moreira, y Abreu (2015), en el artículo "A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance", al evaluar algoritmos reconocidos de predicción mediante minería como k-Nearest, vecinos cercanos, bosque aleatorio, AdaBoost, árboles de clasificación y regresión (CART), máquinas de vectores de soporte, determina que la mayoría de algoritmos, puede lograr resultados acertados, mientras se tome en cuenta las variables correctas y se cree un modelo adecuado, por lo que distintos modelos deben ser probados previo a la elección del adecuado.

Daud et al. (2018), en el artículo "Predicting Student Performance using Advanced Learning Analytics". investiga características de estudiantes desertores de diferentes universidades de Pakistán, aplicando los algoritmos de minería de datos máquina de vectores de soporte (SVM), árbol de clasificación y regresión (CART), red Bayes (BN) en variables de gasto familiar , ingresos familiares, información personal del estudiante (Género, Estado civil, tipo de institución anterior) y activos familiares, los resultados muestran que el gasto familiar y las características de información personal, tienen un impacto significativo en el desempeño del estudiante.

Viloria y Parody (2016), en el artículo "Methodology for obtaining a predictive model academic performance of students from first partial note and percentage of absence", busca evaluar el impacto de la calificación del primer parcial y el porcentaje de no asistencia en los estudiantes de grado final usando un algoritmo ajustado en R y las variables nota parcial, porcentaje de ausencias, nota final de los estudiantes, determina que la primera calificación parcial y el porcentaje de ausencias, son variables directamente asociadas con el rendimiento académico final del estudiante, las cuales permiten la predicción acertada.

La minería de datos, por sus niveles de predicción, se propone como una herramienta para medir el desempeño académico, Han y Watts (2017), en el artículo "Predicting the Academic Performance of International Students on an Ongoing Basis", estudia la manera de predecir el desempeño académico de estudiantes internacionales, para su estudio emplea algoritmos como REPTree, J48 Tree y Modelo de Árbol Logístico(LMT), analiza información demográfica de los estudiantes internacionales, antecedentes académicos, promedio, registros de asistencia semanal, género, etnia, edad, asistencia media, número de ponencias, asistencia semanal, pasado éxito/fracaso, encontrando que de acuerdo con los modelos, un estudiante con un promedio más alto, radio de logros, y asistencia semanal tiene una mejor oportunidad de pasar el curso actual, por lo tanto estas dos variables estan directamente relacionadas.

Los aportes de Oyedotun, Tackie, Olaniyi, y Khashman (2015), en el artículo "Data Mining of Students' Performance: Turkish Students as a Case Study", quienes crean un modelo de predicción mediante redes neuronales, el cual, busca predecir la deserción, para ello analiza variables relacionadas con atributos académicos, de estudiantes de la universidad de Ankara en Turquía, concluyen que las veces que un estudiante haya repetido, son factores cruciales de deserción y que su implementación en la creación del modelo, produjeron mayor exactitud en la predicción.

Con el fin de evitar la deserción y predecir el número de graduados de una promoción, Ismail Abdulla (2015), en el artículo "Design and implementation of an intelligent system to predict the student graduation AGPA", diseña un modelo de minería de datos, para predecir el número de graduados de una promoción, en su estudio, analiza variables de tipo académicas (calificaciones del estudiante y número de alumnos por clase), el modelo se creó mediante la metodología para la minería de datos (CRISP-DM) sus resultados mostraron una lista de variables potenciales de deserción, lo que permite agrupar y crear estrategias que permitan alcanzar la graduacióny evitar abandono escolar.

Así mismo, Thakar, Metha, y N, (2015), en el artículo "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", estudia los factores de deserción estudiantil mediante uso de árboles de decisión, redes bayesianas, técnica de clustering, algoritmo de análisis de asociación, prueba de Chi-cuadrado usando

variables académicas y personales como inteligencia emocional, experiencia académica, autogestión encontrando que, factores como estar satisfecho con un curso en particular, el tipo de evaluaciones, comprensión de la clase, planificación de la clase e inteligencia emocional, son los que influyen más significativamente en la permanencia o deserción.

Las investigaciones de Altujjar, Altamimi, Al-Turaiki, y Al-Razgan (2016), en el artículo "Predicting Critical Courses Affecting Students Performance: A Case Study", estudian la forma de predecir a los estudiantes que están por desertar, para ello analiza variables de datos personales y calificaciones, mediante arboles de decisión, encontrando que las calificaciones son altamente importantes, al momento de predecir, así mismo, propone utilizar el 75% de datos para crear el modelo, mientras que el 25% se usa para la verificación de funcionamiento del modelo.

En la ciudad de Ambato se ha aplicado técnicas de minería de datos con K-means para el análisis de datos en el sector educativo en la evaluación docente. En el artículo "Minería de datos aplicada a la evaluación docente. Caso práctico Uniandes.", se encuentra que, K-means, logra altas tasas de predicción de resultados, mediante el análisis del historial de evaluación docente (Martinez, 2018).

Regresión logística.

En las investigaciones de Carvajal y González (2018), en el artículo "Variables Sociodemográficas y Académicas Explicativas de la Deserción de Estudiantes.", se aplica regresión logística, que permite analizar variables dicotómicas (que pueden tomar dos valores), para predecir la sucesión de un evento; su estudio busca conocer, la influencia de variables sociodemográficas y académicas en la deserción de estudiantes universitarios, para ello analizan los registros de 169 estudiantes, mediante técnicas estadísticas de tipo descriptivas, encuentran que las variables más importantes de predicción, son el número de niveles aprobados, el estado civil y profesión de los padres, la regresión al permitir analizar variables dicotómicas, es una estrategia de minería eficiente en variables relacionados con la deserción.

Henríquez y Escobar (2016), en el artículo "Construcción de un modelo de alerta temprana para la detección de estudiantes en riesgo de deserción de la universidad metropolitana de ciencias de la educación", en su estudio sobre alerta temprana para la detección de estudiantes en riesgo de deserción, utilizan el modelado de regresión logística, se analiza los factores que influyen en la deserción, estudiando variables pedagógicas y personales de una población de 1006 estudiantes, encuentran que las variables relacionadas con calificaciones, son determinantes importantes en la predicción, el modelo presento altos índices de predicción mediante la evaluación con la curva de característica de funcionamiento del receptor (ROC).

Logit

Es un algoritmo matemático basado en el logaritmo natural, que utiliza R para la regresión logística.

1.2.Objetivos

General

- Diseñar un modelo predictivo de deserción estudiantil universitaria basado en regresión.

Específicos

- Seleccionar la base de datos que servirá como base para la generación del modelo predictivo.

  Se realiza un diagnóstico previo con el número de estudiantes que posee las Carreras de la Facultad de Ciencias Humanas y de la Educación en el periodo septiembre 2019-enero 2020, en función de los datos mostrados en la página web de la Universidad Técnica de Ambato: www.uta.edu.ec

  Se selecciona la Carrera de Turismo por poseer mayor cantidad de datos, además porque se cuenta con el visto bueno de la Coordinación de la Carrera.

- Aplicar la metodología (KDD) para el descubrimiento de conocimiento en bases de datos.

  Por ser una metodología enfocada específicamente a la minería de datos y por contar con una estructura eficiente en el desarrollo de esta técnica, así mismo

por ser la más completa en comparación con otras metodologías de minería, se adopta la metodología KDD.

Se descarta metodologías como CRISP-DM (Proceso estándar de la industria cruzada para minería de datos), por estar enfocada exclusivamente a los negocios.

Además, no se toma en cuenta a la metodología SEMMA (Muestreo, exploración, modificación, modelado y evaluación), debido a que la metodología no contempla el análisis y comprensión del problema.

- Diseñar el modelo predictivo de deserción estudiantil basado en regresión logística en la carrera de turismo.

  Se aplica el algoritmo de minería de datos denominado regresión logística con el software R para la generación del modelo predictivo de deserción estudiantil con los datos históricos de los estudiantes de la Carrera de Turismo. Se utiliza R por ser un software es de libre distribución y estudios previos dan soporte a su funcionalidad.

  Al tratarse de un método predictivo de Minería de Datos, lo que se intenta es en función de los datos históricos, detectar a posibles futuros desertores, de manera que con esta información las autoridades, y estudiantes puedan prevenir y evitar la deserción.

CAPÍTULO II
METODOLOGÍA.

2.1 Materiales

Recursos humanos:

- Victor Xavier Vicente Guerrero
- Tutor: Ing. Mg. Blanca Cuji

Recursos materiales:

Hardware

- Computador
- Impresora

Software

- Base de datos
- Programa R
- Paquete de Office

Recursos Institucionales:

- UTA (laboratorios)

2.2. Costos

Se muestra la tabla de costos (Ver Tabla 1), en la que se detalla cada uno de los recursos utilizados.

Tabla 1-Analisis de costos

N°	Descripción	Costo
1	Planificación	$30
2	Análisis	$60
3	Aplicación	$160
4	Análisis de resultados	$50
5	Reporte	$20
Costo Total		$320

2.3. Métodos

2.3.1. Enfoque de la investigación

El estudio está dirigido por el enfoque mixto, al momento de implementar el análisis de los resultados de las encuestas, cuantitativo que permite mostrar datos numéricos, para presentar un punto de vista en el trabajo elaborado, cualitativo que permita describir la información.

2.3.2. Modalidad de la investigación

La modalidad es de campo ya que se va a analizar la información de la población seleccionada y los datos tomados desde las fuentes en las que se obtienen.
Bibliográfica, se basará en contenidos abordados en publicaciones de libros existentes en diferentes repositorios tanto de la universidad como repositorios de la ciudad.
Documental que permitirá conocer los estudios que dan soporte a la investigación, estudios, revistas, publicaciones en línea que refieren al tema que abordará la investigación.

2.3.3. Tipo de investigación

Descriptivo, ya que los resultados presentados serán analizados y descritos dando un punto de vista. Mediante la descripción escrita, lo cual logra que los lectores conozcan lo que se ha encontrado claramente.

2.3.4. Población y Muestra

Población

85 estudiantes de la Universidad Técnica de Ambato, Facultad de Ciencias Humanas, Carrera de Turismo y Hotelería.

2.3.5. Metodología

Cuji, Gavilanes, y Sanchez (2017), en el artículo "Modelo predictivo de deserción estudiantil basado en arboles de decisión", menciona que para la creación del modelo, se hace uso de la metodología de descubrimiento de conocimiento en Bases de datos (KDD), la cual ha sido probada en estudios anteriores, esta permite analizar datos para descubrir patrones en base a variables, los cuales muestran los resultados de la variable tomada como predictora en el análisis, su estructura está integrada por 5 fases:

selección, procesamiento, transformación, minería de datos y evaluación, (Ver Figura 1).

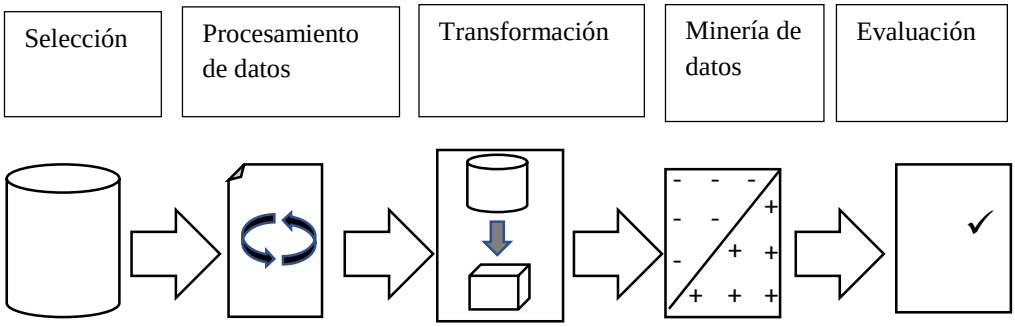


Figura 1-Modelo KDD

a. Selección

Permite comprender el problema, en esta fase se realiza los procesos de recolección e integración de datos

Recolección de datos

Los datos seleccionados para el estudio pertenecen a 509 estudiantes de la carrera de Turismo y Hotelería de la Universidad Técnica de Ambato(UTA), los registros proporcionados por la institución están entre los años 2013 a 2016, se ha utilizado un identificador numérico del 1…n reemplazando los campos de identificación con el fin de mantener la confidencialidad de los participantes del estudio, según se establece en la ley del sistema nacional de registros públicos, Nacional Pleno (n.d.). Además, campos de género y datos académicos identificados con el curso matriculado y notas del primer parcial(nota1) y segundo parcial(nota2)(Ver Figura 2).

Identificador	Genero	nota1p	nota2p	nota1s	nota2s	nota1t	nota2t	nota1c	nota2c
1	1	8.09	7.97	7.72	7.69	7.88	7.92	7.83	8.12
2	1	8.09	7.97	7.72	7.69	7.88	7.92	7.83	8.12
3	2	8.09	7.97	7.72	7.69	7.88	7.92	5.50	8.50
4	2	8.09	7.97	7.72	7.69	7.88	7.92	6.20	7.30
5	2	8.09	7.97	7.72	7.69	7.88	7.92	7.71	8.80
6	1	8.09	7.97	7.72	7.69	7.88	7.92	8.19	9.27

Figura 2-Matriz: datos generales-académicos

Integración de datos

Los registros obtenidos para el análisis, fueron 13 768, los mismos que fueron integrados en una hoja única de cálculo, para poder ser utilizados en la minería de datos, a su vez se agregaron atributos de género y desertor, se asignaron las etiquetas de nota 1, nota 2 para identificar el primero y segundo parcial y los correspondientes sufijos de identificación de nivel(p=primer, s=segundo, t=tercero, c=cuarto, q= quinto, se=sexto, sp=séptimo, o=octavo y n=noveno)(Ver Figura 3)

periodo al q	carrera	nivel	genero	nota1p	nota2p	nota1s	nota2s
marzo/13-A	turismo y hoteleria	cuarto	femenino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	femenino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	masculino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	masculino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	masculino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	femenino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	femenino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	femenino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	masculino	8.09	7.97	7.72	7.69
marzo/13-A	turismo y hoteleria	cuarto	masculino	8.09	7.97	7.72	7.69

Figura 3-Matriz integral con atributos como género, nota 1, nota 2, desertor.

Dada la matriz inicial, se dedujo otros atributos en la matriz, de la siguiente manera:

Género: tomando en cuenta el nombre asignado al estudiante

Desertor: en base al curso matriculado, las calificaciones obtenidas y el lapso transcurrido comparado con el ultimo nivel alcanzado.

nota1p, nota2p, nota1s, nota2s…n: se obtuvieron, en base al promedio de calificaciones de todos los módulos de un mismo nivel. Repitiendo el proceso por cada nivel que el estudiante cursó.

b. Procesamiento de datos

Para poder ser utilizados en la minería de datos, los registros se someten a dos procesos que son la limpieza y el almacenamiento de datos.

Limpieza de los datos

Al momento de etiquetado de columnas identificadoras de los registros, se cuenta con un total de 536 datos, se detectó datos atípicos (datos extremadamente grandes o pequeños, Daniel y Cesar (2007)) por lo que se procedió a eliminar dichos registros,

con el fin de evitar resultados erróneos, posterior a la limpieza e integración, se obtiene 509 datos.

c. Transformación de datos

Al trabajar en minería de datos, los registros deben ser cuantitativos, por lo que algunos debieron ser modificados, es el caso de variables como género y desertor, para ello se creó las ponderaciones necesarias en cada caso (Ver Tabla 2)

Tabla 2-Ponderaciones de variables

Nombre de variable	Ponderación
Género	1= Femenino
	2=Masculino
Desertor	0=no
	1=si

Se hizo necesario crear ponderaciones, para dos variables del tipo cualitativas, para transformarlas a un formato adecuado, de tipo cuantitativo. Para el género de los participantes, se utiliza 1 para "femenino" y 2 para "masculino", en el caso de la variable desertor, se utiliza 1 para estudiantes que han desertado y 0 para quienes no. (Ver Tabla 2).

d. Minería de datos

Determinar la tarea de minería.

Se usa regresión, en la que se supervisa cada uno de los pasos de aprendizaje del modelo, Analizadas las características del conjunto de datos a los que se aplica la minería, considerando de que la regresión permite a partir de un conjunto de variables predecir una nueva variable, la variable predictora.

Selección del algoritmo.

Analizadas las características del problema a solucionar, los algoritmos disponibles y recursos, se elige Regresión logística. Logit Silva Ayçaguer (2000).

Generación del modelo

Los datos disponibles, se dividen en proporciones, la una utilizada para la generación y entrenamiento del modelo, otro porcentaje se utiliza para realizar las pruebas de funcionamiento (Altujjar, Altamimi, Al-Turaiki, & Al-Razgan, 2016), los valores "TRUE" representan el 75% de los datos para entrenar el modelo y los valores "FALSE" representan el 25% que se usan para las pruebas (Ver Figura 4).

```
> split<-sample.split(estudiantes, SplitRatio = 0.75)
> split
 [1]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
[18]  TRUE
>
```

Figura 4-División de datos

Con la ayuda del software R, se procede a generar el modelo de deserción estudiantil, para ello se usa regresión logística cuya función *Logit*, genera las interacciones necesarias hasta determinar cómo se ve influenciada la predicción por cada una de las variables seleccionadas.

- Se cargan todas las variables disponibles y se determina las menos influyentes, basándose en el incremento estimado por unidad de aumento en la variable (Estimate) (Ver Figura 5), otro factor determinante es ver las medidas de dispersión de la variable en torno a la media (Pr(>|z|), mientras más cercano se encuentre a cero, más significante este será, es importante considerar el error en cada caso. Basándose en los parámetros establecidos, se observa que variables como género, nota2sp, nota1o y nota2o, son poco significantes en la predicción del modelo (Ver Figura 5).

```
Deviance Residuals:
      Min          1Q      Median          3Q         Max
-4.465e-05  -2.100e-08  -2.100e-08  -2.100e-08   3.025e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.166e+01  3.939e+05   0.000    1.000
genero      -4.167e+00  2.452e+04   0.000    1.000
notalp       1.706e+00  6.805e+04   0.000    1.000
nota2p      -1.045e+00  8.274e+04   0.000    1.000
notals      -9.471e-01  2.625e+04   0.000    1.000
nota2s       7.490e-01  2.166e+04   0.000    1.000
notalt      -3.775e-01  5.364e+04   0.000    1.000
nota2t       3.254e-01  5.240e+04   0.000    1.000
notalc      -5.754e-01  1.862e+04   0.000    1.000
nota2c       2.231e-01  1.225e+04   0.000    1.000
notalq       8.221e-01  3.593e+04   0.000    1.000
nota2q      -3.584e-01  3.409e+04   0.000    1.000
notalse      3.363e+00  2.309e+04   0.000    1.000
nota2se     -1.869e+00  1.689e+04   0.000    1.000
notalsp     -3.885e+00  2.334e+04   0.000    1.000
nota2sp      1.002e+00  1.671e+04   0.000    1.000
notalo      -1.072e+01  7.266e+03  -0.001    0.999
nota2o       1.444e+00  4.594e+03   0.000    1.000
```

Figura 5-Determinar variables menos influyentes

- Se propone un nuevo modelo, excluyendo las variables menos significativas, lo cual resulta en un nuevo modelo (modelo1) el cual arroja los siguientes resultados (Ver Figura 6):

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7142  -0.1038  -0.0746  -0.0280   3.4557

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 15.43841   14.22425   1.085  0.27776
nota1p       0.16748    1.30562   0.128  0.89793
nota2p       0.48264    1.28201   0.376  0.70656
nota1s      -2.56511    1.63098  -1.573  0.11578
nota2s       1.94316    1.50581   1.290  0.19690
nota1t      -1.13740    1.00832  -1.128  0.25931
nota2t      -0.04619    1.15650  -0.040  0.96814
nota1c      -0.16134    0.32854  -0.491  0.62336
nota2c       0.68198    0.75519   0.903  0.36650
nota1q       0.19601    0.77519   0.253  0.80038
nota2q      -0.75392    0.69773  -1.081  0.27991
nota1se     -0.87882    0.80876  -1.087  0.27720
nota2se      0.34764    0.75022   0.463  0.64309
nota1sp     -0.97152    0.37655  -2.580  0.00988 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 6-Modelo resultante

- Se vuelve a verificar en la gráfica la significancia de las variables, para esta interacción se observa que tanto nota1p, nota2p, nota2t y nota2se son menos significativas, por lo que también se suprimen del modelo, los resultados obtenidos en esta interacción tienen un mayor grado de significación (Ver Figura 7)

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7704  -0.1083  -0.0809  -0.0332   3.4075

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  19.0521    11.9495   1.594  0.11085
nota1s       -2.2358     1.3433  -1.664  0.09604 .
nota2s        1.5565     1.2378   1.257  0.20858
nota1t       -0.9929     0.8506  -1.167  0.24306
nota1c       -0.1529     0.3252  -0.470  0.63816
nota2c        0.6095     0.7450   0.818  0.41327
nota1q        0.1266     0.7365   0.172  0.86346
nota2q       -0.5322     0.5063  -1.051  0.29325
nota1se      -0.5686     0.5618  -1.012  0.31153
nota1sp      -0.9385     0.3049  -3.078  0.00208 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 7-Variables para el modelo

En esta interacción las medidas de dispersión de la variable en torno a la media (Pr(>|z|) están aproximadamente entre 0,5 y 0 por lo que se plantea como modelo de predicción. Las variables que no han sido excluidas son las que servirán como predictoras en el modelo creado.

e. Evaluación

- El modelo generado, se analiza en el 25% de datos que se han separado para este fin, para ello se accede a la variable que guarda la matriz y se le aplica el modelo, se muestra la tabla de las predicciones logradas por el modelo. (Ver Figura 8). Se observa 109 participantes no desertores, correctamente clasificados por el modelo, un estudiante clasificado como no desertor fue mal clasificado, mientras que los participantes desertores, 31 fueron correctamente clasificados mientras que uno estuvo mal clasificado. Por lo tanto, el modelo logra una eficiencia del 98,59% con lo que el modelo se considera válido.

```
> confmatrix<-table(actual_value=test$desertor, predicted_value=res >0.5)
> confmatrix
            predicted_value
actual_value FALSE TRUE
           0   109    1
           1     1   31
> (confmatrix[[1,1]]+confmatrix[[2,2]])/sum(confmatrix)
[1] 0.9859155
>
```

Figura 8-Validación del modelo

2.4.Hipótesis

La minería de datos permite la predicción de las variables que influyen en la deserción estudiantil en la carrera Turismo y Hotelería de la Universidad Técnica de Ambato.

CAPÍTULO III
RESULTADOS Y DISCUSIÓN

3.1. Análisis y discusión de los resultados.

3.1.1. Análisis

A partir de la aplicación de encuestas a 85 de los estudiantes de primero, tercero y cuarto semestre de la carrera de Turismo y Hotelería de la Universidad Técnica de Ambato, se muestra un análisis completo de los resultados encontrados en relación con la percepción de la deserción estudiantil y los factores que la provocan.

A continuación, se visualiza e interpreta los resultados obtenidos en cada una de las preguntas:

Pregunta 1.- ¿Cree usted que la regresión logística (explicada previo a la aplicación del cuestionario), es un buen mecanismo de predicción de deserción estudiantil?

Tabla 3-Regresión logística, como mecanismo de predicción

Ítem	Frecuencia	Porcentaje
Siempre	26	31
A veces	54	63
No	4	5
Otros	1	1
Total	**85**	**100**

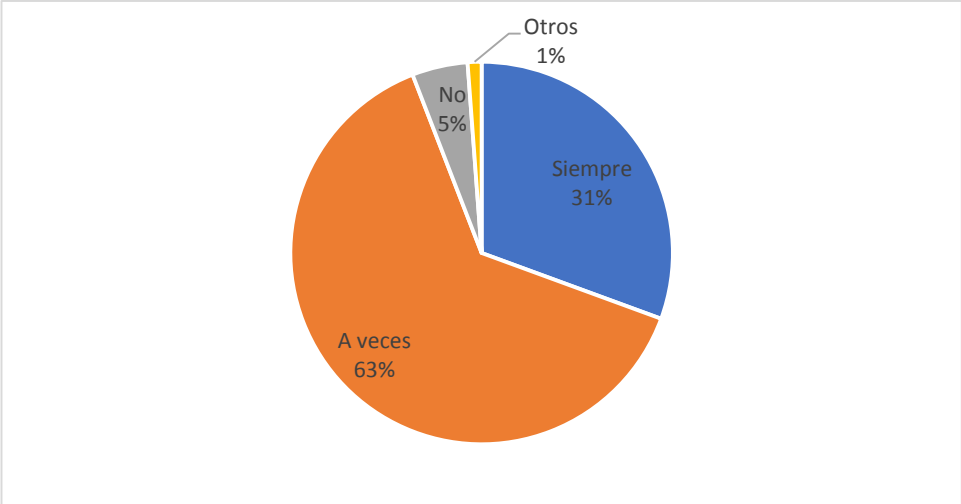


Figura 9-Regresión logística, como mecanismo de predicción

Análisis

En esta pregunta, el 63% de participantes cree que la regresión logística, a veces es un buen mecanismo de predicción de deserción estudiantil, el 31% de los estudiantes encuestados consideran que la regresión logística, es un buen mecanismo de predicción de deserción estudiantil y un 4% opinan que la regresión logística, no es un buen mecanismo de predicción de deserción estudiantil, el 1% de estudiantes no responde ni afirmativa ni negativamente a la pregunta (Ver Tabla 3).

Interpretación

Al aplicar la encuesta, se observa que la mayoría de los estudiantes, consideran la regresión logística un buen mecanismo de predicción de deserción, sin embargo, existe un pequeño número de estudiantes encuestados que no consideran la regresión un mecanismo adecuado de predicción (Ver Figura 9).

Pregunta 2 ¿Considera usted que es importante la investigación para detectar a posibles desertores de la carrera?

Tabla 4-Investigación sobre deserción estudiantil

Ítem	Frecuencia	Porcentaje
Siempre	64	75
A veces	18	21
No	3	4
Otros	0	0
Total	**85**	**100**

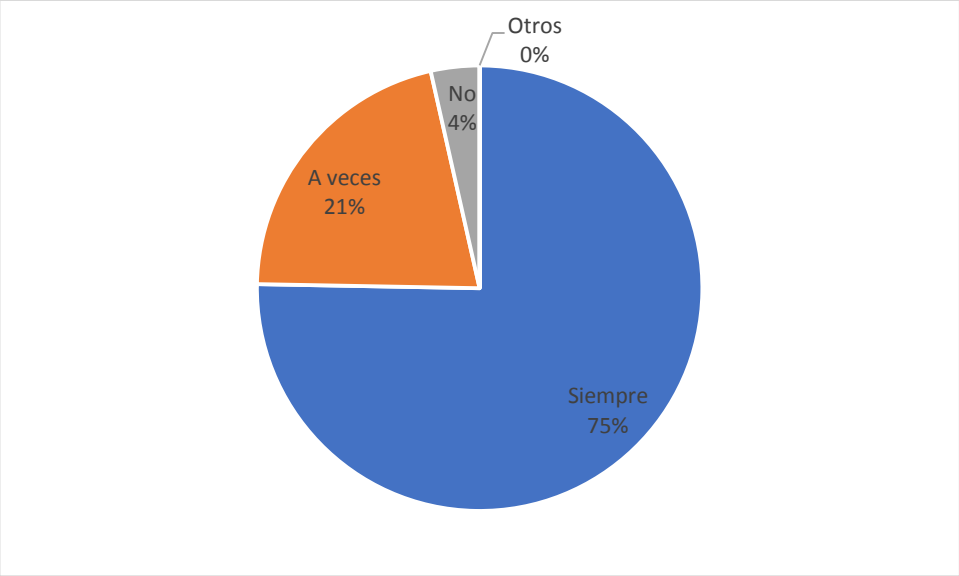


Figura 10-Investigación sobre deserción estudiantil

Análisis

En esta pregunta el 75% de los estudiantes encuestados indican que es importante la investigación para detectar a posibles desertores de la carrera, sin embargo, el 21% de las personas consideran que únicamente a veces es necesario aplicar estos mecanismos, el 4% de los encuestados opinan que no es importante la investigación con este fin. (Ver Tabla 4).

Interpretación

La información que se obtiene por medio de la encuesta es que es importante para la mayoría de los estudiantes las investigaciones con el fin de detectar a posibles desertores. Sin embargo, también se observó que existe una cantidad menos significativa de estudiantes, quienes no consideran importante las investigaciones con el fin de predecir la deserción en la carrera (Ver Figura 10).

Pregunta 3 ¿Conocer a tiempo las causas que provocan la deserción ayuda a las autoridades y estudiantes a prevenir el abandono estudiantil?

Tabla 5-Causas de la deserción

Ítem	Frecuencia	Porcentaje
Siempre	54	64
A veces	29	34
No	2	2
Otros	0	0
Total	**85**	**100**

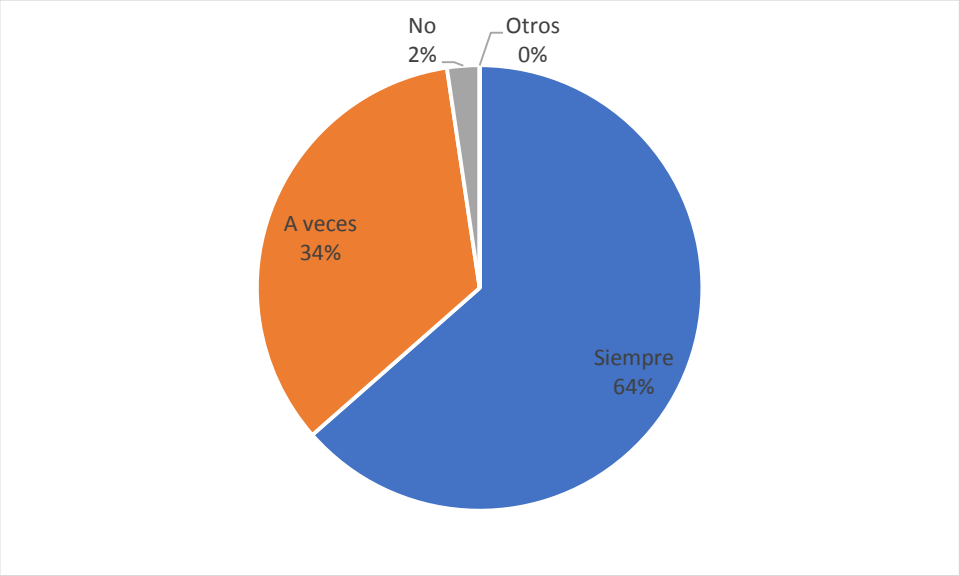


Figura 11-Causas de la deserción

Análisis

En esta pregunta el 64% de los estudiantes encuestados opinan que conocer a tiempo las causas que provocan la deserción ayuda a las autoridades y estudiantes a prevenir el abandono estudiantil, el 34% de las personas consideran que a veces, mientras que el 2% de los encuestados opinan que conocer a tiempo las causas que provocan la deserción no ayuda a las autoridades y estudiantes a prevenir el abandono estudiantil (Ver Tabla 5).

Interpretación

A partir de esta pregunta, se obtiene que un gran porcentaje de estudiantes considera importante conocer las causas que provocan la deserción, lo cual ayuda a las

autoridades y estudiantes a prevenir el abandono estudiantil, así mismo, una cantidad significativa de encuestados considera que este mecanismo únicamente es importante a veces y un pequeño número de ellos respondieron que no ayuda a prevenir la deserción el conocer las causas que la provocan (Ver Figura 11).

Pregunta 4: ¿Considera usted que las calificaciones son uno de los factores que mayor deserción estudiantil provocan?

Tabla 6-Las calificaciones como factores de deserción

Ítem	Frecuencia	Porcentaje
Siempre	28	33
A veces	53	62
No	4	5
Otros	0	0
Total	**85**	**100**

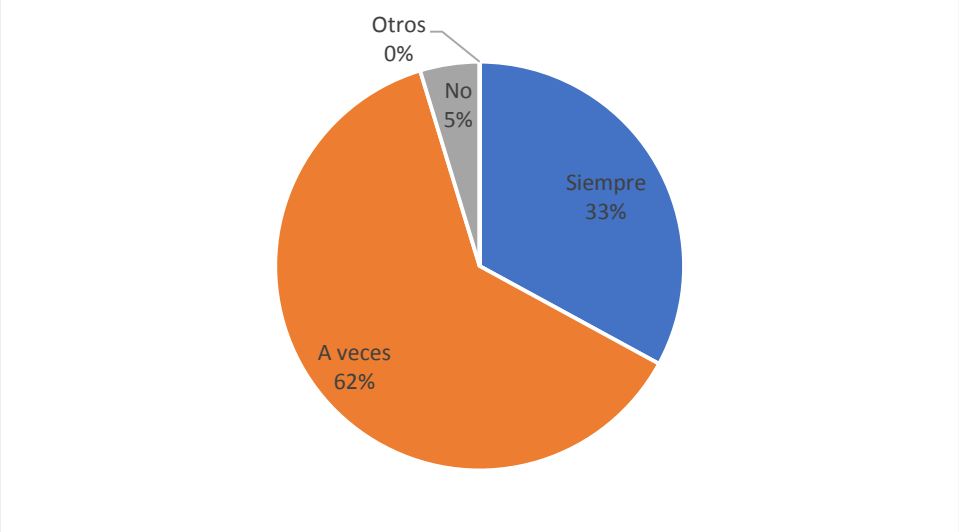


Figura 12-Las calificaciones como factores de deserción

Análisis

En esta pregunta el 62% de las personas consideran que a veces las calificaciones son uno de los factores que mayor deserción estudiantil provocan, mientras que el 33% indican que siempre las calificaciones son factores que provocan deserción, una cantidad menos significativa del 5% de los encuestados opinan que las calificaciones no son factores relacionados con la deserción (Ver Tabla 6).

Interpretación

La pregunta muestra que las calificaciones, son uno de los elementos clave para la deserción en un gran número de estudiantes, ya que la mayoría respondieron que siempre o a veces si influye las calificaciones en la decisión de abandonar los estudios, solamente un número reducido de estudiantes no considera a las calificaciones como un factor determinante para el abandono escolar (Ver Figura 12).

Pregunta 5 ¿Cree que las calificaciones de los primeros semestres inciden significativamente en la deserción de abandono de estudios universitarios?

Tabla 7-Calificaciones de los primeros semestres y la deserción

Ítem	Frecuencia	Porcentaje
Siempre	23	27
A veces	40	47
No	22	26
Otros	0	0
Total	**85**	**100**

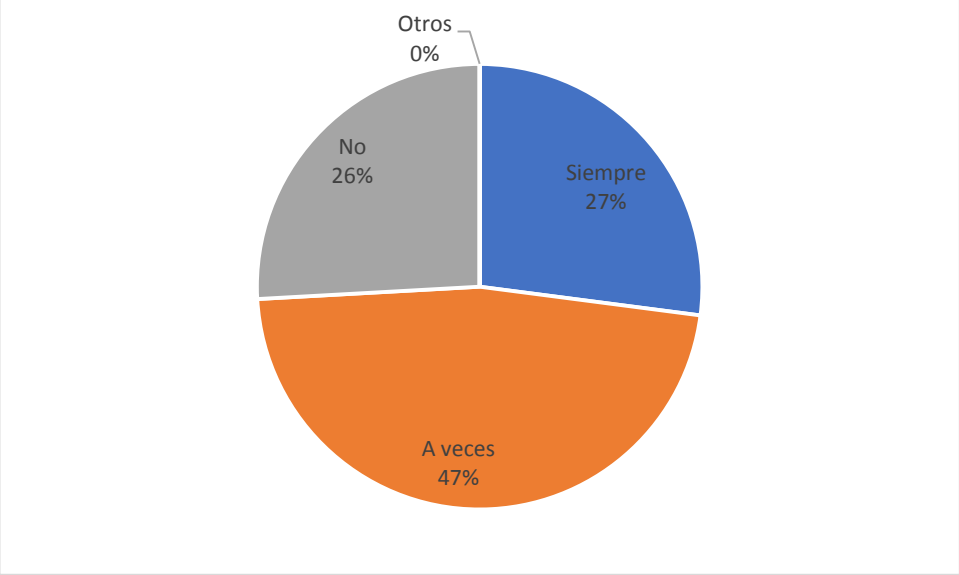


Figura 13-Calificaciones de los primeros semestres y la deserción

Análisis

En esta pregunta el 47% de encuestados, respondieron que a veces las calificaciones de los primeros semestres inciden en la decisión de abandonar los estudios, así mismo el 27% de los estudiantes encuestados indican que las calificaciones de los primeros

semestres inciden significativamente en la deserción de abandono de estudios universitarios, un 26% dice que no considera influyentes dichas calificaciones (Ver Tabla 7).

Interpretación

Los resultados de esta pregunta muestran que las calificaciones de los primeros semestres influyen únicamente algunas veces en la decisión de abandonar los estudios, mientras que, otros encuestados dijeron que siempre se relacionan las calificaciones de primeros semestres con el abandono, hubo una cantidad significativa que considera que esta variable no esta relacionada con el abandono (Ver Figura 13).

Pregunta 6 ¿Considera el estado civil del estudiante como un factor que influye en la deserción?

Tabla 8-Estado civil del estudiante como un factor de la deserción

Ítem	Frecuencia	Porcentaje
Siempre	19	23
A veces	53	62
No	13	15
Otros	0	0
Total	**85**	**100**

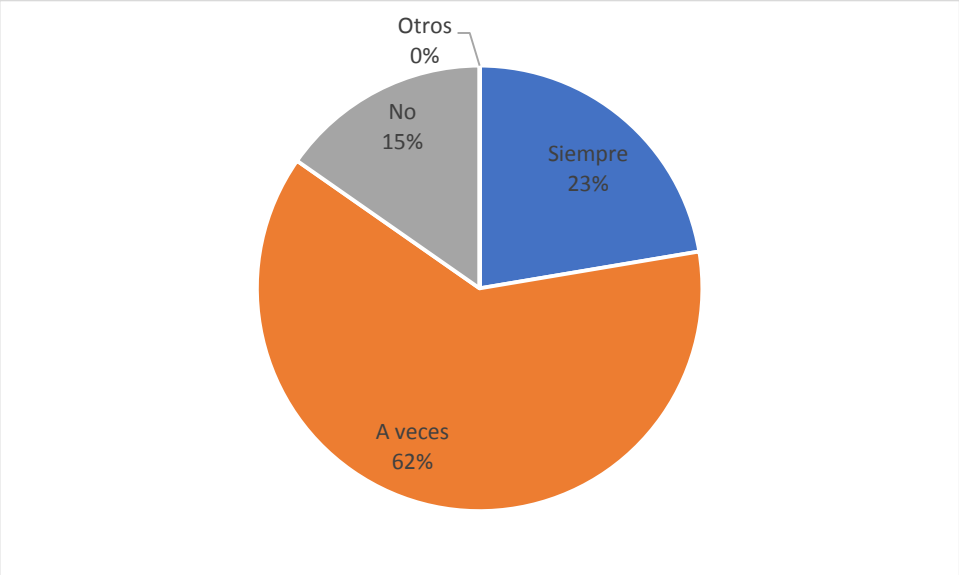


Figura 14-Estado civil del estudiante como un factor de la deserción

Análisis

En esta pregunta el 62% de las personas dice que a veces el estado civil influye en el abandono estudiantil, el 23% de los estudiantes encuestados indican que el estado civil del estudiante es en definitiva un factor que influye en la deserción, sin embargo, el 15% de estudiantes no relaciona estas dos variables (Ver Tabla 8).

Interpretación

La pregunta muestra que en la mayoría de los casos, únicamente a veces el estado civil es determinante para el abandono estudiantil mientras que un numero significativo de estudiantes si lo consideran como un factor de abandono al estado civil, algunos de ellos creen que no es importante (Ver Figura 14).

Pregunta 7 ¿Considera que los hombres abandonan los estudios más que las mujeres?

Tabla 9-Los hombres abandonan los estudios más que las mujeres

Ítem	Frecuencia	Porcentaje
Siempre	14	16
A veces	39	46
No	32	38
Otros	0	0
Total	**85**	**100**

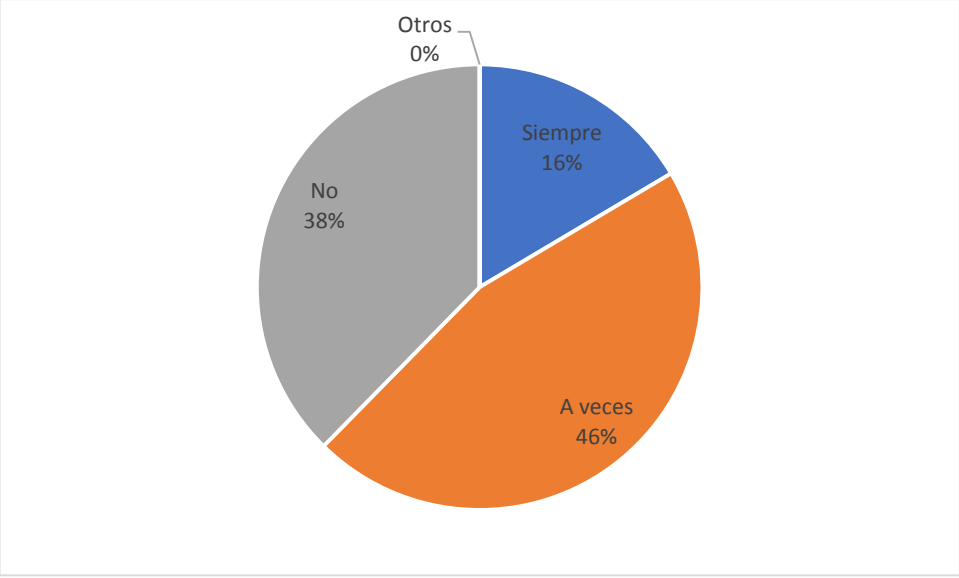


Figura 15-Los hombres abandonan los estudios más que las mujeres

Análisis

En esta pregunta el 46% de los estudiantes encuestados indican que los hombres abandonan los estudios más que las mujeres únicamente a veces mientras que, el 38% indican que no es determinante esta variable en el abandono y solamente el 16% de los encuestados dijo que si considera a los hombres como la población que abandona más los estudios (Ver Tabla 9).

Interpretación

A partir de los resultados obtenidos en la pregunta, se puede concluir que los estudiantes únicamente a veces consideran que el género masculino es más propenso a desertar, mientras que una proporción menos significativa dice que si mientras que los demás encuestados, dijeron que el ser hombre no incide en el abandono estudiantil (Ver Figura 15).

Pregunta 8 ¿Considera que las mujeres abandonan los estudios más que los hombres?

Tabla 10-Las mujeres abandonan los estudios más que los hombres

Ítem	Frecuencia	Porcentaje
Siempre	10	12
A veces	48	56
No	27	32
Otros	0	0
Total	**85**	**100**

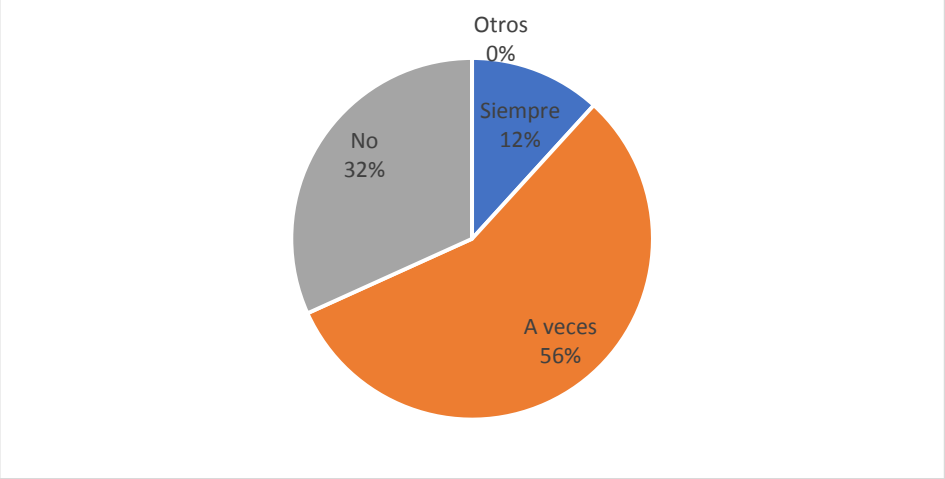


Figura 16-Las mujeres abandonan los estudios más que los hombres

Análisis

En esta pregunta el 56% de los estudiantes encuestados indican que únicamente a veces las mujeres abandonan los estudios más que los hombres mientras que el 32% creen que no es determinante esta variable para el abandono, el 12% de estudiantes, si consideran que el abandono estudiantil está relacionado con el hecho de ser mujer (Ver Tabla 10).

Interpretación

Los resultados de la pregunta muestran que el género femenino, solamente a veces es una variable que incide en el abandono estudiantil, mientras que un gran número de encuestados no considera esta variable como causante del abandono estudiantil, únicamente para una pequeña población encuestada, si depende de esta variable la deserción estudiantil (Ver Figura 16).

3.1.2. Discusión

Como lo muestran los estudios de Cuji, Gavilanes, y Sánchez (2017), en el artículo "Modelo predictivo de deserción estudiantil basado en arboles de decisión", KDD muestra eficiencia en la construcción del modelo de predicción, sus lineamientos permiten la construcción del modelo más adecuado, acorde con los requerimientos, los resultados que permite encontrar son los siguientes:

En el estudio de Bonaldo y Pereira (2016) titulado "Dropout: Demographic Profile of Brazilian University Students", los investigadores encontraron que el género, junto con las calificaciones y otras variables, son las más influyentes del modelo, mientras que, los resultados de la aplicación de KDD y las encuestas en la carrera de Turismo de la Universidad Técnica de Ambato, muestran que el género de los participantes no influye significativamente en si aprueba o no el curso; por otro lado, las calificaciones si son influyentes.

Una variable significativa en la creación del modelo son las calificaciones pertenecientes al segundo semestre (nota1s+nota2s), las cuales son en los dos parciales significativas para el modelo.

De igual manera se observa que las calificaciones de tercer semestre están divididas, las calificaciones del primer parcial (nota1t) son más importantes para el modelo mientras que la calificación del segundo parcial (nota1t) no influyen en gran medida.

Al igual que en las investigaciones de Viloria y Parody (2016), en el artículo "Methodology for obtaining a predictive model academic performance of students from first partial note and percentage of absence", en el estudio se encontró que las calificaciones del primer parcial de cada semestre son las más influyentes del modelo, es el caso de: cuarto (nota1c, nota2c), quinto (nota1q, nota2q) y primer parcial de sexto (nota1se) y séptimo (nota1sp).

Las calificaciones de los niveles segundo tercero y cuarto, se pueden considerar como variables importantes debido a que su crecimiento sigue un patrón constante ascendente (Ver Figura 17)

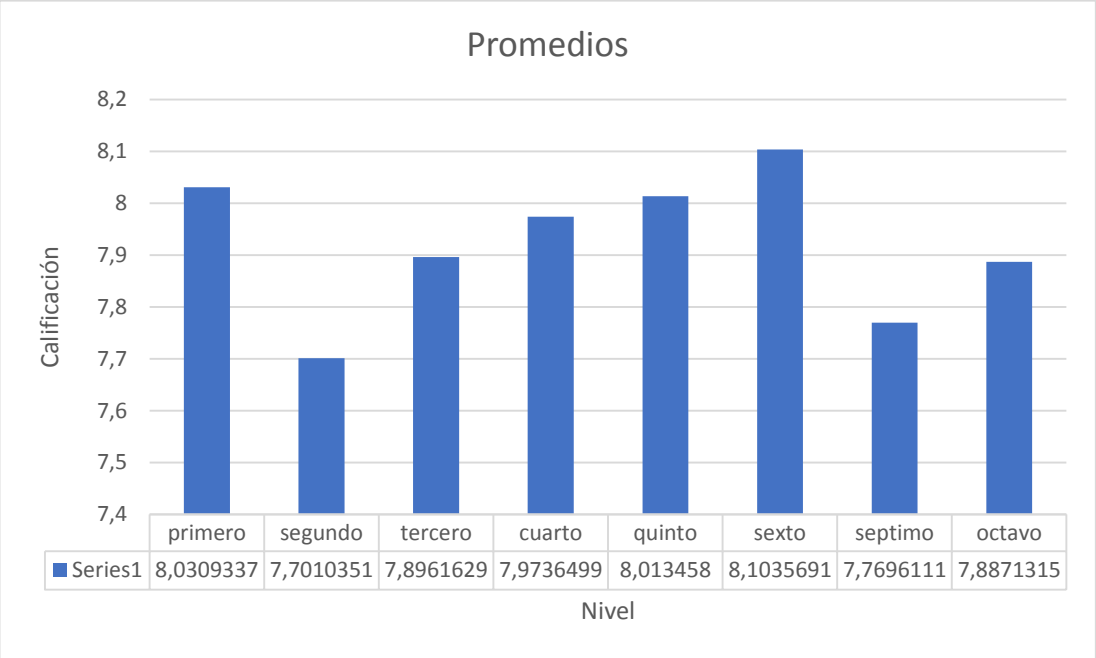


Figura 17-Tabla de promedios por nivel

Las calificaciones correspondientes al segundo parcial de tercer semestre (nota2t), segundo parcial de sexto semestre (nota2se), segundo parcial de séptimo semestre (nota2sp) y octavo semestre (nota1o y nota2o), no son representativas en la generación del modelo.

3.2. Verificación de hipótesis

3.2.1. Señalamiento de variables

Variable independiente: Minería de datos

Variable dependiente: Deserción estudiantil

Con el fin de verificar la hipótesis, se procede a comprobar la relación entre las variables, se ha procedido a aplicar la prueba del Chi-Cuadrado (X2), para ello se utiliza los datos de la encuesta aplicada a los estudiantes de la carrera de Turismo y Hotelería de la Universidad Técnica de Ambato.

a) Modelo lógico

Hipótesis Nula (H0) = La minería de datos NO permite la predicción de las variables que influyen en la deserción estudiantil en la carrera Turismo y Hotelería de la universidad Técnica de Ambato.

Hipótesis Alterna (H1) = La minería de datos SI permite la predicción de las variables que influyen en la deserción estudiantil en la carrera Turismo y Hotelería de la universidad Técnica de Ambato.

b) Modelo matemático.

Ho: O = E

Hi: O ≠ E

c) Modelo Estadístico

$$X^2 = \Sigma \left[\frac{(O - E)^2}{E} \right]$$

X^2 =Chi cuadrado

Σ=Sumatoria

O= Frecuencia observada

E= Frecuencia esperada

Chi-cuadrado prueba estadística que afirma o refuta la relación entre las variables seleccionadas.

Al analizar la relación entre las 85 respuestas a las variables del problema: regresión y deserción, se encuentra los resultados (Ver Figura 18)

regresion * desercion [recuento, fila %, columna %, total %].

	desercion			
regresion	Nunca	A veces	Siempre	Total
Nunca	1.00	1.00	3.00	5.00
	20.00%	20.00%	60.00%	100.00%
	33.33%	5.26%	4.76%	5.88%
	1.18%	1.18%	3.53%	5.88%
A veces	1.00	8.00	45.00	54.00
	1.85%	14.81%	83.33%	100.00%
	33.33%	42.11%	71.43%	63.53%
	1.18%	9.41%	52.94%	63.53%
Siempre	1.00	10.00	15.00	26.00
	3.85%	38.46%	57.69%	100.00%
	33.33%	52.63%	23.81%	30.59%
	1.18%	11.76%	17.65%	30.59%
Total	3.00	19.00	63.00	85.00
	3.53%	22.35%	74.12%	100.00%
	100.00%	100.00%	100.00%	100.00%
	3.53%	22.35%	74.12%	100.00%

Figura 18-Relación de variables

Tomando los valores de Chi cuadrado de Pearson (10,38), obtenidos en el programa PSPP, se observa que el valor es menor al valor Chi cuadrado de tablas (5,99) en un valor de p=0.05 (Ver Figura 19)

Pruebas Chi-cuadrado.

Estadístico	*Valor*	*df*	*Sig. Asint. (2-colas)*
Chi-cuadrado de Pearson	10.38	4	.034
Razón de Semejanza	8.17	4	.086
Asociación Lineal-by-Lineal	1.10	1	.295
N de casos válidos	85		

Figura 19-Prueba Chi Cuadrado

Por lo tanto, se rechaza H0 y se acepta H1, es decir: **La minería de datos SI permite la predicción de las variables que influyen en la deserción estudiantil en la carrera Turismo y Hotelería de la universidad Técnica de Ambato.**

CAPITULO IV

CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- La selección de los datos debe hacerse de manera exhaustiva, conociendo las limitaciones y alcances del estudio, debe tomarse en cuenta características como el espacio temporal en el que se sitúan los datos.
- Al momento de seleccionar la base de datos, esta debe pertenecer a una población que pueda beneficiarse de esta, por lo que se descartan bases de datos con carreras en proceso de ser cerradas, así mismo, esta debe contener un número significativo de registros, para evitar que los resultados obtenidos generen modelos inestables.
- La metodología (KDD) permite la creación de modelos predictivos con un alto grado de efectividad. Esta permitió la creación de un modelo con una efectividad del 98,59%.
- Tomando en cuenta que unas variables serán más importantes mientras que otras no aportarán de manera significativa en la predicción, se crea un modelo predictivo eficiente con las variables más influyentes, la matriz de confusión muestra la relación entre los valores predichos correctamente por el modelo y los valores originales de los datos de prueba (Ver Figura 10).

```
> confmatrix<-table(actual_value=test$desertor, predicted_value=res >0.5)
> confmatrix
            predicted_value
actual_value FALSE TRUE
           0   109    1
           1     1   31
> (confmatrix[[1,1]]+confmatrix[[2,2]])/sum(confmatrix)
[1] 0.9859155
>
```

Figura 20-Matriz de confusión

4.2 Recomendaciones

- Aplicar la técnica de minería de datos, para predecir la deserción estudiantil universitaria en la carrera de Turismo y Hotelería.

- Solicitar al estudiante al momento de la matricula información: cargas familiares, financiamiento de estudios entre otras, pues se observa estas variables en otras investigaciones.
- Crear un histórico, en el que se almacene los tipos de metodología empleados por los docentes para la enseñanza, con el fin de poder hacer uso de este recurso en futuras investigaciones sobre la deserción estudiantil.
- Emplear las siguientes metodologías: SEMMA y Catalyst (P3TQ) en futuras investigaciones, con el fin de observar los cambios que se producen en los resultados arrojados por estas y compararlos con los resultados obtenidos con KDD.

B. MATERIALES DE REFERENCIA

Referencias Bibliográficas

Abbas, A., Sarker, K. U., Mahmood, S., Hasan, R., & Palaniappan, S. (2015). The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. *International Journal of Business Information Systems*, *1*(1), 1. https://doi.org/10.1504/ijbis.2020.10020425

Altujjar, Y., Altamimi, W., Al-Turaiki, I., & Al-Razgan, M. (2016). Predicting Critical Courses Affecting Students Performance: A Case Study. *Procedia Computer Science*, *82*(March), 65–71. https://doi.org/10.1016/j.procs.2016.04.010

Bonaldo, L., & Pereira, L. N. (2016). Dropout: Demographic Profile of Brazilian University Students. *Procedia - Social and Behavioral Sciences*, *228*(June), 138–143. https://doi.org/10.1016/j.sbspro.2016.07.020

Carvajal, C. M., & González, J. A. (2018). *Variables Sociodemográficas y Académicas Explicativas de la Deserción de Estudiantes. 11*(2), 3–12.

Daniel, S., & Cesar, P. (2007). *Minería de datos. Técnicas y herramientas*. 13–14.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2018). *Predicting Student Performance using Advanced Learning Analytics*. (October), 415–421. https://doi.org/10.1145/3041021.3054164

Goswami, S., & Chakrabarti, A. (2015). Feature Selection: A Practitioner View. *International Journal of Information Technology and Computer Science*, *6*(11), 66–77. https://doi.org/10.5815/ijitcs.2014.11.10

Han, B., & Watts, M. J. (2017). *Predicting the Academic Performance of International Students on an Ongoing Basis*. (July 2016). Retrieved from http://unitec.researchbank.ac.nz/bitstream/handle/10652/3578/2016CITRENZ_1_Han_IntAcademicPerf_17-2.pdf?sequence=1

Henríquez, N., & Escobar, D. (2016). Construcción de un modelo de alerta temprana para la detección de estudiantes en riesgo de deserción de la universidad metropolitana de ciencias de la educación. *Revista Mexicana de Investigacion Educativa*, *21*(71), 1221–1248.

Iam-On, N., & Boongoen, T. (2017). Generating descriptive model for student

dropout: a review of clustering approach. *Human-Centric Computing and Information Sciences*, *7*(1), 1–24. https://doi.org/10.1186/s13673-016-0083-0

Ismail, S., & Abdulla, S. (2015). Design and implementation of an intelligent system to predict the student graduation AGPA. *Australian Educational Computing*, *30*(2).

Kerby, M. B. (2015). Toward a new predictive model of student retention in higher education: An application of classical sociological theory. *Journal of College Student Retention: Research, Theory and Practice*, *17*(2), 138–161. https://doi.org/10.1177/1521025115578229

Martelo, R. J., Herrera, K., & Villabona, N. (2017). *Estrategias para disminuir la deserción universitaria mediante series de tiempo y multipol*.

Martinez, C. (2018). Minería de datos aplicada a la evaluacion docente. Caso práctico Uniandes. *Órbita Pedagógica*, 61–76.

Nacional Pleno. (n.d.). *Ley del sistema nacional de registros públicos.* Retrieved from https://www.telecomunicaciones.gob.ec/wp-content/uploads/downloads/2012/11/LEY-DEL-SISTEMA-NACIONAL-DE-REGISTRO-DE-DATOS-PUBLICOS.pdf

Oyedotun, O. K., Tackie, S. N., Olaniyi, E. O., & Khashman, A. (2015). Data Mining of Students' Performance: Turkish Students as a Case Study. *International Journal of Intelligent Systems and Applications*, *7*(9), 20–27. https://doi.org/10.5815/ijisa.2015.09.03

Ricard, N. C., & Pelletier, L. G. (2016). Dropping out of high school: The role of parent and teacher self-determination support, reciprocal friendships and academic motivation. *Contemporary Educational Psychology*, *44–45*, 32–40. https://doi.org/10.1016/j.cedpsych.2015.12.003

Silva Ayçaguer, L. C. (2000). *Excursión a la regresión logística en ciencias de la salud.* Ediciones Díaz de Santos.

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, *9*(4), 1–5. https://doi.org/10.17485/ijst/2016/v9i4/87032

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *Proceedings of the 8th International Conference on Educational Data Mining*, (June), 392–395.

Thakar, P., Metha, A., & N, M. (2015). *Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue*. *110*(15), 60–68. Retrieved from http://arxiv.org/abs/1509.05176

Viloria, A., & Parody, A. (2016). Methodology for obtaining a predictive model academic performance of students from first partial note and percentage of absence. *Indian Journal of Science and Technology*, *9*(46). https://doi.org/10.17485/ijst/2016/v9i46/107369

Anexos

a. Imágenes

Figura 21-Encuestas

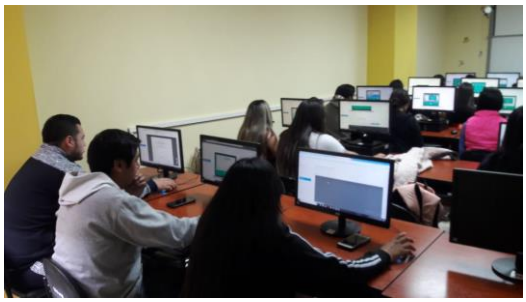


Figura 22-Encuestas

b. Cuestionario

Universidad Técnica de Ambato

Carrera de Docencia en Informática

A. Instrucciones

- Esta encuesta es anónima y personal, lea las preguntas planteadas y conteste de acuerdo con lo que considera verdad para usted mismo.
- Agradecemos dar su respuesta con la mayor transparencia y veracidad a las diversas preguntas del cuestionario.
- Para responder, señale con una X en el espacio correspondiente a su elección

B. Conceptos

- **Regresión logística**: A partir de un conjunto de variables como las calificaciones previas, condición socioeconómica, se puede predecir si un estudiante aprobará o no un semestre. (Ver ejemplo: Tabla)

Nota 1	**Nota 2**	**Nota 3**	**Aprueba**
8.6	4.2	4.1	No
8.9	9.4	8.1	Si

C. Cuestionario

Pregunta	**Siempre**	**A veces**	**No**
1) ¿Cree usted que la regresión logística (explicada previo a la aplicación del cuestionario), es un buen mecanismo de predicción de deserción estudiantil?			
2) ¿Considera usted que es importante la investigación para detectar a posibles desertores de la carrera?			
3) ¿Conocer a tiempo las causas que provocan la deserción ayuda a las autoridades y estudiantes a prevenir el abandono estudiantil?			
4) ¿Considera usted que las calificaciones son uno de los factores que mayor deserción estudiantil provocan?			
5) ¿Cree que las calificaciones de los primeros semestres inciden significativamente en la deserción de abandono de estudios universitarios?			
6) ¿Considera el estado civil del estudiante como un factor que influye en la deserción?			
7) ¿Considera que los hombres abandonan los estudios más que las mujeres?			
8) ¿Considera que las mujeres abandonan los estudios más que los hombres?			

c. Manual de aplicación.

Manual de Usuario

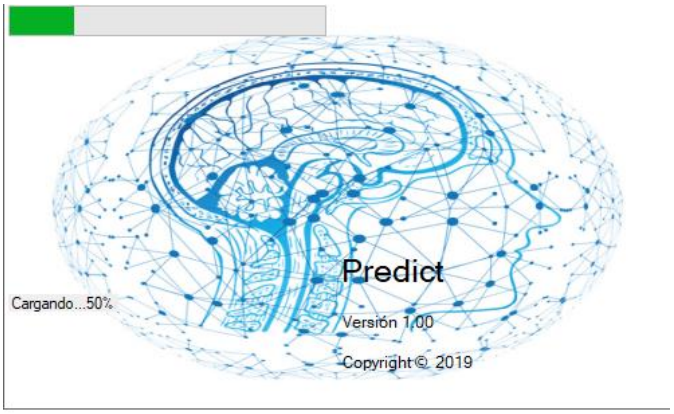
"Predict"

Introducción

El objetivo primordial de este Manual es ayudar y guiar al usuario a utilizar el programa para detectar a posibles estudiantes desertores; y comprende:

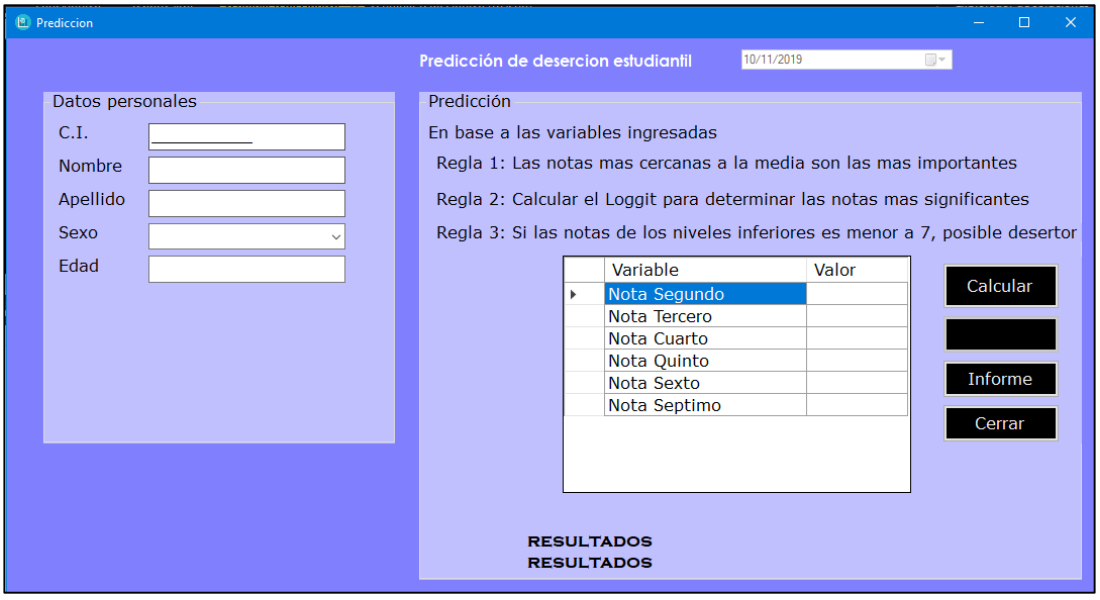
Bienvenida

Como primera pantalla al ejecutar el sistema, se abre una máscara de bienvenida, la cual carga los componentes del programa, esta muestra la versión, nombre y año de lanzamiento del programa.



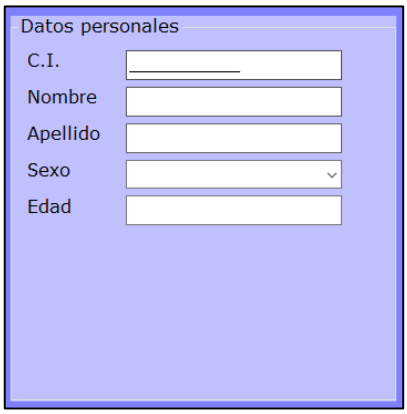
Pantalla principal

En esta pantalla, el sistema permite que nuevos usuarios analicen las variables relacionadas con sus calificaciones para predecir la deserción, así como el botón para acceder a la ventana adicional de reportes

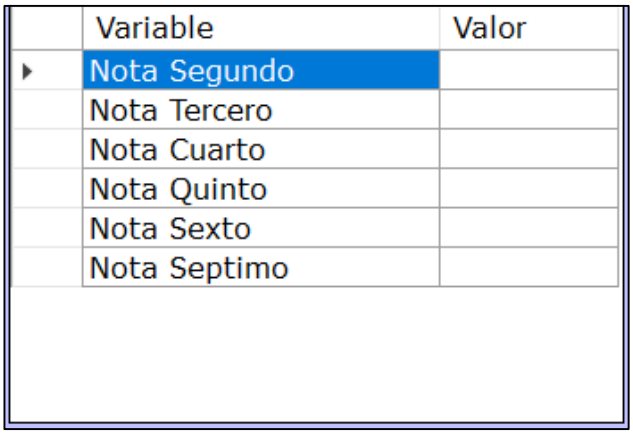


En la ventana principal, se muestra la fecha y las reglas de predicción del modelo, siga los siguientes pasos para analizar un nuevo conjunto de variables relacionadas a un usuario.

- En la parte izquierda de la pantalla, llene los campos que se piden, con la información personal: cédula, nombre, apellido, sexo y edad; de la persona cuyos datos van a ser analizados.

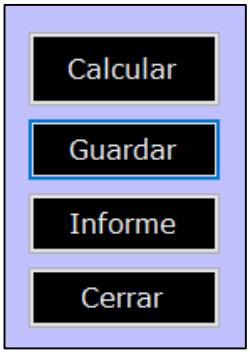


- Ingrese las calificaciones correspondientes a los niveles que se piden en el programa, para las calificaciones con decimales, utilizar punto(.) para separar los enteros de los decimales.



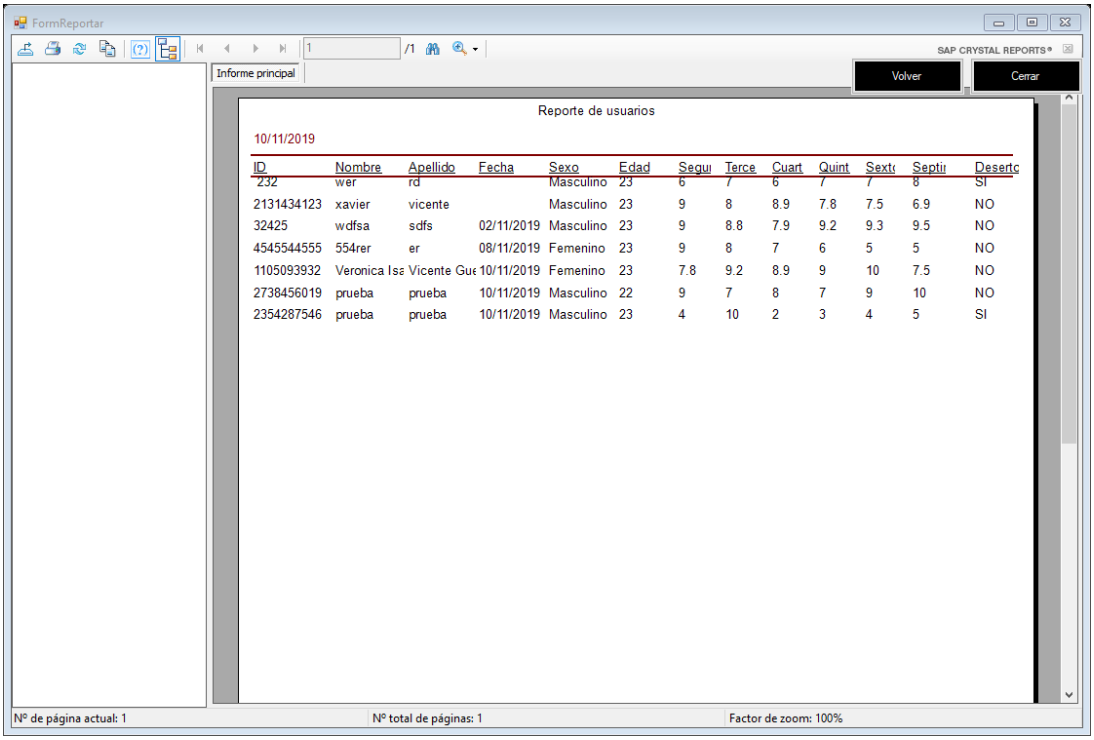
- Una vez ingresadas las calificaciones, presiona el botón calcular el cual, a partir de las instrucciones programadas, muestra los resultados obtenidos al usuario.

- Cuando se ha presionado el botón calcular, se habilita la opción de guardar los datos en la base (Ver figura), así también se puede acceder al informe de usuarios del programa o cerrar la aplicación



- En la parte inferior se muestra los resultados que el programa encontró en los análisis

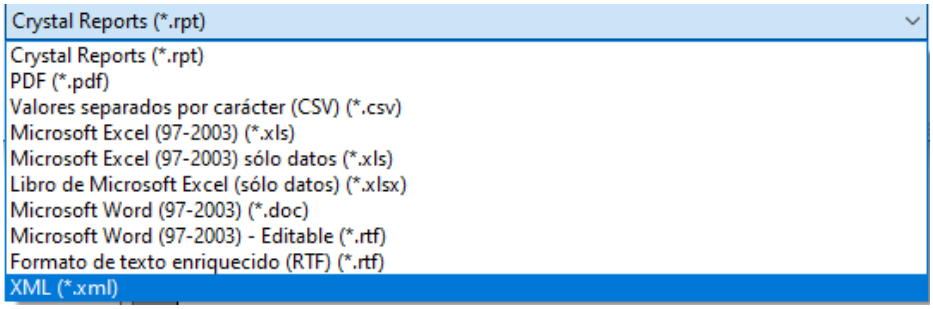
Informe



En la pantalla del informe se muestra los datos de los usuarios que han utilizado el programa y los resultados de los análisis. Se puede regresar a la ventana de predicción o cerrar la aplicación.

Los informes presentan una cinta de opciones en las que se puede navegar por los registros, buscar un dato en particular, exportar los informes, imprimir, actualizar los registros y copiar.

Los informes se pueden exportar en diferentes formatos.



Ayuda

Esta opción nos permitirá visualizar el manual de ayuda que posee la aplicación para poder manipular cada una de las opciones existentes dentro del programa.