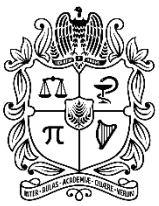


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Propuesta de un modelo estadístico para caracterizar y predecir la deserción estudiantil Universitaria

Jorge Iván Madrid Echeverry

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ingeniería de la Organización

Medellín, Colombia

2017

Propuesta de un modelo estadístico para caracterizar y predecir la deserción estudiantil Universitaria

Ing. Jorge Iván Madrid Echeverry

Tesis presentada como requisito parcial para optar al título de:

Magister en Ingeniería Administrativa

Director:

Ph.D. Henry Laniado Rodas

:

Universidad Nacional de Colombia

Facultad, Departamento de Ingeniería de la Organización

Medellín, Colombia

2017

“La simplicidad es lo más difícil de conseguir en este mundo, es el último límite de la experiencia y el último esfuerzo del genio.”

George Sand

Agradecimientos

Quiero expresar mis agradecimientos a mi tutor, el Profesor Henry Laniado por su compromiso, apoyo constante y por la disposición que siempre tuvo para ayudarme en el desarrollo de este trabajo.

A la Profesora Diana Luz Ceballos Gómez, por el apoyo y confianza que me brindo al permitirme acceder a la información.

A la Universidad Nacional de Colombia por darme la oportunidad de adelantar mis estudios de Maestría.

Resumen

El objetivo de la investigación se fundamentó en utilizar técnicas estadísticas multivariadas: Máquinas Vector Soporte (SVM), Análisis Discriminante (AD), K-vecinos más próximos (KNN) y Regresión Logística (RL) para clasificar a los estudiantes de pregrado de la Universidad Nacional de Colombia Sede Medellín en dos poblaciones (con posibilidad o no de desertar) a partir de la información que se tenía disponible de las variables definidas e identificadas como determinantes de la deserción estudiantil Universitaria.

Para el estudio se utilizó la información que suministraron los estudiantes que ingresaron a la Universidad Nacional de Colombia Sede Medellín desde el primer semestre del año 2009 hasta el primer semestre del año 2016, su correspondiente rendimiento académico en cada periodo matriculado y la identificación de cuáles de ellos perdieron la calidad de estudiante en la Universidad por bajo rendimiento y cuáles continuaron con sus estudios. Lo que permitió contar con un porcentaje de datos que fueron utilizados para el entrenamiento de los modelos y el resto de los datos como validación.

Los resultados permitieron identificar la técnica que permite obtener el modelo con menor porcentaje de error y mayor sensibilidad, y que podría ser utilizada para hacer predicciones de deserción en nuevos individuos a partir de la información de las variables seleccionadas.

Palabras clave: Deserción Universitaria, Estadística Multivariada, Máquina Vector Soporte, Análisis Discriminante, Regresión logística.

Abstract

The objective of the research was based on the use of multivariate statistical techniques: support vector machines (SVMs), Discriminant Analysis, k-nearest neighbors (kNN) and Logistic Regression for classify the pregrade students of the Universidad Nacional de Colombia Sede Medellin in two Populations (with or without possibility of deserting) taking the information that was available of the variables defined and identified as determinants of student dropout

For the study was used the information supplied by the students that entered in the National University of Colombia in Medellin from the first semester of 2009 until the second semester of 2016, their corresponding academic performance in each registered period and the identification of which of them lost de student quality in the university for poor performance and which of them continued with their studies. This allowed that was used a percentage of data for the training of the models and the rest of the data as validation.

The results allowed identify the technique that allows obtain the model with lower percentage of error and greater sensitivity, and that could be used to make predictions of desertion in new individuals from the information of the selected variables.

Keywords: University Desertion, Multivariate Statistics, Support Vector Machines, Discriminant Analysis, Logistic Regression.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

La deserción estudiantil ha sido un problema que ha estado presente en la mayoría de Instituciones de Educación Superior (IES) de Colombia y que ha sido objeto de un número considerable de investigaciones y estudios que han buscado hacer una caracterización de los factores que influyen en ella.

La deserción tiene consecuencias sociales al no cumplirse las expectativas de los estudiantes y sus familias; y consecuencias económicas tanto para las personas como para el sistema en su conjunto. En el año 2009 la deserción le costó a Colombia 778 mil millones de pesos, cifra que represento el 43% de las transferencias del Estado a las universidades públicas, de esta cifra, 221 mil millones representan la inversión de las Instituciones de educación superior (IES) públicas en desertores, 337 mil millones fue lo que invirtieron las familias de los desertores y 220 mil millones lo que dejaron de recibir en matrículas las IES privadas. (MEN, 2013).

Según las estadísticas de deserción y graduación presentadas en el Sistema para la Prevención de la Deserción de la Educación Superior (SPADIES) en el año 2015 la tasa de deserción universitaria en Colombia fue del 46.1%, es decir que de cada 100 estudiantes que son admitidos a una IES aproximadamente la mitad no culminan sus estudios universitarios.

En la Universidad Nacional de Colombia (UNAL) según la información presentada en el plan Global de desarrollo 2016-2018 (UNAL 2015 p.114), muestran que el 36% (34 de 95) de los programas de pregrado presentan deserciones acumuladas al semestre 15 superiores a un 50%. Se presenta en el 14% (7/49) de los programas de pregrado de la sede Bogotá, en el 70% (19/27) de los programas de pregrado de la sede Medellín, en el 42% (5/12) de los programas de la sede Manizales y en el 43% (5/12) de los programas de pregrado de la sede Palmira.

Haciendo una revisión a los índices de deserción que tiene la Sede Medellín, en comparación con las demás Sedes de la UNAL, toma una gran importancia el poder establecer mecanismos que permitan la identificación de los estudiantes que podrían tener riesgo de deserción, con el fin de que la institución a partir de integración de éstos estudiantes en los programas de bienestar y de apoyo estudiantil busque la disminución de las tasas de deserción.

El trabajo se desarrolló en 5 capítulos, en el primer capítulo, se presentan los antecedentes del problema de la deserción en la Universidad Nacional y más específicamente en la Sede Medellín, en el segundo capítulo se aborda el marco teórico sobre deserción, se examina la definición de deserción desde varias perspectivas y se hace una identificación de los factores que más inciden en que se presente este fenómeno tomando en cuenta las investigaciones previas. Luego se hace una introducción teórica a la definición de las técnicas estadísticas multivariadas de Máquina Vector Soporte, Análisis Discriminante, k-vecinos más próximos y Regresión Logística. En el tercer capítulo se hace la identificación de las variables y se presenta un análisis exploratorio de los datos de cada una de las variables analizadas desde el punto de vista de la deserción. En el cuarto capítulo se presenta los resultados obtenidos en la aplicación de cada uno de los modelos entrenados con las técnicas multivariantes y se hace la selección de los modelos que mejor podrían predecir la deserción en cada uno de los semestres y por último en el capítulo 5 se presentan las conclusiones derivadas de la investigación.

1 Antecedentes

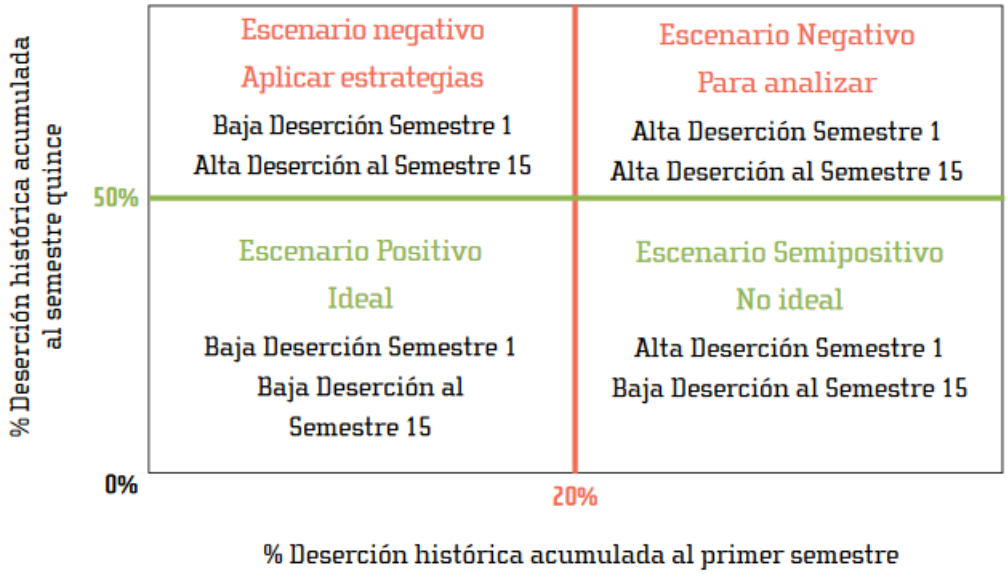
El Centro de Estudios para el Desarrollo Económico de la Universidad de los Andes (CEDE) creó una herramienta informática que permite hacer el seguimiento al problema de la deserción en la educación Superior denominada Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior (SPADIES). El cual según la información del MEN (2007) "*consolida y ordena información que permite hacer seguimiento a las condiciones académicas y socioeconómicas de los estudiantes que han ingresado a la educación superior en el país. De esta manera, permite conocer el estado y evolución de la caracterización y del rendimiento académico de los estudiantes, lo cual es útil para establecer los factores determinantes de la deserción, para estimar el riesgo de deserción de cada estudiante y para diseñar y mejorar las acciones de apoyo a los estudiantes orientados a fomentar su permanencia y graduación."* Pero en un informe presentado por centro de investigaciones y documentación socioeconómica (CIDSE, 2011) advierten que, en términos de la cuantificación del fenómeno, el SPADIES cumple con su objetivo, es una herramienta que además permite realizar un análisis descriptivo. Sin embargo, argumentan que, tanto desde el punto de vista conceptual como metodológico, se presentan serias limitaciones en su capacidad predictiva y en especial cuando se tratan de hacerlas a nivel desagregado para cada institución. Las estimaciones de los riesgos la obtienen utilizando la información suministrada, asumiendo que las IES no presentan diferencias entre ellas y que tampoco existe una diferenciación respecto a los programas académicos que se imparten en las instituciones.

Tomando en cuenta esas advertencias cobra sentido hacer la búsqueda del mejor modelo estadístico que permita determinar los estudiantes que están en riesgo de desertar y que se ajuste a las características particulares de la IES que sería objeto de estudio y los factores identificados como determinantes en la deserción.

La UNAL en su plan Global de desarrollo 2016-2018 (UNAL 2015), presentó cifras estadísticas sobre el comportamiento de la deserción estudiantil para los programas de pregrado de cada una de sus Sedes, para ello utilizaron información extraída del SPADIES y emplearon el enfoque de medición por cohortes, la cual estima las probabilidades de deserción desde el primer semestre hasta el semestre quince, después de haber sido admitido un estudiante en la universidad.

Para hacer la evaluación de los resultados presentan 4 posibles escenarios que se presentan en la figura 1-1 en la cual se presentan escenarios para la evaluación de los resultados del análisis de la deserción definitiva por cohorte en la Universidad Nacional de Colombia.

Figura 1-1 Comparación deserción histórica acumulada al primero y al semestre quince.



Fuente: UNAL (2015).

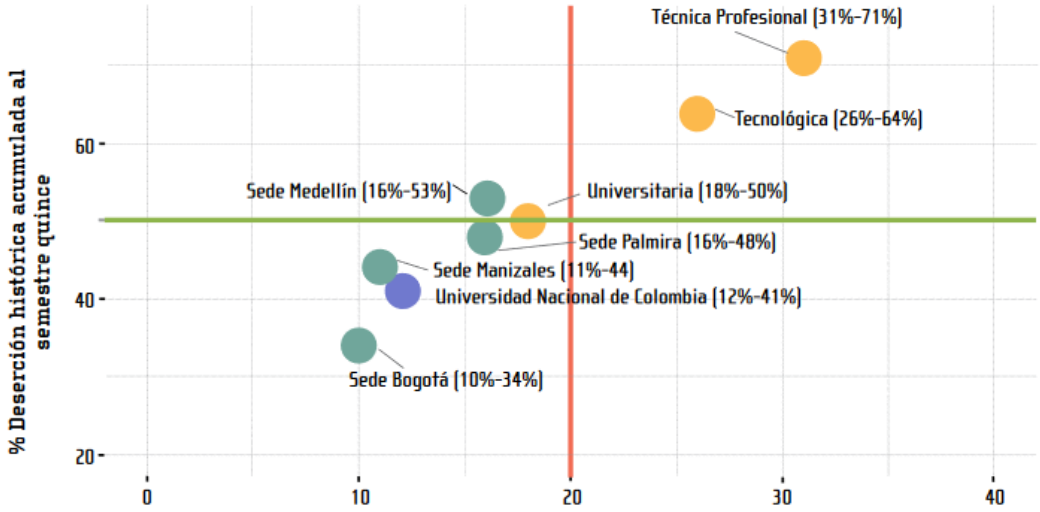
1. Escenario Positivo (ideal): universidades, sedes de la universidad y programas académicos de pregrado que cuentan con una deserción definitiva acumulada al semestre 15 menor al 50% y una deserción definitiva acumulada al primer semestre menor o igual al 20%
2. Escenario semipositivo (no ideal): universidades, sedes de la universidad y programas académicos de pregrado que cuentan con una deserción definitiva

acumulada al semestre 15 menor o igual a un 50 % y una deserción definitiva acumulada a primer semestre mayor del 20 %.

3. Escenario negativo (aplicar estrategias): universidades, sedes de la universidad y programas académicos de pregrado que cuentan con una deserción definitiva acumulada al semestre 15 mayor a un 50 % y una deserción definitiva acumulada a primer semestre menor o igual al 20%
4. Escenario negativo (para analizar): universidades, sedes de la universidad programas académicos de pregrado que cuentan con una deserción definitiva acumulada al semestre 15 mayor a un 50% y una deserción definitiva acumulada a primer semestre mayor del 20%

En la Figura 1-2 se muestra la tasa histórica acumulada de deserción definitiva al primero y al semestre 15 del país, de la UNAL y de cada una de sus Sedes Andinas (Bogotá, Medellín, Manizales y Palmira). En este punto llama la atención que la Sede Medellín es la única Sede de la UNAL que se encuentra ubicada en el escenario negativo y la que tiene una mayor deserción tanto en el primer semestre (16%) como en el semestre 15 (53%)

Figura 1-2 Tasa histórica acumulada de deserción definitiva al primero y al semestre 15: país, Universidad Nacional de Colombia y sedes andinas de la Universidad

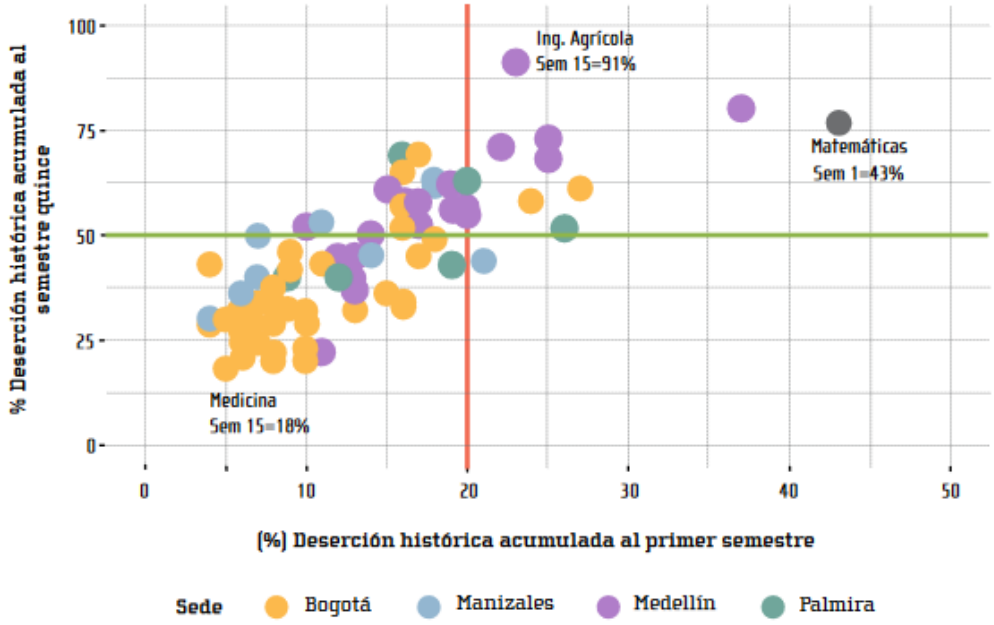


Fuente: UNAL (2015).

En la evaluación de la deserción por programas (Figura 1-3) muestran que el 36% (34 de 95) de los programas de pregrado presentan deserciones acumuladas al semestre 15

superiores a un 50%. Se presenta en el 14% (7/49) de los programas de pregrado de la sede Bogotá, en el 70% (19/27) de los programas de pregrado de la sede Medellín, en el 42% (5/12) de los programas de la sede Manizales y en el 43% (5/12) de los programas de pregrado de la sede Palmira. Lo que implica que la Sede Medellín se presenta aproximadamente el 56% de la deserción a pesar de que solo tiene el 28% de los programas de pregrado la universidad. Tomando en cuenta estas cifras de deserción de la UNAL, el estudio se va a centrar en analizar los datos correspondientes a los estudiantes de la Sede Medellín, para lo cual se tomará la información de los estudiantes que ingresaron desde el primer semestre de 2009 hasta el segundo semestre de 2016.

Figura 1-3 Deserción Acumulada por Sede en el primer semestre



Fuente: UNAL (2015).

2 Marco Teórico

2.1 Conceptos sobre deserción

La deserción estudiantil es uno de los problemas que aborda la mayoría de las instituciones de educación superior de toda Latinoamérica (UNESCO, 2004). Es un fenómeno que ha sido estudiado por diferentes autores e Instituciones de Educación Superior (IES). Dentro de los primeros trabajos que se encuentran en la literatura están los modelos sociológicos desarrollados por Bean en 1980 y Spady 1970 y Tinto 1975 (MEN 2009, p. 25), en los que explican los motivos por los cuáles los estudiantes deciden abandonar una IES a partir de dos conjuntos de factores: los factores ajenos a la institución y el grado de integración del estudiante con el ambiente académico y social de la institución. Según el MEN (2009), en Colombia, los estudios se han centrado en su gran mayoría en un desarrollo conceptual y metodológico y una identificación de las variables que inciden en la deserción estudiantil.

Para Tinto (1989), la deserción se puede definir desde tres puntos de vista: desde el punto de vista individual, el cual debe referirse a las metas y propósitos que tienen las personas al incorporarse al sistema de educación superior en el que desertar significa el fracaso para completar un determinado curso de acción o alcanzar una meta deseada; Pero aun con un compromiso adecuado del estudiante, en las IES se tiene un nivel de exigencia intelectual y habilidades sociales mayores que los requeridos en la educación media. A esto se suma el hecho de que según las estadísticas presentadas en el 2015 por el programa para la Evaluación Internacional de Alumnos (PISA) el cual es realizado por la Organización para la Cooperación y el Desarrollo Económicos (OCDE), Colombia está entre los países de Latinoamérica que presenta bajos niveles de rendimiento en las áreas de matemáticas, lectura y ciencias. Para Tinto (1989), las habilidades sociales son de igual importancia que las intelectuales para la retención en la universidad. Ya que le permiten al alumno localizar y utilizar los recursos disponibles en la institución e interactuar con ellos, afirma que la carencia de habilidades sociales, en especial entre los sectores desfavorecidos del

estudiantado, aparece como un factor importante en relación con el fracaso para mantener un nivel adecuado de rendimiento académico.

El segundo punto de vista es el institucional en el que la deserción se basa en la identificación de cuáles formas de abandono merecen atención y que conlleven por lo tanto a tomar acciones institucionales. Existen varios periodos críticos en los que el riesgo de deserción es más alto: el primero se presenta en la etapa de admisión en la que el estudiante tienen el primer contacto con la universidad y en el que se podría presentar una deserción debido a que por alguna causa el estudiante no continúe con su proceso de admisión, y el segundo se presenta en los primeros semestres del programa, en los cuáles se lleva a cabo el proceso de adaptación social y académica y en el que algunos estudiantes no logran buena adaptación o deciden retirarse. En los semestres siguientes si bien se presenta deserción, en términos generales esta tiende a disminuir de un semestre a otro.

Por último, desde el punto de vista estatal o nacional en la que solo aquellas formas de abandono universitario que significan a la vez el abandono de todo el sistema formal de educación superior son probablemente consideradas como deserciones, ya que las transferencias interinstitucionales pueden considerarse como migraciones internas de alumnos dentro del sistema educativo.

La definición de deserción puede variar según las partes interesadas y los objetivos y áreas en la que se esté analizando. El MEN adopto para la medición y seguimiento de la deserción la siguiente definición: *"situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos lo cual equivale a un año de inactividad académica. En algunas investigaciones este comportamiento se denomina como "primera deserción" (first drop-out) ya que no se puede establecer si pasado este periodo el individuo retomará o no sus estudios o si decidirá iniciar otro programa académico"* (MEN 2009, p. 22).

2.1.1 Factores determinantes de la deserción

Castaño, Gallón, Gómez y Vásquez (2007, citados por MEN 2009), presentaron un resumen del estado del arte de los factores determinantes de la deserción en el cual resumieron los autores y perspectivas del análisis en el estudio de la deserción y agruparon las variables más utilizadas en cuatro categorías: individuales, Académicos, Institucionales y socioeconómicos.

Tabla 2-1 Factores determinantes de la deserción

DETERMINANTES DE LA DESERCION			
INDIVIDUALES	**ACADÉMICOS**	**INSTITUCIONALES**	**SOCIOECONOMICOS**
• Edad, género y estado Civil • Posición dentro de los hermanos • Entorno familiar • Calamidad y problemas de salud • Integración social • Incompatibilidad Horaria con actividades extra académicas • Expectativas no satisfechas • Embarazo	• Orientación profesional • Tipo de colegio • Rendimiento Académico • Calidad del Programa • Métodos de estudio • Resultado examen de ingreso • Insatisfacción con el programa u otros factores • Número de materias	• Normalidad Académica • Becas y Formas de financiamiento • Recursos universitarios • Orden público • Entorno político • Nivel de interacción personal con profesores • Apoyo académico • Apoyo psicológico	• Estrato • Situación laboral • Situación Laboral de los padres e ingresos • Dependencia económica • Personas a cargo • Nivel educativo de los padres • Entorno macroeconómico del país

Adaptado de Castaño, et al. (Citados por MEN,2009),

El MEN (2014) afirma que existe consenso en que la deserción estudiantil es el resultado del efecto no de una sola categoría, sino del efecto individual y de la interacción de diferentes categorías de factores. Estos factores serán utilizados dentro del desarrollo de la investigación, pero las categorías o variables que se utilizarán serán las que están disponibles en el Sistema de información Académico (SIA) que dispone en la Sede Medellín en los resultados recopilados de los procesos de Admisión a la Universidad y en los datos extraídos del ICFES, los cuáles serán descritos más adelante.

2.2 Análisis Multivariante

Cuadras (2014), define el análisis multivalente (AM) como la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos. Las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás.

2.2.1 Técnicas de Análisis Multivariante

El análisis multivariante de acuerdo a la finalidad que se busque, se puede dividir en tres grupos: Técnicas funcionales o de dependencia, Técnicas estructurales y Técnicas de interdependencia.

En los métodos de dependencia las variables analizadas están divididas en dos grupos: las variables dependientes y las variables independientes. El objetivo de estos métodos consiste en determinar si un conjunto de variables independientes afecta al conjunto de variables dependientes y de qué forma.

En los métodos estructurales las variables están divididas en dos grupos: las variables dependientes y las variables independientes. El objetivo de estos métodos es analizar, no sólo como las variables independientes afectan a las variables dependientes, sino también cómo están relacionadas entre sí.

En los métodos de Interdependencia, no se diferencia entre variables dependientes e independientes y su objetivo es identificar las variables que se encuentran relacionadas, como los están y porqué. (Salvador, 2000)

2.2.2 Métodos de Clasificación

La clasificación es una de las metodologías fundamentales de la estadística multivariada y consiste en encontrar un modelo matemático capaz de reconocer la pertenencia de un objeto a una clase. Una vez el modelo de clasificación ha sido obtenido, la pertenencia de

nuevas observaciones a una clase definida puede ser predicha. En estos casos, adicional a la matriz tradicional de datos, se cuenta con un vector de clases el cual asigna la pertenencia a cada observación dada en la matriz, ya sea un origen geográfico o un atributo cualitativo (bueno/malo), (si/no). (ZULUAGA, 2011)

Existes dos tipos de sistemas de clasificación, los sistemas de clasificación supervisados los cuáles a partir de un conjunto de datos clasificados (datos de entrenamiento), asignan una clasificación a un segundo conjunto de datos (datos de validación o nuevos datos) y los sistemas de clasificación no supervisados que son aquellos en los que no se cuenta con datos clasificados previamente, sino que únicamente a partir de las propiedades de los datos se les asigna una agrupación según su similitud. (Sancho, 2017)

Para el desarrollo de la presente investigación se van a emplear cuatro técnicas o métodos de clasificación supervisada que son: Regresión Logística (RL), las máquinas vector soporte (SVM), el Análisis Discrimínate (AD) y k-vecinos más próximos (KNN).

2.2.2.1 Regresión Logística

El análisis de regresión se utiliza para modelar relaciones entre variables y para realizar pronósticos o predicciones de respuestas a partir de variables explicativas y de este modo crear un modelo donde se seleccionan las variables que están vinculadas con la respuesta, descartando aquellas que no aportan información. Además, permite detectar interacciones entre las variables independientes que afectan a la variable dependiente o predicha. La regresión logística es un tipo especial de regresión que se utiliza para explicar una variable dicotómica en función de variables independientes que pueden ser cuantitativas o cualitativas. (Castejón, 2011)

El modelo de regresión logística establece la relación entre la probabilidad de que ocurra el suceso dado que el individuo presenta los valores $(X = x_1, X = x_2, \cdots, X = x_k)$:

$$P[Y = 1 / x_1, x_2, \cdots x_k] = \frac{1}{1 + e^{(-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_k x_k)}} \quad (2\text{-}1)$$

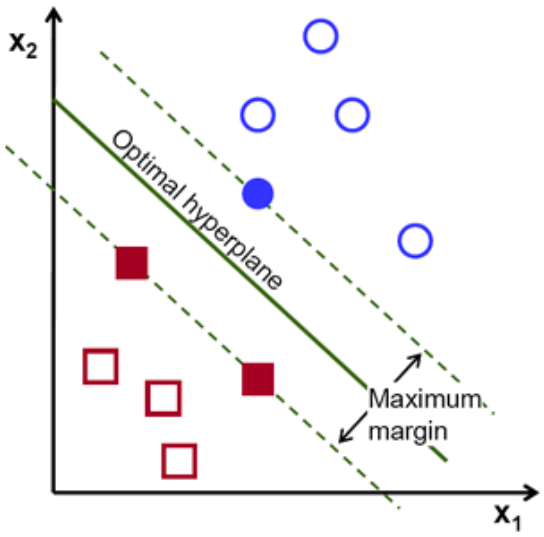
El objetivo es hallar los coeficientes $(\beta_0, \beta_1, ..., \beta_k)$ que mejor se ajusten a la expresión. El procedimiento de estimación de estos coeficientes, se basa en el método de máxima verosimilitud.

2.2.2.2 Máquinas Vector Soporte

La SVM se basan en los fundamentos de la teoría de aprendizaje estadístico desarrollada por Vapnik y pertenecen a la categoría de los clasificadores lineales ya que establecen separadores lineales o hiperplanos, ya sean en el espacio original de los datos de entrada, o en un espacio transformado. (Carmona,2014)

Dado un conjunto de vectores $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ donde $x_i \in R^d$ y $y_i \in \{-1, 1\}$ para $i = 1, \cdots, n$ se define separable si existe algún hiperplano en R^d que separa los vectores $X = \{x_1, \cdots, x_n\}$ con etiqueta $y_i = 1$ de aquellos con etiqueta $y_i = -1$

Figura 2-1 Hiperplano separador en R^2 SVM



Fuente: OpenCV dev team (2014)

La ecuación del Hiperplano está dada por $f(x) = x'\beta + b = 0$ donde $\beta \in R^d$ y b es un número real. Si los datos no son separables linealmente en el espacio se pueden utilizar funciones Kernel (no lineales).

- Polinomial

$$k(x_i, x_j) = (x_i . x_j)^n \quad (2\text{-}2)$$

- Radial Gaussiana

$$k(x_i, x_j) = e^{\frac{(x_i . x_j)^2}{2\sigma^2}} \quad (2\text{-}3)$$

Estas funciones resuelven el problema de clasificación trasladando los datos a un espacio donde el hiperplano de solución es lineal y, por tanto, más sencillo de obtener. Una vez conseguido, la solución se transforma, de nuevo, al espacio original. El SVM se hace inicialmente con datos de entrenamiento para los cuáles ya se conoce el valor de la variable dependiente, luego de este entrenamiento, se define un modelo que permitirá clasificar otro individuo. (Carmona,2014)

2.2.2.3 Análisis Discriminante

El Análisis Discriminante es una técnica estadística multivariante cuya finalidad es analizar si existen diferencias significativas entre grupos de individuos respecto a un conjunto de variables medidas, para que en el caso de que existan, explicar en qué sentido se dan y facilitar procedimientos de clasificación de nuevas observaciones en uno de los grupos analizados. (Fernández,2011).

Partiendo de *q* grupos donde se asignan a una serie de objetos y de *p* variables medidas sobre ellos ($x_1, x_2, \ldots, x_p$), se trata de obtener para *cada objeto* una *serie de puntuaciones* que indican el grupo al que pertenecen ($y_1, y_2, \ldots, y_m$), de modo que sean funciones lineales de $x_1, x_2, \ldots, x_p$

$$\begin{aligned} y_1 &= w_{11}x_1 + w_{12}x_2 + \ldots + w_{1p}x_p + w_{10} \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_m &= w_{m1}x_1 + w_{m2}x_2 + \ldots + w_{mp}x_p + w_{m0} \end{aligned} \quad (2\text{-}4)$$

donde $m = mín(q-1, p)$ tales que discriminen o separen lo máximo posible a los *q* grupos. Estas combinaciones lineales de las *p* variables deben maximizar la varianza *entre* los grupos y minimizar la varianza *dentro* de los grupos.

La variabilidad total de la muestra se puede descomponer en variabilidad *dentro* de los grupos y *entre* los grupos. Para ello, se parte:

$$\text{cov}(x_j, x_{j'}) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \qquad (2\text{-}5)$$

se puede considerar la media de la variable x_j en cada uno de los grupos $I_1,...,I_q$, es decir,

$$\bar{x}_{kj} = \frac{1}{n_k}\sum_{i\in I_k} x_{ij} \qquad \text{para } k = 1,...,q \qquad (2\text{-}6)$$

De esta forma, la media total de la variable x_j se puede expresar como función de las medias dentro de cada grupo. Asi,

$$\sum_{i\in I_k} x_{ij} = n_k \bar{x}_{kj} \qquad (2\text{-}7)$$

entonces

$$x_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{n}\sum_{i=1}^{q}\sum_{i\in I_k}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \qquad (2\text{-}8a)$$

$$x_j = \frac{1}{n}\sum_{i=1}^{q} n_k \bar{x}_{kj} = \sum_{k=1}^{q}\frac{n_k}{n}\bar{x}_{kj} \qquad (2\text{-}8b)$$

Así,

$$\text{cov}(x_j, x_{j'}) = \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in I_k}(x_{ij} - \bar{x}_j)(x_{ij'} - \bar{x}_{j'}) \qquad (2\text{-}9)$$

Si en cada uno de los términos se pone:

$$(x_{ij} - \overline{x}_j) = (x_{ij} - \overline{x}_{kj}) + (\overline{x}_{kj} - \overline{x}_j) \tag{2-10}$$

$$(x_{ij'} - \overline{x}_{j'}) = (x_{ij'} - \overline{x}_{kj'}) + (\overline{x}_{kj'} - \overline{x}_{j'}) \tag{2-11}$$

Al simplificar se obtiene

$$\begin{aligned} \operatorname{cov}(x_j, x_{j'}) &= \frac{1}{n}\sum_{k=1}^{q}\sum_{i\in I_k}(x_{ij} - \overline{x}_{kj})(x_{ij'} - \overline{x}_{kj'}) + \sum_{k=1}^{q}\frac{n_k}{n}(\overline{x}_{kj} - \overline{x}_j)(\overline{x}_{kj'} - \overline{x}_{j'}) \\ \operatorname{cov}(x_j, x_{j'}) &= d(x_j, x_{j'}) + e(x_j, x_{j'}) \end{aligned} \tag{2-12}$$

Es decir, la covarianza total es igual a la covarianza dentro de grupos más la covarianza entre grupos.

La aplicación del Análisis Discriminante (AD) a la clasificación de individuos en el caso de que se puedan asignar solamente a dos grupos a partir de k variables discriminadoras. Fue resuelta por Fisher mediante su función discriminante:

$$D = w_1x_1 + w_2x_2 + ... + w_kx_k \tag{2-13}$$

Las puntuaciones discriminantes son los valores que se obtienen al dar valores a $(x_1, x_2, ...x_k)$ en la ecuación anterior.

Se busca obtener los coeficientes de ponderación w_j. Si se considera N observaciones, la función discriminante $D_i = w_1x_{1i} + w_2x_{2i} + ... + w_kx_{ki}$ para $\forall_i = 1, ..., N$. D_i es la puntuación discriminante correspondiente a la i-ésima observación. La función discriminante en forma matricial:

$$\begin{pmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{pmatrix} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{1N} & X_{2N} & \cdots & x_{kN} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{pmatrix} \tag{2-14}$$

Expresando el modelo en función de las desviaciones a la media resulta:

$$\begin{pmatrix} D_1 - \bar{d}_1 \\ D_2 - \bar{d}_2 \\ \vdots \\ D_N - \bar{d}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ x_{1N} & X_{2N} & \cdots & x_{kN} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{pmatrix} \quad (2\text{-}15)$$

Es decir $d = Xw$. La variabilidad de la función discriminante (suma de cuadrados de las desviaciones de las variables discriminantes con respecto a su media) se expresa como la suma de cuadrados explicada por esta función $d'd = w'X'Xw$, $X'X$ es una matriz simétrica que expresa las desviaciones cuadráticas respecto a la media de las variables (suma de cuadrados total).

Se puede descomponer en suma de cuadrados entre grupos F y suma de cuadrados dentro de los grupos V: $T = X'X$ será la matriz de suma de cuadrados y productos cruzados (varianzas y covarianzas) para el conjunto de observaciones

$$T = X'X = F + V \quad (2\text{-}16)$$

Con lo cual,

$$d'd = w'X'Xw = w'(F+V)w = w'Fw + w'Vw \quad (2\text{-}17)$$

Los ejes discriminantes vienen dados por los vectores propios asociados a los valores propios de la matriz $(V^{-1}F)$ ordenados de mayor a menor.

Las puntuaciones discriminantes corresponden con los valores obtenidos al proyectar cada punto del espacio k-dimensional de las variables originales sobre el eje discriminante.

Los coeficientes w se obtienen como

$$\max \lambda = \frac{w'Fw}{w'Vw} = \frac{\text{separacion entre grupos}}{\text{separacion dentro grupos}} \quad (2\text{-}18)$$

Para realizar la clasificación, se obtienen puntuaciones discriminantes d_i para cada observación introduciendo los correspondientes valores de las k variables en la función discriminante. Aplicando el criterio de clasificación:

$$\begin{aligned} d_i < C \ (d_i - C < 0) &\rightarrow \text{ pertenece al grupo I} \\ d_i > C \ (d_i - C > 0) &\rightarrow \text{ pertenece al grupo II} \end{aligned} \tag{2-19}$$

Otra forma de realizar la clasificación seria obtener funciones discriminantes para cada grupo y se clasifica la observación en el grupo en el que la función correspondiente arroja mayor valor.

2.2.2.4 K-vecinos más Próximos

La clasificación por K-NN clasifica las observaciones en función de su probabilidad de pertenecer a un grupo u otro. Los datos de entrenamiento, son vectores en un espacio característico multidimensional, cada dato está descrito en términos de *p* atributos considerando *q* clases para la clasificación. Los valores de los atributos del i - eximo dato se representan por el vector *p*-dimensional.

$$x_i = (x_{1i}, x_{2i}, \cdots, x_{pi}) \in X \tag{2-20}$$

La distancia es el criterio de comparación utilizado para determinar la clase a la cual pertenece una nueva observación (Ramos et al. 2004 p. 10), es decir se basa en la idea de que las nuevas observaciones serán clasificadas en la clase a la cual pertenezca la mayor cantidad de vecinos más cercanos del conjunto de entrenamiento más cercano a él.

Algunos de los modelos utilizados para medir la distancia son:

- Distancia Euclidiana

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2-21}$$

- Distancia de Manhattan

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i| \tag{2-22}$$

- Distancia de Chebychev

$$d(x,y) = \max_{i=1\cdots n} |x_i - y_i| \quad (2\text{-}23)$$

- Distancia del coseno

$$d(x,y) = ar\cos\left(\frac{x^T y}{\|x\|.\|y\|}\right) \quad (2\text{-}24)$$

- Distancia de Mahalanobis

$$d(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)} \quad (2\text{-}25)$$

- Distancia Spearman

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2\text{-}26)$$

Dependiendo del modelo de medición utilizado, los resultados pueden variar.

2.3 Detección de valores atípicos multivariantes

Peña (2005) define los datos atípicos como aquellas observaciones que parecen haberse generado de forma distinta al resto de los datos. En muchos casos, las observaciones multivariantes no pueden ser detectadas como valores extremos, cuando cada variable se considera de forma independiente. La detección de valores atípicos sólo es posible cuando se realiza un análisis multivariado y las interacciones entre las diferentes variables son comparadas dentro de las clases de datos. (Ben-Gal, 2005). Una de las técnicas utilizadas en la detección de outliers multivariantes es la distancia de Mahalanobis, la cual permite tener en cuenta las diferentes escalas entre las variables y las correlaciones entre estas, pero para dimensiones mayores no es recomendable utilizar únicamente la distancia de Mahalanobis, ya que si existen grupos de atípicos, pueden distorsionar la estimación del centro y la dispersión de los datos enmascarando los atípicos y quizás señalando como atípicos a puntos que no lo son Peña (2005).

Peña (2005) propone un procedimiento para detección de outliers a partir de la distancia de Mahalanobis que consiste en proyectar los datos (cualquier observación atípica multivariante debe aparecer como atípica al menos en la dirección de proyección definida

por la recta que une el centro de los datos con el dato atípico) buscando p direcciones ortogonales de máxima curtosis y p direcciones ortogonales de mínima curtosis eliminando provisionalmente los datos extremos en esas direcciones y calculando la media y la matriz de covarianzas con los datos no sospechosos para después identificar los datos atípicos como aquellos que son extremos con la distancia de Mahalanobis calculada con las estimaciones no contaminadas.

El procedimiento propuesto por Peña y que se utilizó en el desarrollo de la investigación es el siguiente:

Dada la muestra multivariante (x1… xn).

Sean $\overline{\boldsymbol{x}}$ y S el vector de medias y la matriz de covarianzas de los datos. Estandarizar los datos de forma multivariante y sean $\boldsymbol{z_i} = \boldsymbol{s_x}^{-\frac{1}{2}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}})$ los datos estandarizados con media cero y matriz de covarianzas identidad. Tomar $\boldsymbol{j = 1\ y\ z_i^{(1)} = z_i}$

Calcular la dirección dj con norma unidad que maximiza el coeficiente de curtosis univariante de los datos proyectados. Llamando $\boldsymbol{y_i^{(j)} = d_j' z_i^{(j)}}$, a los datos proyectados sobre la dirección dj

Proyectar los datos sobre un espacio de dimensión $p - j$ definido como el espacio ortogonal a la dirección dj. Para ello tomar $\boldsymbol{zz^{(j+1)} = \left(I - d_j d_j'\right) z^{(j)}}$. Hacer $\boldsymbol{j = j + 1}$

Repetir (2) y (3) hasta obtener *p* direcciones, d1, … ,dp

Repetir (2) y (3) pero ahora minimizando la curtosis en lugar de maximizarla para obtener otras p direcciones dp+1,…,d2p

Considerar como sospechosos aquellos puntos que en alguna de estas 2p direcciones están claramente alejados del resto, es decir, verificar

$$\frac{\left|y_i^{(j)} - med(y^j)\right|}{Meda(y^j)} > 5 \qquad (2\text{-}27)$$

Donde $\boldsymbol{med}(\boldsymbol{y^j})$ se refiere a la mediana que es el estimador robusto del centro de los datos proyectados y $\boldsymbol{Meda}(\boldsymbol{y^j})$ es la mediana de las desviaciones absolutas de los datos proyectados $\left|\boldsymbol{y_i^{(j)}} - \boldsymbol{med}(\boldsymbol{y^j})\right|$ A continuación se eliminan todos los valores sospechosos detectados y se vuelve a (2) para analizar los datos restantes. La estandarización multivariante ahora se realiza con la nueva media y la matriz de covarianzas de los datos restantes. Los pasos (2) a (6) se repiten hasta que no se detecten más datos atípicos. Una vez la muestra no contenga más valores sospechosos con el procedimiento anterior, se calcula el vector de medias $\bar{x}_R$ y la matriz de covarianzas SR, de los datos no sospechosos, y las distancias de Mahalanobis para los sospechosos como:

$$\boldsymbol{d_R^2}(\boldsymbol{x_i}, \overline{\boldsymbol{x}}_{\boldsymbol{R}}) = (\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{R}})\boldsymbol{S_R^{-1}}(\boldsymbol{x_i} - \overline{\boldsymbol{x}}_{\boldsymbol{R}})' \qquad (2\text{-}28)$$

Los valores mayores que $p + 3\sqrt{2p}$ se consideran atípicos siendo *p* el número de variables utilizadas en la muestra. Para otras definiciones, se puede consultar Peña (2005).

2.4 Estandarización y transformación de variables

La estandarización de variables resulta muy útil para eliminar la dependencia de estas respecto a las unidades de medida empleadas. Equivale a una transformación lineal mediante la siguiente ecuación:

$$z = \frac{x-\mu}{\sigma} = \frac{1}{\sigma}x - \frac{\mu}{\sigma} \qquad (2\text{-}29)$$

Siendo $z = ax + b$ donde $a = \frac{1}{\sigma}$ y $b = \frac{\mu}{\sigma}$

La variable estandarizada, expresa el número de desviaciones típicas que dista de la media cada observación. Por ello, se puede comparar la posición relativa de los datos de diferentes distribuciones.[1]

[1] Extraído de http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/EDescrip/tema4.pdf

2.5 Validación de los resultados del modelo

En esta investigación se utilizó para evaluar el desempeño de los modelos la matriz de confusión, en esta matriz cada columna representa el número de predicciones de cada clase, y cada fila representa el número de eventos en la clase real.

Figura 2-2 Matriz de Confusión

		CLASIFICACIÓN DEL MODELO	
		Sin Deserción	Con Deserción
VALORES REALES	Sin Deserción	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Con Deserción	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuente: elaboración Propia

Los aciertos serán la proporción de predicciones verdaderas sobre el total de datos utilizados en la validación

$$Aciertos = \frac{VP+VN}{VP+FN+FP+VN} \quad (2\text{-}30)$$

La precisión en la clasificación de los estudiantes que no tendrían deserción (especificidad) viene dada por:

$$Especificidad = \frac{VP}{VP+FN} \quad (2\text{-}31)$$

La precisión en la clasificación de los estudiantes con deserción (Sensibilidad) viene dada por:

$$Sensibilidad = \frac{VN}{VN+FP} \quad (2\text{-}32)$$

La tasa de error será la proporción de Falsos positivos y Falsos negativos del total de datos analizados

$$Tasa\ de\ error = \frac{FN+FP}{VP+FN+FP+VN} = 1 - Aciertos \quad (2\text{-}33)$$

3 Metodología y Análisis de los datos

Para el desarrollo de la investigación, lo primero que se hizo fue una identificación de los factores determinantes de la deserción basado en los estudios realizados por varios autores (tabla 1-1) y de esos factores o variables se seleccionaron aquellas cuya información se podía extraer del Sistema de Información de la Universidad. Una vez identificados los factores o variables a utilizar, se realizó el análisis exploratorio y gráfico de los datos, haciendo énfasis en las características de los estudiantes que presentan deserción. Por último se hizo la identificación de los datos atípicos multivariantes y se aplicó cada una de las técnicas multivariantes, utilizando para ello una proporción de los datos (75%) para entrenamiento y el resto (25%) para la validación del modelo realizando este proceso n-veces para grupos diferentes de datos de entrenamiento y validación y seleccionando a partir de los resultados obtenidos la técnica que permitiese obtener el mejor rendimiento en la predicción de estudiantes con deserción en cada una de los semestres analizados.

Figura 3-1 Metodología para la aplicación de las Técnica Multivariantes

3.1. Identificación de las variables

Las variables que se identificaron y que fueron tomadas en cuenta para la investigación se clasificaron en cada uno de los factores individuales, académicos y socioeconómicos

3.1.1 Factores Individuales

Dentro los factores individuales que se pudieron identificar se encuentran la edad, el género y el estado civil

- Edad: La edad fue tomada como el número de años entre la fecha de nacimiento del estudiante y la fecha correspondiente al periodo de ingreso en el plan, dado que en la universidad existen dos periodos regulares por año: 01 y 03, para hacer el cálculo de la edad, el periodo 01 se tomó como fecha el 01 de febrero y para el periodo 03 el 01 de agosto del año correspondiente.

- Género: Es una variable categórica que toma dos posibles valores, Hombre y Mujer, éstos fueron identificados dentro de los datos con un 0 para las mujeres y un 1 para los hombres.

- Estado Civil: Se tomó como el estado reportado por el estudiante al momento de ser admitido a la universidad. A cada una de las categorías se le asignó un valor correspondiente el cual se ilustra en la tabla 3-1

Tabla 3-1 Categorías de la variable Estado Civil

CATEGORIA	VALOR
Casado	1
Soltero	2
Separado	3
Unión Libre	4
Viudo	5
Divorciado	6
No definido	7

3.1.2 Factores Académicos

- Tipo de colegio: el tipo de colegio corresponde a la clasificación dada por el ICFES a las instituciones educativas en el periodo en que los estudiantes presentaron el examen saber 11[2] Si el dato no estaba presente se tomó la última clasificación dada al colegio. A cada una de las categorías se le asignó un valor

Tabla 3-2 Categorías de la variable tipo colegio

CATEGORIA	VALOR
A+	1
A	2
B	3
C	4
D	5
No definido	6

- Rendimiento Académico: Para el rendimiento académico se tomó el valor del Promedio Aritmético Ponderado Acumulado (PAPA) que es la medida de rendimiento académico utilizada por la universidad[3]. Dado que para el primer semestre los estudiantes no tienen un valor de PAPA, esta variable fue considerada en el segundo y tercer semestre.

- Puntaje Componente Matemáticas: corresponde al puntaje obtenido por el estudiante al momento de presentar el examen de admisión a la universidad en el componente de matemáticas, el cual evalúa los temas de Pensamiento numérico, Pensamiento espacial y métrico, Pensamiento aleatorio y pensamiento Variacional[4]

[2] Se puede ampliar la información en la página web http://www.mineducacion.gov.co/cvn/1665/w3-article-343956.html

[3] Para calcular el PAPA se multiplica cada calificación definitiva por el número correspondiente de créditos de la asignatura cursada y luego se suman todos los productos anteriores y el resultado se divide por la suma total de créditos cursados.

[4] Extraído de http://admisiones.unal.edu.co/pregrado/panel-2-informacion-sobre-las-pruebas/prueba-de-admision/

- Puntaje Componente de Ciencias Naturales: corresponde al puntaje obtenido por el estudiante al momento de presentar el examen de admisión a la universidad en el componente de ciencias naturales, el cual evalúa los temas de Física, Química y Biología.

- Puntaje Componente de Ciencias Sociales: corresponde al puntaje obtenido por el estudiante al momento de presentar el examen de admisión a la universidad en el componente de ciencias naturales, el cual evalúa los temas de Historia, Geografía y Filosofía.

- Puntaje Componente de Análisis Textual corresponde al puntaje obtenido por el estudiante al momento de presentar el examen de admisión a la universidad en el componente de Análisis textual.

- Puntaje Componente de Análisis de Imagen corresponde al puntaje obtenido por el estudiante al momento de presentar el examen de admisión a la universidad en el componente de Análisis de imagen.

- Nivela Matemáticas Básicas: Corresponde a la información de si el estudiante quedo clasificado o no para nivelar matemáticas básicas según los resultados obtenidos en el examen de admisión.

- Nivela Lectoescritura: Corresponde a la información de si el estudiante quedo clasificado o no Lectoescritura según los resultados obtenidos en el examen de admisión.

- Resultado en Matemáticas Básicas: Corresponde a la información si el estudiante aprueba o no el curso de Matemáticas Básicas. Se le asignara un 1 para los que la perdieron y un 0 a los que la aprobaron.

- Resultado Lectoescritura: Corresponde si el estudiante aprueba o no el curso de Lectoescritura. Se le asignara un 1 para los que la perdieron y un 0 a los que la aprobaron.

- Asignaturas no aprobadas: corresponde al número de asignaturas no aprobadas durante el semestre inmediatamente anterior, se incluirá esta variable a partir del segundo semestre.

- Indicador de Eficiencia: a partir del segundo semestre se va a utilizar un indicador de eficiencia calculado como la razón entre el promedio de número de créditos aprobados por periodo (CAP) dividido en promedio de créditos que debe matricular (CRP), es decir:

$$CAP = \frac{CREDITOS\ APROBADOS}{NUMERO\ DE\ MATRICULAS} \quad (2\text{-}34)$$

$$CRP = \frac{CREDITOS\ REQUERIDOS\ PLAN}{NUMERO\ DE\ PERIODOS\ PLAN} \quad (2\text{-}35)$$

$$EFICIENCIA = \frac{CAP}{CRP} \quad (2\text{-}36)$$

El valor de CRP es una constante para los estudiantes de un mismo plan de estudio, el valor del CAP varía según el rendimiento del estudiante en los periodos matriculados.

3.1.3 Factores Socioeconómicos

Para los factores socioeconómicos se tomó la información suministrada por los estudiantes la cual es usada para calcular el valor de matrícula dentro de la Universidad, las variables que se consideraron fueron las siguientes:

- Valor de la pensión mensual: corresponde al valor de la pensión mensual pagada durante el último año de secundaria, expresado como un múltiplo del salario mínimo legal mensual vigente (SMLV) del año respectivo, los puntajes son asignados según la tabla 3-3[5]

[5] Extraído de http://www.legal.unal.edu.co/sisjurun/normas/Norma1.jsp?i=34298

Tabla 3-3 Cálculo del Puntaje asignado para el Valor de la pensión

Valor de la pensión en múltiplos de salarios mínimos mensuales (VP)	PUNTAJE
de 0 a 0.03	500 x VP
mayor de 0.03 a 0.10	15+ 300 x (VP- 0.03)
mayor de 0.10 a 0.20	36+ 240 x (VP- 0.10)
mayor de 0.20 a 0.30	69+ 100 x (VP- 0.20)
mayor de 0.30 a 0.50	70 + 80 x (VP- 0.30)
más de 0.50	86 + 50 x (VP- 0.50)
no informa	100

- Estrato: Corresponde al estrato del lugar de residencia del responsable de la manutención del estudiante

Tabla 3-4 Categorías de la variable estrato

ESTRATO	PUNTAJE
1	0
2	13
3	30
4	55
5	80
6	100
7 (NO INFORMA)	100

- Ingresos (A3): De acuerdo con los ingresos familiares mensuales hasta un máximo de 100 puntos, según la expresión: A3 = 9xING donde ING son los ingresos mensuales expresados como un múltiplo del salario mínimo legal mensual vigente en el año respectivo.

- Carácter del colegio (B1): indica el tipo de colegio en el que el estudiante termino sus estudios de secundaria

Tabla 3-5 Categorías de la variable carácter del Colegio

CARÁCTER DEL COLEGIO	PUNTAJE
Plantel Oficial	0
Plantel Privado	1

- Lugar de Residencia (B2): corresponde al lugar de residencia de los padres o responsables de la manutención del estudiante

Tabla 3-6 Categorías de la variable Lugar de Residencia

LUGAR DE RESIDENCIA	PUNTAJE
Fuera del perímetro urbano	0.80
Dentro del perímetro urbano	1

- Propiedad de la vivienda (B3): Corresponde al estado de propiedad de la vivienda donde habitan los responsables de la manutención del estudiante.

Tabla 3-7 Categorías de la variable Propiedad de la Vivienda

PROPIEDAD DE LA VIVIENDA	PUNTAJE
Sin propiedad raíz	0.9
Pagando crédito Hipoteca	0.95
Vivienda propia	1

- Número de hijos dependientes (B4): corresponde al número de hijos menores de 18 años o estudiantes regulares que dependen económicamente de los responsables de la manutención del estudiante.

Tabla 3-8 Categorías de la variable Número de Hijos

NÚMERO DE HIJOS DEPENDIENTES DEL INGRESO FAMILIAR	PUNTAJE
Siete o más	0.8
Cinco o Seis	0.85
Tres o Cuatro	0.90
Uno o Dos	1
No informa	1

3.1.4 Variable Respuesta

La variable respuesta corresponde a si el estudiante presentó o no deserción en el periodo analizado, en caso de que no haya tenido deserción, se le asignó un valor de 0 y en caso contrario un valor de 1.

3.2 Selección de los individuos

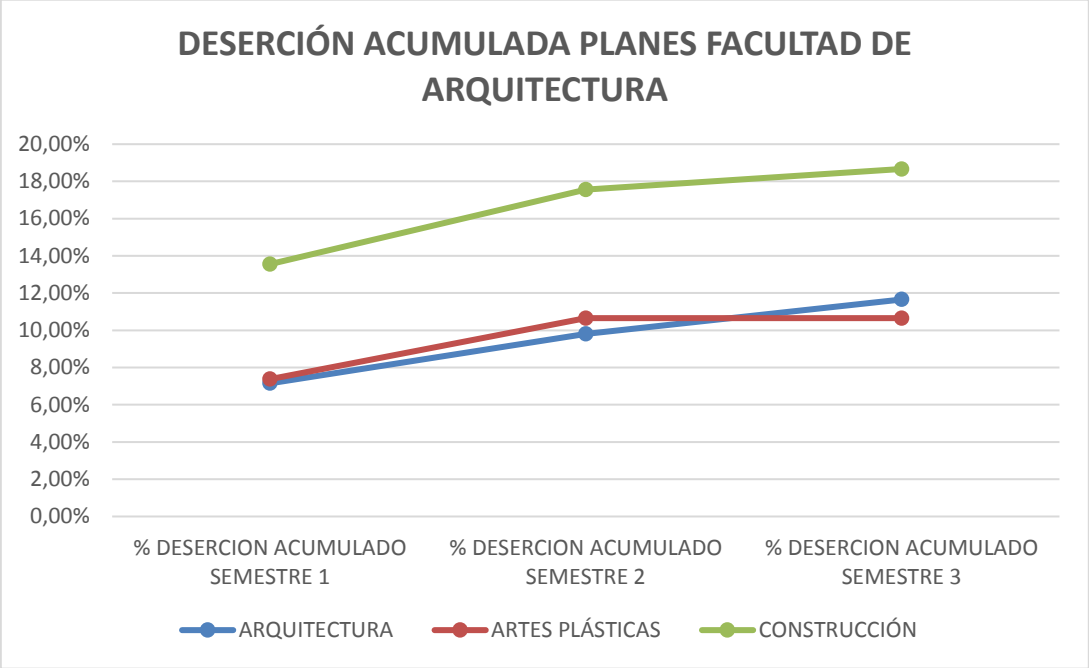
La investigación se basó en la información recolectada de 18150 estudiantes que fueron admitidos y se matricularon en la universidad nacional de Colombia Sede Medellín desde el periodo 2009-01 hasta el periodo 2016-01[6]. Se hizo la identificación de cuáles de ellos salían en cada uno de los tres periodos analizados.

Con el fin de presentar una aproximación al nivel de deserción por bajo rendimiento académico[7] que se presenta en cada uno de los programas durante en el primer, segundo y tercer semestre matriculado, se ilustra en las figuras 3-1 a 3-5 el porcentaje de deserción para cada una de las Facultades de la Universidad y en la figura 3-6 un comparativo de la deserción acumulada para los tres primeros semestres de cada Facultad y Sede.

Comparando los planes de la Facultad de Arquitectura que se ilustran en la figura 3-2, se observa que el porcentaje de deserción[8] en el programa de construcción está muy por encima de la deserción en arquitectura y artes plásticas, adicionalmente, no se presenta deserción para el programa de Artes Plásticas entre el segundo y tercer semestre, y éste junto con Arquitectura son los que presentan menor porcentaje de deserción en la Universidad para los tres semestres analizados.

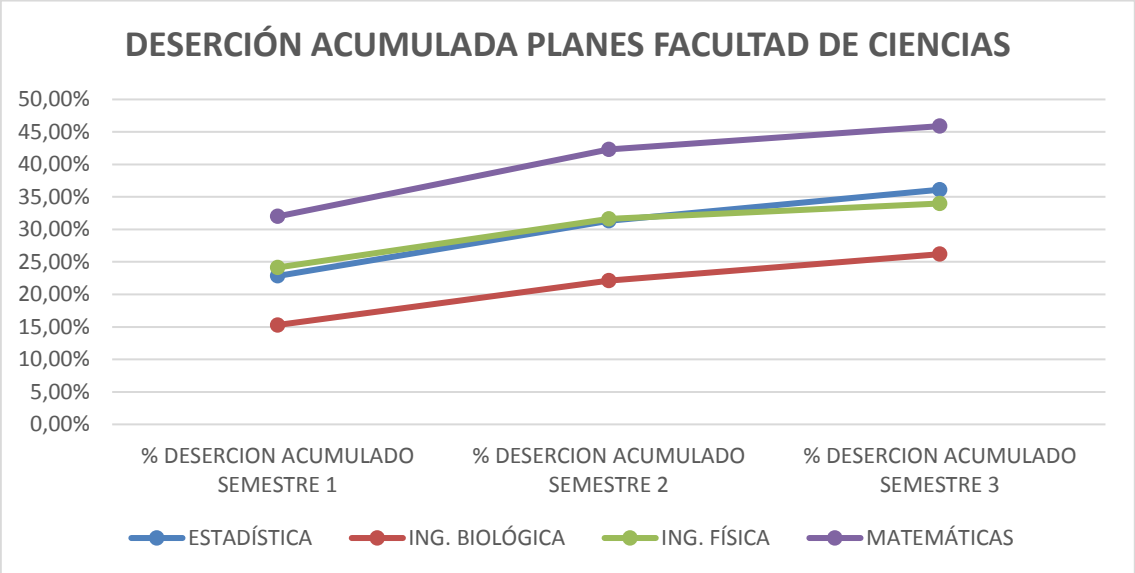
Figura 3-2 Deserción acumulada de los planes de la Facultad de Arquitectura

[6] En la ANEXO A se encuentra un resumen de los datos por plan de estudios y Facultad que serán considerados.
[7] Corresponde a los estudiantes cuyo PAPA es menor a 3.0 y los estudiantes cuyo cupo de créditos no es suficiente para cursar los créditos pendientes de aprobación
[8] Los datos presentados son en proporción al número de estudiantes de cada uno de los programas



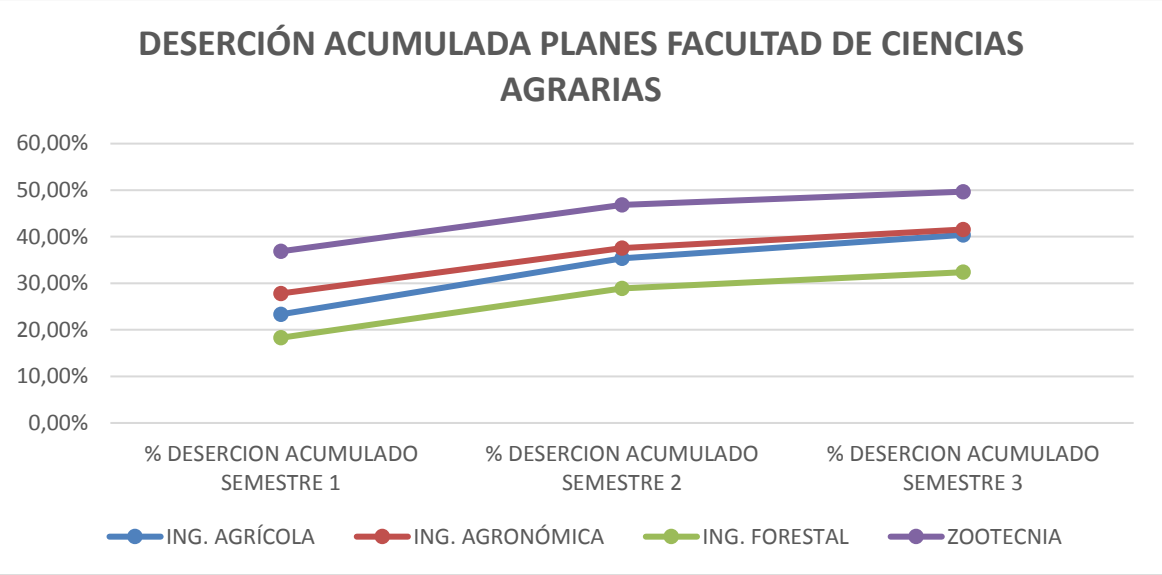
En la Figura 3-3 se presenta la deserción acumulada por semestre para los programas de la Facultad de Ciencias, allí se observa que el programa de Matemáticas es el que presenta un mayor porcentaje de deserción (45.89%). La deserción en el segundo semestre es muy similar entre los planes de Ingeniería física (31.63%) y estadística (31.33%), pero para el tercer semestre la deserción en Estadística es mayor. Ingeniería Biológica es el plan que presenta un menor porcentaje de deserción en los tres semestres analizados.

Figura 3-3 Deserción acumulada de los planes de la Facultad de Ciencias



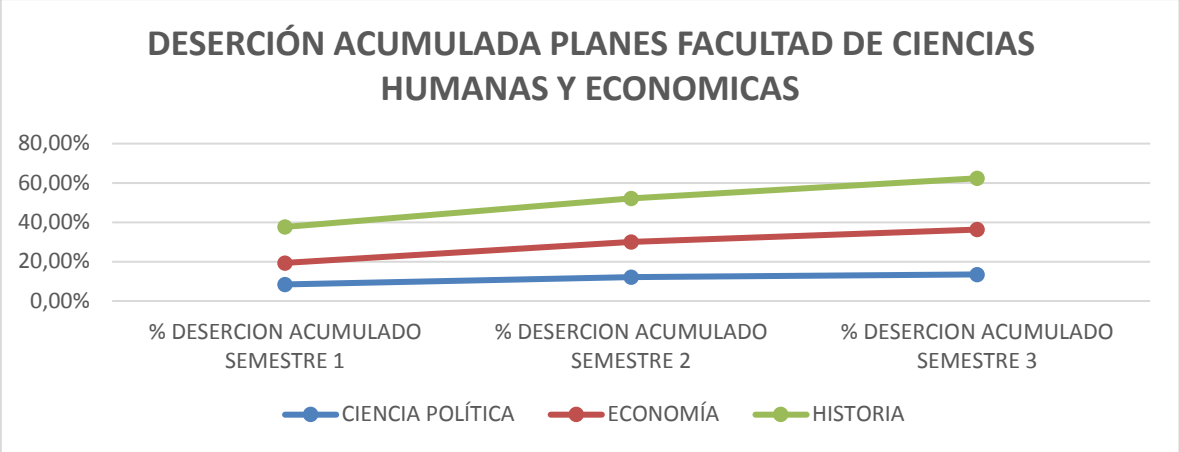
En los planes de la Facultad de Ciencias Agrarias Figura 3-4, la deserción acumulada para el programa de Zootecnia es uno de los programas que presenta mayor porcentaje de deserción en la Sede durante los tres periodos analizados. En esta Facultad, la deserción está por encima del promedio de la Sede para todos los programas, siendo Ingeniería Forestal el que presenta una menor deserción.

Figura 3-4 Deserción acumulada de los planes de la Facultad de Ciencias Agrarias



La deserción para el programa de Ciencia Política (figura 3-5), es menor respecto a los programas de Economía e Historia, siendo éste último el que presenta un mayor porcentaje de deserción.

Figura 3-5 Deserción acumulada de los planes de la Facultad de Ciencias Humanas y Económicas



Para los programas de la Facultad de Minas, se observa en la figura 3-6 que el programa de Ingeniería de Control es el que mayor deserción presenta en el semestre 1. Pero para el segundo y tercer semestre la deserción del programa de Ingeniería de Sistemas e Informática es mucho mayor que los demás programas de la Facultad. El programa que presenta una menor deserción en los tres semestres analizados es el programa de Ingeniería Civil.

Figura 3-6 Deserción acumulada de los planes de la Facultad de Minas

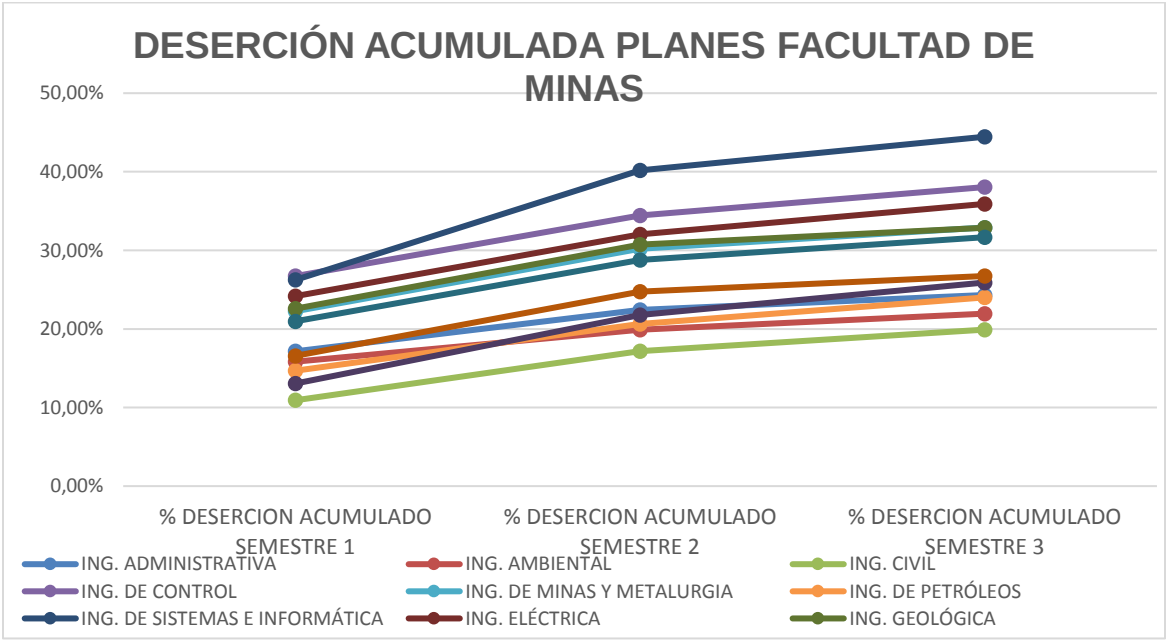
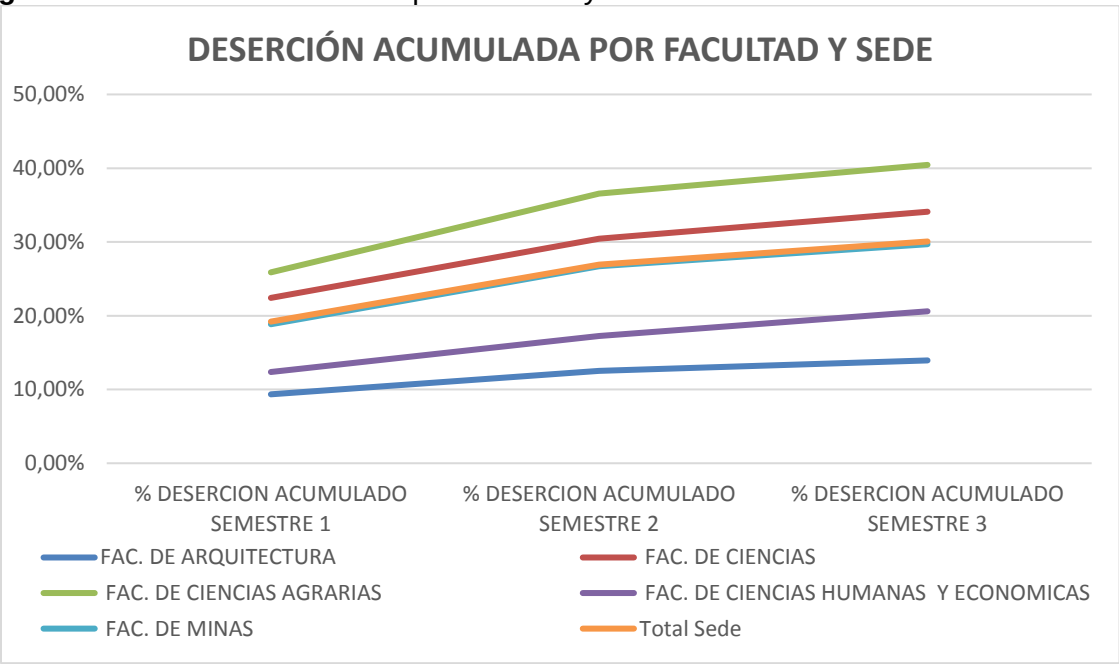


Figura 3-7 Deserción acumulada por Facultad y Sede



3.3 Análisis Exploratorio y gráfico de los datos

A continuación, se presentan algunos análisis descriptivos para las variables consideradas, los cuáles serán presentados discriminados en cada una de las Facultades.

3.3.1 Componentes del examen

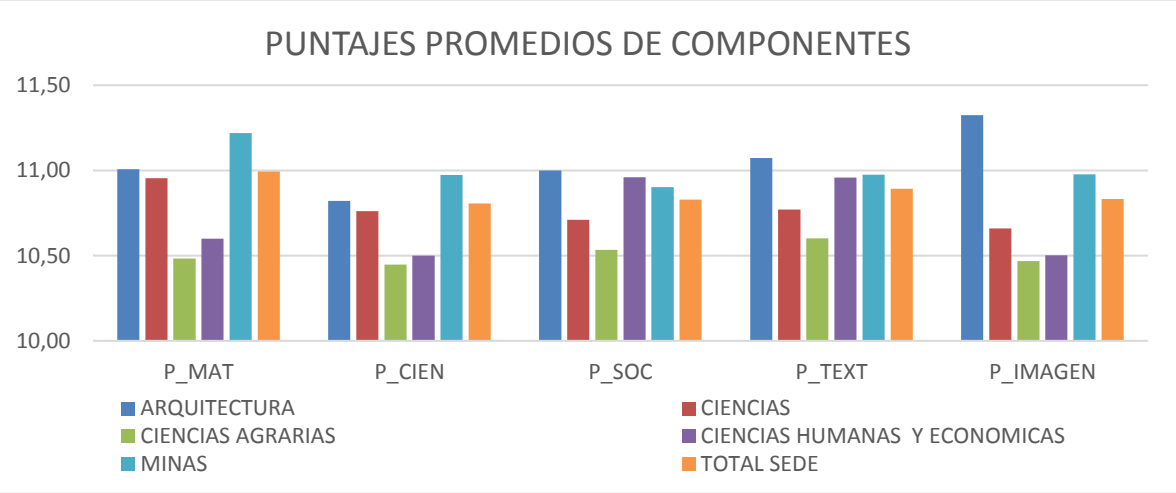
En la tabla 3-9 y en la figura 3.7 se presentan los puntajes promedios por Facultad y por Sede, en ésta se observa que en promedio los puntajes obtenidos por la Facultad de Ciencias Agrarias, son menores en todos los componentes en comparación con las demás Facultades, para las componentes de matemáticas y Ciencias, la Facultad de Minas es la que presenta mayor puntaje promedio. Para los otros tres componentes la Facultad que tiene los mejores puntajes es la Facultad de Arquitectura. Las Facultades de Minas y Arquitectura presentan puntajes en todos los componentes por encima del promedio de la Sede, para la Facultad de Ciencias Agrarias los puntajes están por debajo en todos los componentes.

Tabla 3-9 Puntaje promedio por Facultad de cada componente del examen

FACULTAD	P_MAT	P_CIEN	P_SOC	P_TEXT	P_IMAGEN
ARQUITECTURA	11.01	10.82	11.00	11.07	11.32
CIENCIAS	10.95	10.76	10.71	10.77	10.66
CIENCIAS AGRARIAS	10.48	10.45	10.53	10.60	10.47
CIENCIAS HUMANAS Y ECONOMICAS	10.60	10.50	10.96	10.96	10.50
MINAS	11.22	10.97	10.90	10.98	10.98
TOTAL SEDE	10.99	10.81	10.83	10.89	10.83

Los resultados para estas variables coinciden con los resultados obtenidos en la deserción por Facultad y por Sede (figura 3.8) en la que se evidencio que la Facultad de Ciencias Agrarias es la que presentan un mayor porcentaje de deserción.

Figura 3-8 Puntaje promedio por Facultad de cada componente del examen



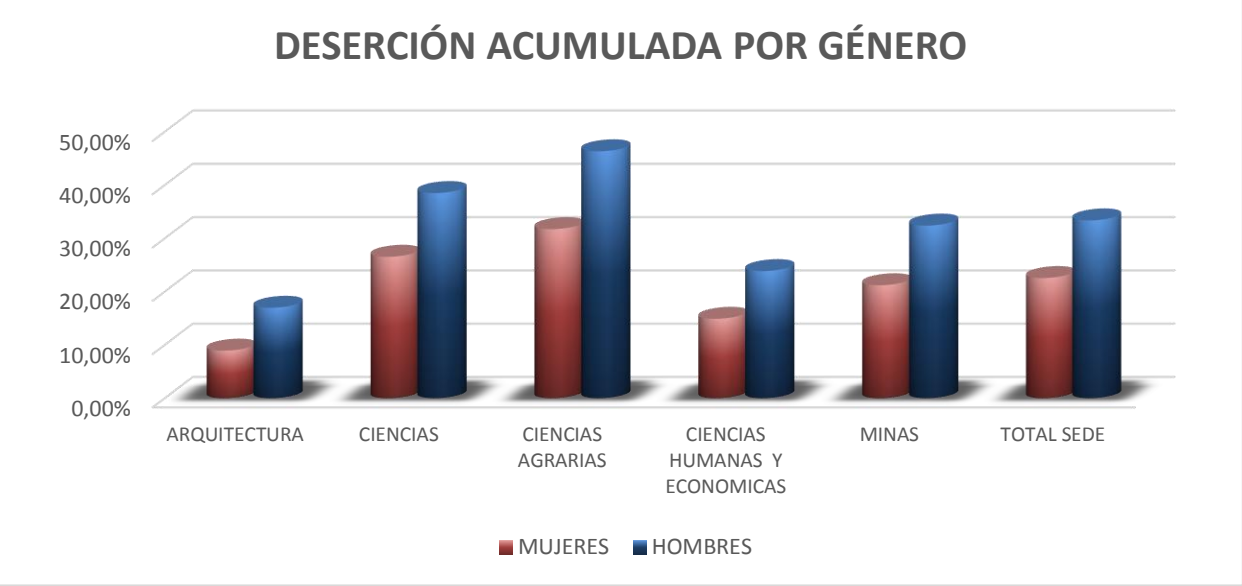
3.3.2 Deserción por Género

En la tabla 3.10. se presenta la comparación de la deserción acumulada por género para cada una de las Facultades y por sede donde se evidencia que la deserción es mayor para los hombres en todas las Facultades. En relación al total de deserción en los tres periodos para cada género (figura 3-9), la deserción acumulada para los hombres es de 33.3% en comparación a un 22.5% para las mujeres. La Facultad de Ciencias Agrarias y la Facultad de Ciencias son las que mayor deserción presentan para ambos géneros.

Tabla 3-10 Deserción por género y periodo

FACULTAD	MUJERES				HOMBRES			
	TOTAL	P1	P2	P3	TOTAL	P1	P2	P3
ARQUITECTURA	483	30	8	5	838	94	34	14
CIENCIAS	1037	168	71	36	1692	439	147	65
CIENCIAS AGRARIAS	1148	219	104	41	1669	508	196	69
CIENCIAS HUMANAS Y ECONOMICAS	625	57	19	17	999	142	60	36
MINAS	2451	303	165	51	7208	1500	593	240
TOTAL SEDE	**5744**	**777**	**367**	**150**	**12406**	**2683**	**1030**	**424**

Figura 3-9 Deserción acumulada por género para los primeros tres periodos



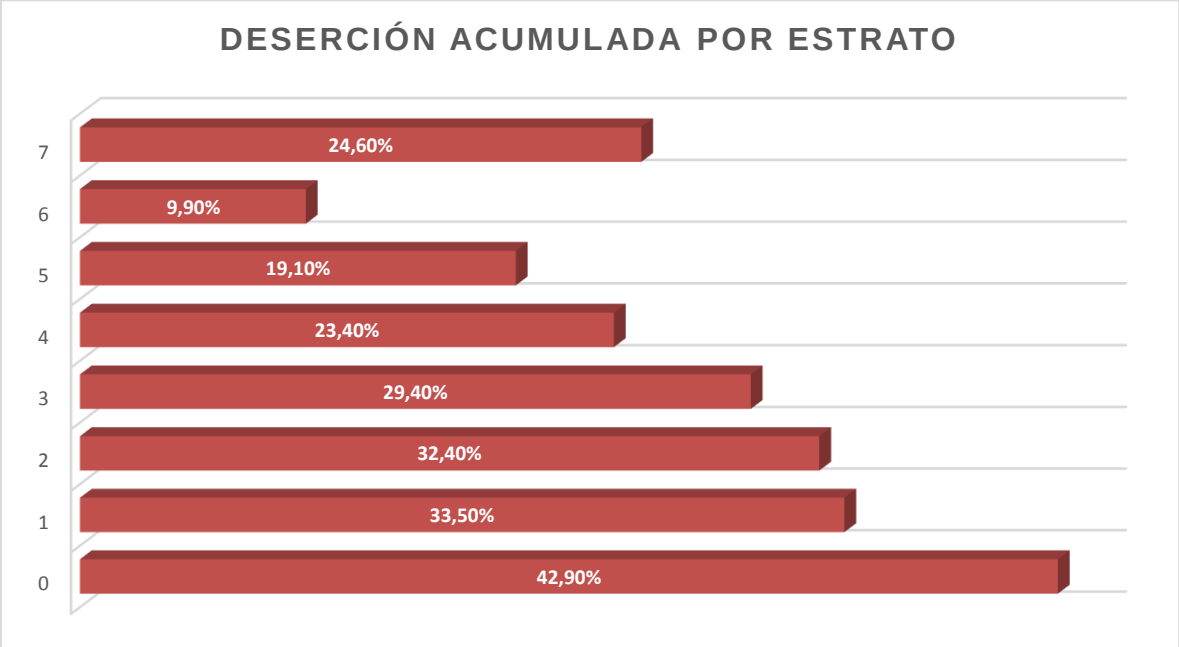
3.3.3 Deserción por estrato

Aproximadamente el 85% de los datos de estudiantes analizados, corresponden a estudiantes de estratos 1, 2 y 3. En la tabla 3-11 se presenta el porcentaje de deserción por estrato respecto al número de estudiantes del mismo estrato. En ésta se observa que el estrato 0 en la Facultad de Ciencias Humanas y económicas y el estrato 3 en la Facultad de Ciencias Agrarias son los que presentas un mayor porcentaje de deserción 100% y 41.3% respectivamente. En la Figura 3-10 se observa que el porcentaje de deserción disminuye en forma inversa al estrato, es decir a estratos altos hay un menor porcentaje de deserción.

Tabla 3-11 Porcentaje de deserción por estrato

FACULTAD	0	1	2	3	4	5	6	7	TOTAL FACULTAD
ARQUITECTURA	0.0%	17.0%	19.8%	11.4%	8.4%	14.3%	0.0%	0.0%	14.0%
CIENCIAS	60.0%	34.7%	34.8%	35.4%	30.0%	17.2%	21.4%	25.6%	33.9%
CIENCIAS AGRARIAS	30.0%	41.1%	40.2%	41.3%	38.9%	27.3%	25.0%	66.7%	40.4%
CIENCIAS HUMANAS Y ECONOMICAS	100.0%	20.9%	21.0%	21.4%	16.7%	11.1%	0.0%	16.7%	20.4%
MINAS	42.1%	34.2%	32.6%	28.5%	21.9%	19.8%	9.4%	18.1%	29.5%
TOTAL SEDE	**42.9%**	**33.5%**	**32.4%**	**29.4%**	**23.4%**	**19.1%**	**9.9%**	**24.6%**	**29.9%**

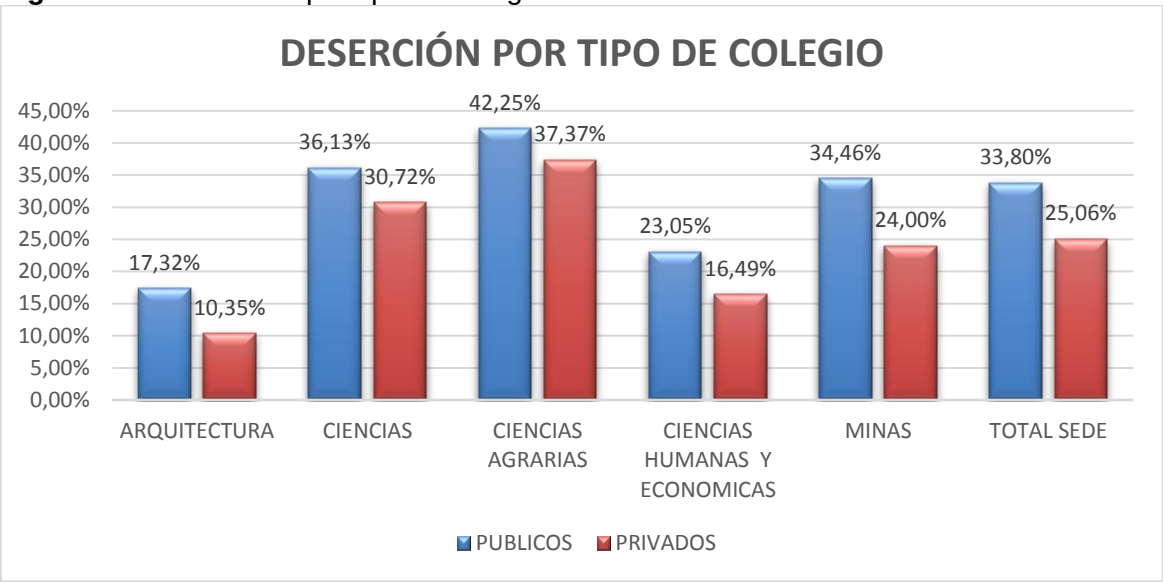
Figura 3-10 Deserción Acumulada al tercer periodo por Estrato



3.3.4 Deserción por tipo de Colegio

Para los datos analizados, el 55.7% corresponden a colegios públicos. De los estudiantes de la Sede que pertenecen a este tipo de colegios el 33.8% presentan deserción, en comparación a la presentada en los colegios privados que corresponde al 25.06%. como se ilustra en la figura 3.11. En todas las Facultades el porcentaje de deserción fue menor para los estudiantes provenientes de un colegio privado.

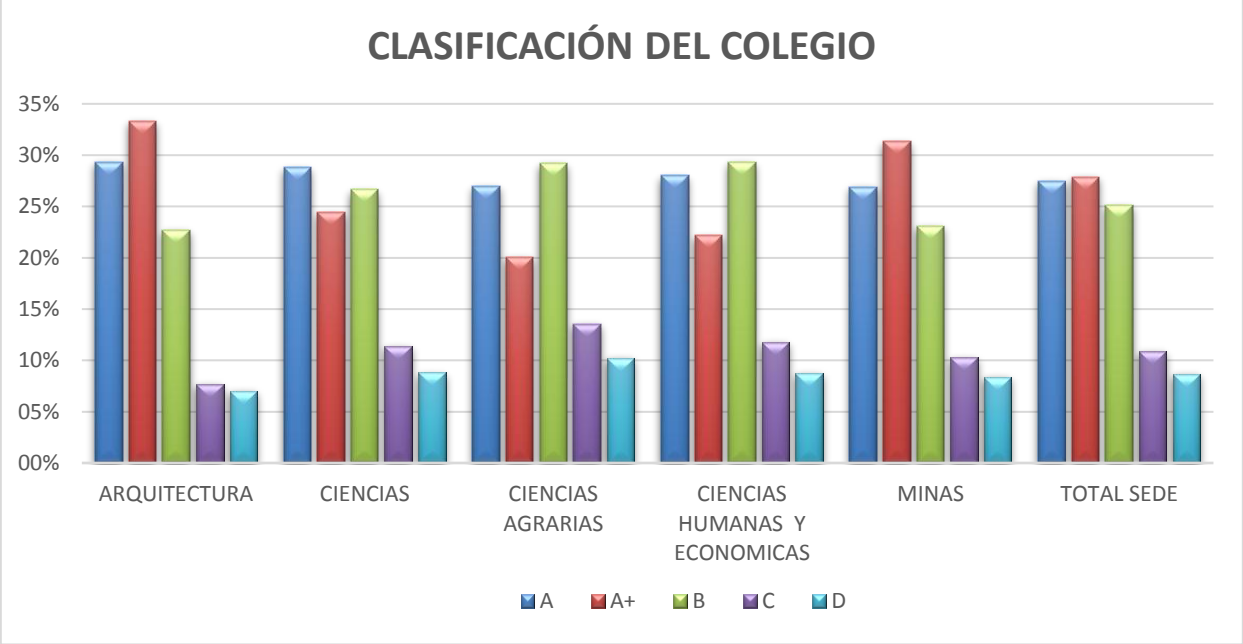
Figura 3-11 Deserción por tipo de colegio



3.3.5 Clasificación del Colegio

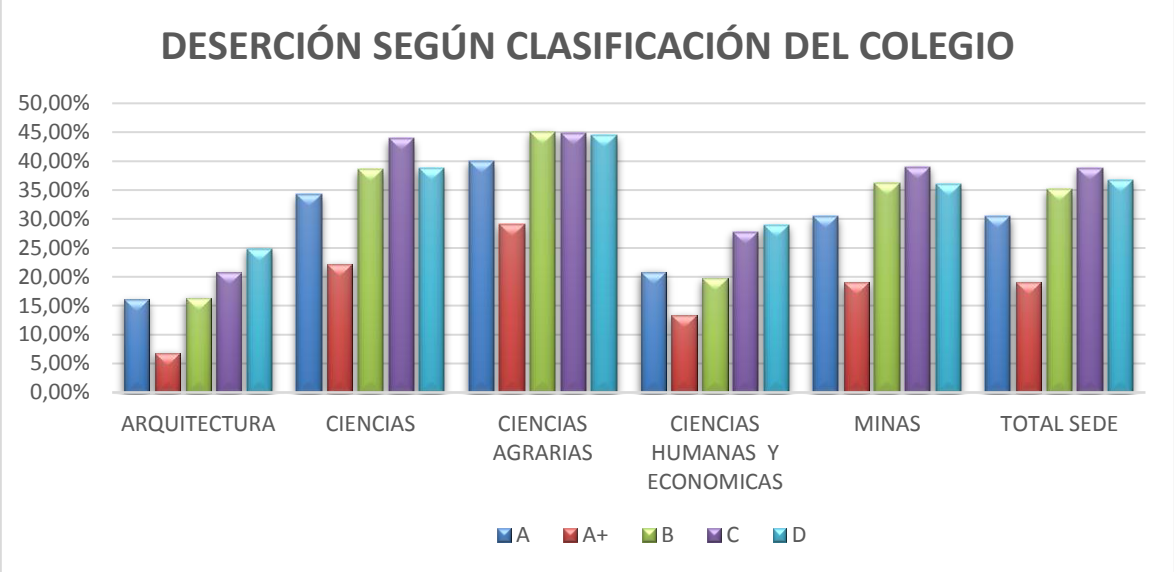
En la figura 3-12, se observa que la Facultad de Arquitectura y la Facultad de Minas son las Facultades que en proporción a su número de estudiantes tienen el mayor porcentaje de estudiantes provenientes de colegios con clasificación A+ y A , seguido por Ciencias, Ciencias Humanas y Económicas y por ultimo Ciencias Agrarias.

Figura 3-12 Distribución de la Clasificación del colegio por Facultad



La deserción para los estudiantes que se graduaron de colegios con clasificación A+ y A es menor para 4 de las 5 Facultades como se ilustra en la figura 3-13. Para la Facultad de Ciencias Humanas y Económicas, la deserción en los colegios con clasificación B es menor que la de los colegios con clasificación A.

Figura 3-13 Deserción según la clasificación del colegio por Facultad



3.3.6 Clasificación por edad

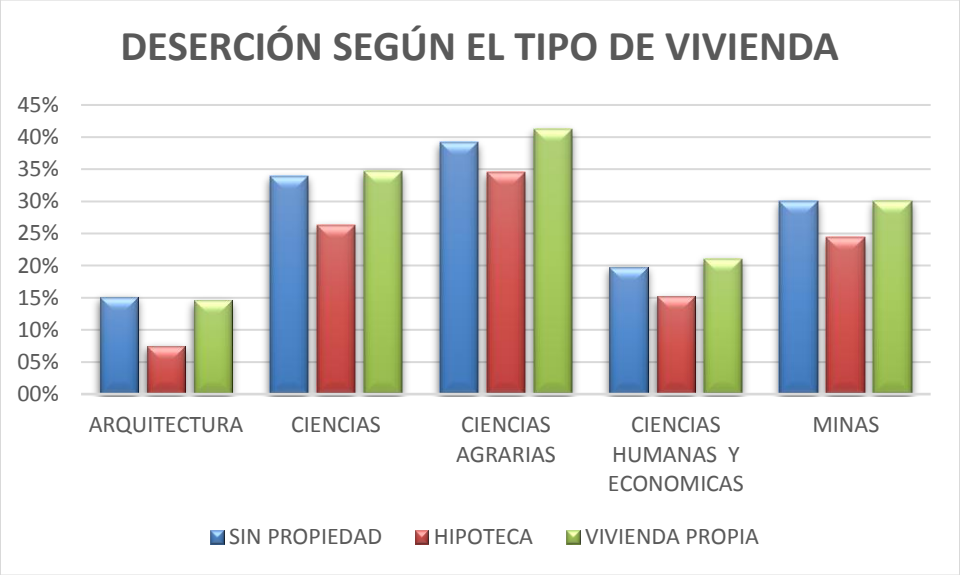
Aproximadamente el 40% de los admitidos que ingresan a la Sede tiene 17 años, el 79% ingresan con edades entre los 16 y 19 años. Pero en proporción a las personas de la misma edad, la mayor deserción se presenta con los estudiantes que ingresan con edades superiores a los 19 años. En la figura 3-14 se presenta la distribución del número de admitidos según la edad de ingreso

Figura 3-14 Admitidos por edad de ingreso.

3.3.7 Tipo de vivienda

Para la variable tipo de vivienda, se observa que el 73% de los estudiantes residen en una vivienda propia, 19% no poseen una propiedad y el 8% tienen hipoteca. En proporción al número de estudiantes que se encuentran en cada una de las tres categorías no se observa una diferencia significativa entre las facultades. En todas, la menor proporción de deserción por categoría está en los estudiantes que poseen hipoteca, como se ilustra en la figura 3-15.

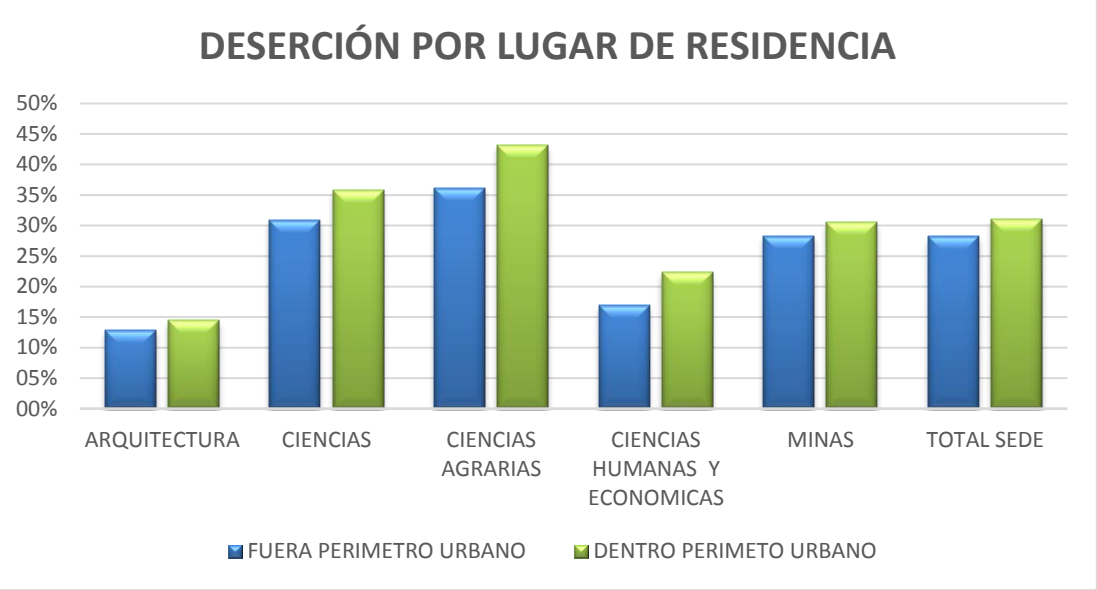
Figura 3-15 Deserción según tipo de vivienda



3.3.8 Lugar de Residencia

Aproximadamente el 60% de los admitidos provienen de lugares dentro del perímetro urbano (Área Metropolitana). De los estudiantes que presentan deserción, un 62% residen dentro perímetro urbano y un 38% en lugares fuera del perímetro urbano. En la figura 3-16 se presenta la distribución del porcentaje de deserción según el número de admitidos de cada categoría.

Figura 3-16 Porcentaje de Deserción según lugar de residencia



3.4 Medidas de tendencia central y dispersión

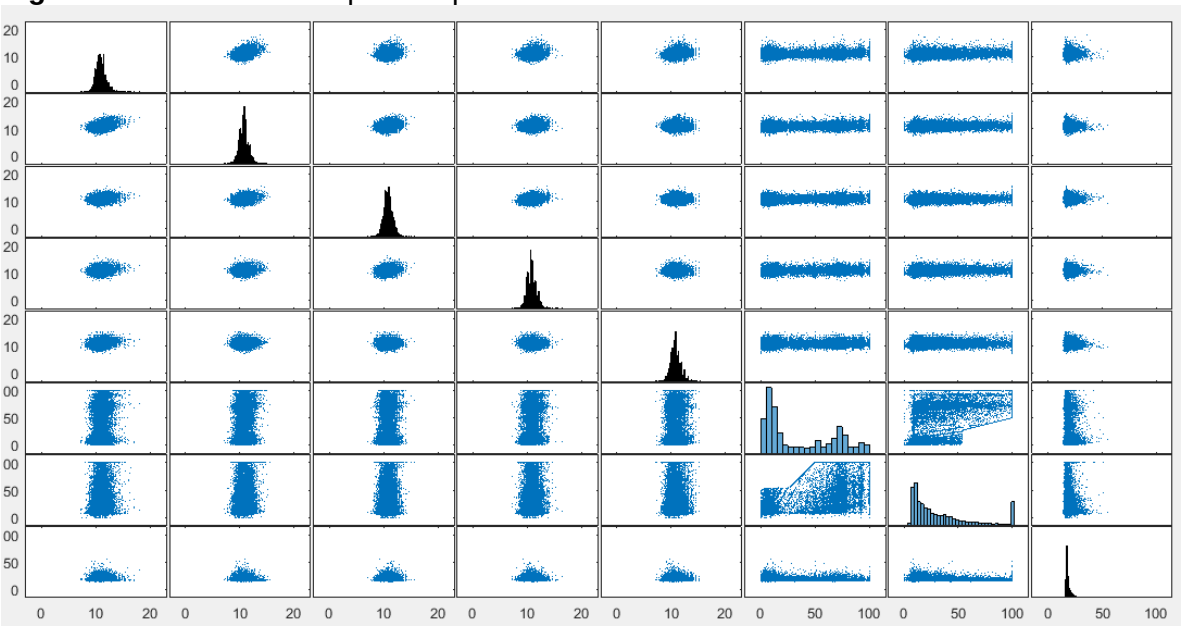
En la tabla 3-12 se resumen las medidas de centralidad y variabilidad para las variables cuantitativas, mediante los valores del coeficiente de variación se puede inferir que la que mayor dispersión la presenta el valor de la pensión (VRPE) seguido por la variable ingresos (INGR). El valor de la asimetría para las variables de los componentes del examen es muy cercano a 0, por lo que se podría decir que estadísticamente estas variables son simétricas. Para la variable edad e ingresos, se observa una asimetría positiva, por lo que se podría afirmar que hay más estudiantes con edades por debajo de la media (18.4)

Tabla 3-12. Medidas de tendencia Central y dispersión

VARIABLE	MEDIA	MEDIANA	DESV. EST	MEDA	COEF. DE VARIACION	ASIMETRIA	CURTOSIS
MATEMATICAS	10.99	10.92	0.94	0.56	0.09	0.69	5.05
CIENCIAS	10.80	10.75	0.88	0.55	0.08	0.34	3.82
SOCIALES	10.83	10.82	0.83	0.52	0.08	0.21	3.47
TEXTUAL	10.89	10.91	0.87	0.58	0.08	0.35	3.80
IMAGEN	10.83	10.76	0.94	0.60	0.09	0.47	3.72
VRPE	35.53	17.90	31.37	14.40	0.88	0.56	1.75
INGR	33.69	23.02	27.88	13.72	0.83	1.18	3.30
EDAD	18.40	17.00	2.94	1.00	0.16	2.98	17.39

En la figura 3-17 se presenta la matriz de gráficos de dispersión para las variables cuantitativas, en éste se observa que aparentemente existe independencia lineal entre las variables

Figura 3-17 Gráfico de dispersión para las variables cuantitativa



Para determinar si las variables cuantitativas son linealmente dependientes se calculó el determinante de la matriz de correlación la cual se presenta en la tabla 13, dando como resultado un valor 0.4794, con lo que se verificó que no se presenta dependencia lineal entre las variables.

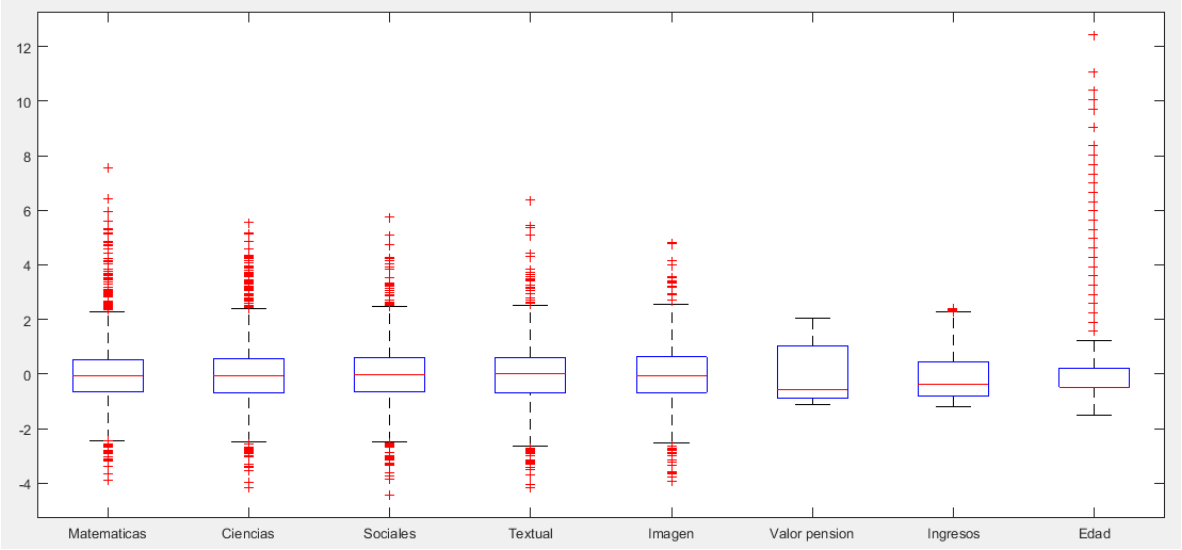
Tabla 3-13 Matriz de correlación

VARIABLE	MATEMATICAS	CIENCIAS	SOCIALES	TEXTUAL	IMAGEN	VRPE	INGR	EDAD
MATEMATICAS	1.00	0.31	0.15	0.15	0.20	0.20	0.15	0.01
CIENCIAS	0.31	1.00	0.18	0.14	0.07	0.19	0.14	-0.05
SOCIALES	0.15	0.18	1.00	0.20	0.03	0.13	0.09	0.02
TEXTUAL	0.15	0.14	0.20	1.00	0.01	0.13	0.08	-0.01
IMAGEN	0.20	0.07	0.03	0.01	1.00	0.12	0.08	-0.01
VRPE	0.20	0.19	0.13	0.13	0.12	1.00	0.55	-0.15
INGR	0.15	0.14	0.09	0.08	0.08	0.55	1.00	-0.17
EDAD	0.01	-0.05	0.02	-0.01	-0.01	-0.15	-0.17	1.00

En la figura 3.18 se presenta el diagrama de cajas para las variables cuantitativas, en éste se observa que el valor de la pensión es el que presenta mayor variabilidad, en la variable edad el valor de la mediana está muy cerca al primer cuartil, lo que quiere decir que los datos están muy concentrados en un valor, esto también se identifica en el histograma de esta variable presentado en la figura 3-16. Se presentan valores atípicos en todas las

variables, a excepción de la variable valor de la pensión. La variabilidad de los datos de los componentes del examen de admisión (matemáticas, ciencias, sociales, textual e imagen) es muy similar.

Figura 3-18 Diagrama de cajas variables cuantitativas



4 Aplicación de las técnicas multivariantes

Dada las diferencias entre las cifras de deserción de las Facultades, analizadas en el capítulo anterior y en particular para las cifras de la Facultad de Ciencias Agrarias y Arquitectura que son las que presentan un mayor y menor porcentaje de deserción respectivamente[9], se identificó la necesidad de generar modelos con las técnicas multivariantes desagregados por Facultad, lo que implicó pasar de 4 modelos que eran los que se había proyectado inicialmente a hacer la identificación de 20 modelos[10], con el fin de buscar disminuir el error y que los grupos de datos fueran más homogéneos. Adicionalmente, la exigencia de las asignaturas que cursan los estudiantes en sus primeros semestres es muy similar entre los programas que hacen parte de una misma Facultad, pero pueden ser diferentes de una Facultad a otra.

La aplicación de cada una de las técnicas multivariantes se hizo mediante la utilización del software Matlab. A través de este software se hizo la identificación de los datos atípicos aplicando el procedimiento descrito en el capítulo 2.3. los cuáles fueron retirados de los datos a utilizar, luego se hizo la estandarización de las variables cuantitativas y con los datos estandarizados, se hizo el entrenamiento clasificado los individuos en dos categorías. Para la construcción del modelo se seleccionaron grupos de datos de entrenamiento y validación aleatorios y se calculó para cada uno de ellos la matriz de confusión y el valor de especificidad y Sensibilidad, para luego seleccionar el modelo que obtuviera un mayor valor de sensibilidad. Los modelos entrenados en Matlab permiten identificar si una nueva observación pertenece a una de las dos categorías.

4.1 Máquina Vector Soporte

Para la aplicación de la técnica máquina Vector Soporte se utilizaron el 25% de los datos para validación y el 75% como entrenamiento, se realizó la aplicación de la técnica utilizando 4 tipos de Kernel: Lineal, cuadrático, cúbico y Gaussiano. Inicialmente se hizo la separación de los datos en cada una de las Facultades y se entrenó el modelo con cada uno de los Kernel, se hizo una exploración inicial para determinar cuál Kernel utilizar,

[9] La proporción de deserción se da en términos del número de estudiantes que posee la Facultad y que fueron objeto de análisis.
[10] Se realizó el entrenamiento de los modelos con las 4 técnicas multivariantes para las 5 Facultades

mediante el entrenamiento de 50 grupos con datos de entrenamiento y validación aleatorios y con diferentes parámetros para cada Kernel, con los resultados obtenidos se determinó cuáles eran los dos mejores y con éstos se generaron los escenarios, para buscar el Kernel con el cual se obtuviera un mayor valor de sensibilidad y menor error.

4.1.1 Estimación del modelo para predecir la deserción en el primer semestre

En la tabla 4-1 se resumen los resultados obtenidos para los grupos de datos que se entrenaron con cada uno de los Kernel, para la Facultad de Arquitectura, los Kernel que presentaron un mayor valor de sensibilidad fueron el cúbico y el Gaussiano, siendo éste último el que presentó menor tasa de error.

A partir de los resultados obtenidos con los Kernel utilizados, para las demás Facultades solo se hizo la validación con los Kernel cúbico y gaussiano, ya que fueron los que mayor sensibilidad presentaron y con éstos se realizó el entrenamiento de los grupos de datos aleatorios, seleccionando el grupo que presentó un mayor porcentaje de sensibilidad y menor tasa de error.

Tabla 4-1 Resultados para SVM Facultad de Arquitectura

KERNEL	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSIBI-LIDAD	TASA DE ERROR
LINEAL	50	1	85,45%	[275 24;24 7]	91,97%	22,58%	14,55%
CUADRATICO	50	6	87,27%	[280 19;23 8]	93,65%	25,81%	12,73%
CÚBICO	1200	425	83,64%	[263 36;18 13]	87,96%	41,94%	16,36%
GAUSSIANO	1230	622	85,45%	[270 29;19 12]	90,30%	38,71%	14,55%

Para la Facultad de Ciencias, el Kernel que presentó una menor tasa de error y un mayor valor de Sensibilidad fue el Gaussiano, como se ilustra en la tabla 4-2.

Tabla 4-2 Resultados para SVM Facultad de Ciencias

KERNEL	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSI-BILIDAD	TASA DE ERROR
CÚBICO	625	48	65,40%	[403 128;108 43]	75,89%	28,48%	34,60%
GAUSSIANO	1000	479	73,90%	[443 87;91 61]	83,58%	40,13%	26,10%

Para la Facultad de Ciencias Agrarias, el Kernel que mejor predijo la posibilidad de deserción para el primer semestre fue el cúbico con una sensibilidad del 38.46% en comparación del 34.81% del modelo entrenado con el Kernel Gaussiano, pero el cúbico posee una mayor tasa de error y dado que la diferencia en la tasa de error entre los dos Kernel es significativa, se eligió como mejor Kernel el Gaussiano.

Tabla 4-3 Resultados para SVM Facultad de Ciencias Agrarias

KERNEL	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSI-BILIDAD	TASA DE ERROR
CÚBICO	400	4	63,49%	[377 145;112 70]	72,22%	38,46%	36,51%
GAUSSIANO	800	499	68,75%	[421 102;118 63]	80,50%	34,81%	31,25%

Para la Facultad de Ciencias Humanas y Económicas, el Kernel que mejor predice la posibilidad de deserción para el primer semestre es el Cúbico con una sensibilidad del 38.78% y un error del 22.26%, pero al igual que para la Facultad de Ciencias Agrarias, el Cúbico presenta una mayor tasa de error.

Tabla 4-4 Resultados para SVM Facultad de Ciencias Humanas y Económicas

KERNEL	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSI-BILIDAD	TASA DE ERROR
CÚBICO	400	196	77,34%	[295 62;30 19]	82,63%	38,78%	22,66%
GAUSSIANO	800	497	83,74%	[321 35;31 19]	90,17%	38,00%	16,26%

Dada la cantidad de datos que se tienen para la Facultad de Minas solo se verificó con el modelo Gaussiano, en vista de que este fue el Kernel que tuvo una mejor relación entre los valores de sensibilidad y tasa de error, con este Kernel la tasa de Sensibilidad para la facultad de Minas fue del 25.78%

Tabla 4-5 Resultados de SVM para la Facultad de Minas

KERNEL	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSIBI-LIDAD	TASA DE ERROR
GAUSSIANO	200	57	75.43%	[1705 259;334 116]	86.81%	25.78%	24.57%

4.1.2 Estimación modelo para deserción en el segundo semestre

Para la estimación de la probabilidad de deserción en el segundo semestre, se tomó la información correspondiente a estudiantes que no salieron en el primer semestre y se adicionó la información de las variables eficiencia, PAPA y número de asignaturas perdidas. Antes de hacer el entrenamiento de los modelos, se siguió el mismo procedimiento que en el primer semestre, es decir, se retiraron los datos atípicos y se hizo la estandarización de las variables con los datos restantes.

En los resultados obtenidos con los datos del primer semestre, se pudo verificar que el Kernel Gaussiano fue el que tuvo un mejor desempeño (tomando en cuenta los valores de sensibilidad y tasa de error) es por esto que para el segundo y tercer semestre se utilizó este Kernel para realizar el entrenamiento del modelo de SVM.

En la tabla 4-6 se presentan los resultados obtenidos para el grupo de datos seleccionado de cada Facultad que obtuvieron el mayor resultado de Sensibilidad, con una menor tasa de error.

Tabla 4-6 Resultados de SVM para el segundo semestre Kernel Gaussiano

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSIBILIDAD	TASA DE ERROR
ARQUITECTURA	800	752	98,14%	[260 2;3 4]	99,24%	57,14%	1,86%
CIENCIAS	800	293	84,58%	[353 41;27 20]	89,59%	42,55%	15,42%
CIENCIAS AGRARIAS	800	379	82,73%	[360 45;36 28]	88,89%	43,75%	17,27%
CIENCIAS HUMANAS Y ECONOMICAS	800	506	92,90%	[279 14;8 9]	95,22%	52,94%	7,10%
MINAS	230	19	85,36%	[1425 163;92 62]	89,74%	40,26%	14,64%

Con los resultados obtenidos se observa que con las variables que se incluyeron, hay una mejora considerable en los valores de sensibilidad de cada de los modelos de SVM utilizados para cada Facultad, en comparación a los resultados obtenidos con los datos del primer periodo.

4.1.3 Estimación modelo para deserción en el tercer semestre

Para la estimación de la deserción en el tercer semestre, se tomaron las mismas variables que el segundo semestre y se siguió el mismo procedimiento, haciendo el cálculo para cada Facultad, en la tabla 4-7 se encuentran los resultados obtenidos. Para este semestre los datos correspondientes a estudiantes que presentaron deserción ya han disminuido, y por esto los valores de la matriz de confusión son bajos y la mayoría de los datos corresponden a estudiantes que no presentan deserción.

Tabla 4-7 Resultados de SVM para el tercer semestre Kernel Gaussiano

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSIBI-LIDAD	TASA DE ERROR
ARQUITECTURA	800	669	99,60%	[248 0;1 2]	100,00%	66,67%	0,40%
CIENCIAS	800	577	93,21%	[336 15;10 7]	95,73%	41,18%	6,79%
CIENCIAS AGRARIAS	800	797	91,28%	[326 20;12 9]	94,22%	42,86%	8,72%
CIENCIAS HUMANAS Y ECONOMICAS	800	140	94,37%	[263 11;5 5]	95,99%	50,00%	5,63%
MINAS	200	12	93,93%	[1435 56;38 19]	96,24%	33,33%	6,07%

4.2 KNN

Al igual que para SVM, para KNN se dejó el 25% de los datos para validación y el restante para entrenamiento. Para cada Facultad y semestre se entrenaron 1000 grupos de datos de entrenamiento y validación aleatorios. Se realizó una exploración inicial a partir de diferentes medidas de distancia, identificando cual es la que mejor resultados de clasificación obtenía.

En la tabla 4-8 se encuentra el resumen de los resultados obtenidos para cada una de las Facultades, con el grupo de datos que obtuvo un mayor porcentaje de sensibilidad y especificidad.

Tabla 4-8 Resultados para KNN primer semestre

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSIBILIDAD	TASA DE ERROR
ARQUITECTURA	1000	86	88,89%	[261 17;17 11]	93,88%	39,29%	11,11%
CIENCIAS	1000	361	73,17%	[388 87;78 62]	81,68%	44,29%	26,83%
CIENCIAS AGRARIAS	1000	422	66,16%	[363 123;98 69]	74,69%	41,32%	33,84%
CIENCIAS HUMANAS Y ECONOMICAS	1000	260	84,88%	[304 30;27 16]	91,02%	37,21%	15,12%
MINAS	1000	540	73,70%	[1511 293;288 117]	83,76%	28,89%	26,30%

Para el segundo semestre la predicción de los estudiantes con deserción disminuyo en 3 de las 5 Facultades (tabla 4-9), respecto a los resultados obtenidos en el primer semestre, pero hubo una disminución considerable en la tasa de error, lo que implica que la predicción del estudiante que no presentarían deserción aumento.

Tabla 4-9 Resultados KNN segundo semestre

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECI-FICIDAD	SENSIBILIDAD	TASA DE ERROR
ARQUITECTURA	1000	110	96,28%	[256 5;5 3]	98,08%	37,50%	3,72%
CIENCIAS	1000	817	84,13%	[351 43;27 20]	89,09%	42,55%	15,87%
CIENCIAS AGRARIAS	1000	516	82,52%	[361 44;38 26]	89,14%	40,63%	17,48%
CIENCIAS HUMANAS Y ECONOMICAS	1000	811	92,26%	[279 13;11 7]	95,55%	38,89%	7,74%
MINAS	1000	810	86,97%	[1460 128;99 55]	91,94%	35,71%	13,03%

En la tabla 4-10 se presentan los resultados obtenidos para los datos del tercer semestre, los valores de Sensibilidad respecto a los del segundo semestre aumentaron para las Facultades de Arquitectura y Ciencias Humanas y Económicas, pero disminuyeron en las demás facultades.

Tabla 4-10 Resultados KNN tercer semestre

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECIFI-CIDAD	SENSIBILIDAD	TASA DE ERROR
ARQUITECTURA	1000	29	98,01%	[244 4;1 2]	98,39%	66,67%	1,99%
CIENCIAS	1000	800	94,29%	[341 10;11 6]	97,15%	35,29%	5,71%
CIENCIAS AGRARIAS	1000	529	92,10%	[331 15;14 7]	95,66%	33,33%	7,90%
CIENCIAS HUMANAS Y ECONOMICAS	1000	311	94,37%	[264 10;6 4]	96,35%	40,00%	5,63%
MINAS	1000	799	93,67%	[1437 54;44 13]	96,38%	22,81%	6,33%

4.3 Análisis Discriminante

Para el entrenamiento de los modelos de AD el Matlab dispone de dos metodologías, el AD lineal y AD cuadrático, siendo este último con el que se obtuvo los mayores resultados de sensibilidad en la clasificación de los grupos de datos y el que se utilizó en en el entrenamiento en los tres semestres

Los resultados del primer semestre se encuentran en la tabla 4-11, allí se observa que para la Facultad de Ciencias, obtuvo el valor de sensibilidad más alto, pero de igual forma con la tasa de error mayor.

Tabla 4-11 Resultados AD primer semestre

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECIFICIDAD	SENSIBILIDAD	TASA DE ERROR
ARQUITECTURA	1000	156	78,76%	[229 48;17 12]	82,67%	41,38%	21,24%
CIENCIAS	1000	418	68,46%	[358 117;77 63]	75,37%	45,00%	31,54%
CIENCIAS AGRARIAS	1000	20	73,35%	[430 57;117 49]	88,30%	29,52%	26,65%
CIENCIAS HUMANAS Y ECONOMICAS	1000	277	80,90%	[291 43;29 14]	87,13%	32,56%	19,10%
MINAS	1000	424	74,74%	[1511 293;265 140]	83,76%	34,57%	25,26%

Para el Segundo semestre (tabla 4-12), El AD no es adecuado para la predicción de deserción con los datos de la Facultad de Arquitectura y la Facultad de Ciencias Humanas y Económicas, en lo que obtuvo una Sensibilidad de 0%, y una especificidad del 100%, es decir el algoritmo clasifico a todos los grupos de validación sin deserción. Par la demás Facultades obtuvo unos valores de sensibilidad relativamente altos.

Tabla 4-12 Resultados AD segundo semestre

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECIFICIDAD	SENSIBILIDAD	TASA DE ERROR
ARQUITECTURA	1000	1	97,40%	[262 0;7 0]	100,00%	0,00%	2,60%
CIENCIAS	1000	456	80,50%	[322 72;14 33]	81,73%	70,21%	19,50%
CIENCIAS AGRARIAS	1000	236	72,07%	[295 110;21 43]	72,84%	67,19%	27,93%
CIENCIAS HUMANAS Y ECONOMICAS	1000	1	94,52%	[293 0;17 0]	100,00%	0,00%	5,48%
MINAS	1000	300	83,93%	[1365 223;50 104]	85,96%	67,53%	15,67%

Para el tercer semestre, los valores de sensibilidad para las Facultades de Minas y Ciencias, continuaron siendo adecuados, pero hay una disminución en la capacidad de predicción para los datos de la Facultad de Ciencias Agrarias en el cual el valor de sensibilidad paso de 67.19% en el segundo semestre a un valor de 23.81%. Para las Facultades de Arquitectura y Ciencias Humanas y Económicas, al igual que el segundo semestre, obtuvo una sensibilidad del 0%,

Tabla 4-13 Resultados AD tercer semestre

FACULTAD	GRUPOS ENTRENADOS	GRUPO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECIFICIDAD	SENSIBILIDAD	TASA DE ERROR
ARQUITECTURA	1000	1	98,80%	[248 0;3 0]	100,00%	0,00%	1,20%
CIENCIAS	1000	777	90,49%	[322 29;6 11]	91,74%	64,71%	9,51%
CIENCIAS AGRARIAS	1000	552	94,82%	[343 3;16 5]	99,13%	23,81%	5,18%
CIENCIAS HUMANAS Y ECONOMICAS	1000	1	96,48%	[274 0;10 0]	100,00%	0,00%	3,52%
MINAS	1000	695	85,40%	[1284 207;19 38]	86,12%	66,67%	14,60%

4.4 Regresión Logística

La regresión logística requiere que las variables sean cuantitativas, por lo que fue necesario crear nuevas variables dummy para las variables categóricas, previo a la aplicación del modelo. En el anexo C, se encuentra la descripción y los valores de estas nuevas variables.

Una vez reemplazadas las variables categóricas con las variables Dummy se realizó el proceso de regresión logística, estimando los coeficientes de cada variable y el p-value correspondiente. Se revisaron las variables cuyo P-value fuera mayor a 0.05 y de estas se retiró la que tuviera el mayor p-value[11] y luego se estimaron nuevamente los coeficientes con las demás variables, repitiendo este procedimiento hasta que todas las variables tuvieran un P-value menor a 0.05.

Para cada una de las Facultades se estimó un modelo y se generaron 1000 corridas con datos aleatorios para la validación. Los coeficientes estimados para cada una de las Facultades se presentan en el anexo D y las variables que fueron significativas en los modelos de cada una se presentan en el Anexo E.

En la tabla 4-14, se presentan los resultados obtenidos a partir de los modelos que se identificaron para la RL, tomando en cuenta las variables que fueron significativas con los datos de cada una de las Facultades. En ésta se observa que la RL para la predicción de la deserción en la Facultad de Arquitectura, no es adecuada, ya que, si bien la tasa de error es la más baja, el valor de sensibilidad es 0%, lo que quiere decir que no hizo ninguna

[11] El p-value verifica la hipótesis nula de que el coeficiente es igual a cero, por lo que un p-value menor que 0.05 indica que puedes rechazar la hipótesis nula.

predicción respecto a los estudiantes que tienen posibilidad de deserción. Para las demás Facultades, también se presenta un valor bajo de sensibilidad

Tabla 4-14 Resultados de RL Primer semestre

FACULTAD	MODELOS ENTRENADOS	MODELO SELECCIONADO	ACIER-TOS	MATRIZ DE CONFUSION	ESPECIFI-CIDAD	SENSIBI LIDAD	TASA DE ERROR
ARQUITECTURA	1000	1	90,61%	[299 0;31 0]	100%	0%	9%
CIENCIAS	1000	371	77,27%	[514 17;138 13]	96,80%	8,61%	22,73%
CIENCIAS AGRARIAS	1000	546	74,86%	[507 15;162 20]	97,13%	10,99%	25,14%
CIENCIAS HUMANAS Y ECONOMICAS	100	9	88,18%	[356 0;48 2]	100,00%	4,00%	11,82%
MINAS	1000	322	79,66%	[1866 98;393 57]	95,01%	12,67%	20,34%

En la tabla 4-15 se presentan los resultados obtenidos para los modelos seleccionados aplicando la técnica de RL para los datos del segundo semestre, en éste se observa un aumento en la sensibilidad para todas las Facultades, pero el valor continúa siendo muy bajo.

Tabla 4-15 Resultados de RL segundo semestre

FACULTAD	MODELOS ENTRENADOS	MODELO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECIFI-CIDAD	SENSI-BILIDAD	TASA DE ERROR
ARQUITECTURA	1000	1	97,24%	[282 0;8 0]	100,00%	0,00%	2,76%
CIENCIAS	1000	377	89,67%	[426 8;42 8]	98,16%	16,00%	10,33%
CIENCIAS AGRARIAS	1000	187	84,57%	[397 32;45 25]	92,54%	35,71%	15,43%
CIENCIAS HUMANAS Y ECONOMICAS	1000	15	94,79%	[308 0;17 1]	1	5,56%	5,21%
MINAS	1000	349	91,25%	[1717 15;152 24]	99,13%	13,64%	8,75%

Para los modelos de Regresión logística con los datos del tercer semestre (tabla 4-16), se observa que también tuvo un aumento en el valor de la sensibilidad respecto a los datos del primer t segundo semestre.

Tabla 4-16 Resultados de RL tercer semestre

FACULTAD	MODELOS ENTRENADOS	MODELO SELECCIONADO	ACIERTOS	MATRIZ DE CONFUSION	ESPECIFI-CIDAD	SENSI-BILIDAD	TASA DE ERROR
ARQUITECTURA	1000	887	98,58%	[275 2;2 2]	99,28%	50,00%	1,42%
CIENCIAS	1000	84	95,61%	[408 1;18 6]	99,76%	25,00%	4,39%
CIENCIAS AGRARIAS	1000	76	94,17%	[402 1;24 2]	99,75%	7,69%	5,83%
CIENCIAS HUMANAS Y ECONOMICAS	1000	604	95,78%	[292 4;9 3]	98,65%	25,00%	4,22%
MINAS	1000	3	96,01%	[1662 0;69 0]	100,00%	0,00%	3,99%

4.5 Resumen de resultados y elección de modelos

A continuación, se presenta el resumen de los resultados obtenidos en los tres semestres, con este resumen, se realizó el análisis comparando los valores de sensibilidad, especificidad y tasa de error de cada modelo entrenado con las técnicas utilizadas y se definió a partir de estos valores cual es el mejor modelo para los datos de cada una de las Facultades.

4.5.1 Primer Semestre

En la tabla 4-17 se encuentra el resumen de los resultados obtenidos en el primer semestre para las medidas de sensibilidad, especificidad y tasa de error de cada uno de las técnicas utilizadas en las 5 facultades.

Tabla 4-17 Resumen de resultados de los modelos para el primer semestre

FACULTAD		ARQUITECTURA	CIENCIAS	CIENCIAS AGRARIAS	CIENCIAS HUMANAS Y ECONOMICAS	MINAS
SVM	Especificidad	90,30%	83,58%	80,50%	90,17%	86.81%
	Sensibilidad	38,71%	40,13%	34,81%	38,00%	25.78%
	Tasa de Error	14,55%	26,10%	31,25%	16,26%	24.57%
KNN	Especificidad	93,88%	81,68%	74,69%	91,02%	83,76%
	Sensibilidad	39,29%	44,29%	41,32%	37,21%	28,89%
	Tasa de Error	11,11%	26,83%	33,84%	15,12%	26,30%
AD	Especificidad	82,67%	75,37%	88,30%	87,13%	83,76%
	Sensibilidad	41,38%	45,00%	29,52%	32,56%	34,57%
	Tasa de Error	21,24%	31,54%	26,65%	19,10%	25,26%
RL	Especificidad	100%	96,80%	97,13%	100,00%	95,01%
	Sensibilidad	0%	8,61%	10,99%	4,00%	12,67%
	Tasa de Error	9%	22,73%	25,14%	11,82%	20,34%

Se observa que, para la Facultad de Arquitectura, los modelos que mejor predicen la posibilidad de deserción son el AD y el KNN con sensibilidades de 41.38% y 39.29% respectivamente. pero el KNN tiene un valor menor de tasa de error, por lo que se selecciona como mejor modelo el KNN. KNN

Para la Facultad de Ciencias, al igual que en la Facultad de Arquitectura los dos modelos que tienen un mayor valor de sensibilidad son el AD y el KNN con valores de 45% y 44.29% respectivamente, pero la tasa de error en KNN (26.83%) es menor que en AD (31.54%), por lo que se selecciona el modelo KNN.

Para la Facultad de Ciencias Agrarias, los modelos que tienen un mayor valor de sensibilidad son el KNN y el SVM con 41.32% y 34.81% respectivamente. El SVM tiene una tasa de error de 31.25% que es menor a la del KNN que tiene una tasa de error de 33.84%, pero dada la diferencia en la Sensibilidad, se escoge como mejor modelo el KNN

Para la Facultad de Ciencias Humanas y Económicas los modelos que presentaron un mayor valor de sensibilidad fueron el SVM (38%) y el KNN (37.21%), pero el KNN tiene la menor tasa de error (15.12%), por lo que el modelo escogido es el entrenado con KNN.

Para la Facultad de Minas, los modelos que obtuvieron un mayor valor de Sensibilidad fueron el KNN y el AD con valores de 28.89% y 34.57% respectivamente, la tasa de error AD es de 25.26%, mientras que para KNN es de 26.30%, por lo que se selecciona el modelo entrenado con AD como el mejor modelo para predecir la deserción en el primer semestre en la Facultad de Minas

Para todas las Facultades, el modelo entrenado con Regresión Logística fue el que obtuvo los resultados más bajos de sensibilidad.

4.5.2 Segundo Semestre

En la tabla 4.18 se encuentran los resultados obtenidos para los modelos que se entrenaron con cada una de las técnicas multivariantes, en éste al igual que en el primer semestre, la sensibilidad para el modelo de Regresión Logística obtuvo resultados muy bajos de Sensibilidad.

Para la Facultad de Arquitectura, el modelo que mayor valor de sensibilidad obtuvo fue el SVM con un valor del 57.14% y también fue este el modelo que obtuvo la menor tasa de error del 1.86%, por lo que para el segundo semestre se escogió éste como mejor modelo. Con una especificidad del 99.24%, predice en casi todos los casos de estudiantes que no presentan deserción y predice un 57.14% de los casos con deserción.

Para la Facultad de Ciencias, el modelo de AD obtuvo un 70.21% de sensibilidad, con una tasa de error del 19.50%, en comparación del modelo de SVM que obtuvo un valor de sensibilidad y tasa de error de 42.55% y 15.42% respectivamente. Se eligió como mejor modelo el AD, ya que el valor de Sensibilidad es por mucho mejor que el del SVM y predice mejor la posibilidad de deserción.

Para la Facultad de Ciencias Agrarias, los modelos que obtuvieron el mayor valor de sensibilidad fueron los entrenados con las técnicas AD y SVM con valores de sensibilidad de 67.19% y 43.75% respectivamente. Al igual que en la Facultad de Ciencias, la diferencia es relativamente alta (23.44%), por lo que se elige como mejor modelo el entrenado con AD.

Tabla 4-18 Resumen de resultados de los modelos para el segundo semestre

FACULTAD		ARQUITECTURA	CIENCIAS	CIENCIAS AGRARIAS	CIENCIAS HUMANAS Y ECONOMICAS	MINAS
SVM	Especificidad	99,24%	89,59%	88,89%	95,22%	89,74%
	Sensibilidad	57,14%	42,55%	43,75%	52,94%	40,26%
	Tasa de Error	1,86%	15,42%	17,27%	7,10%	14,64%
KNN	Especificidad	98,08%	89,09%	89,14%	95,55%	91,94%
	Sensibilidad	37,50%	42,55%	40,63%	38,89%	35,71%
	Tasa de Error	3,72%	15,87%	17,48%	7,74%	13,03%
AD	Especificidad	100,00%	81,73%	72,84%	100,00%	85,64%
	Sensibilidad	0,00%	70,21%	67,19%	0,00%	66,23%
	Tasa de Error	2,60%	19,50%	27,93%	5,48%	16,07%
RL	Especificidad	100,00%	98,16%	92,54%	1	99,13%
	Sensibilidad	0,00%	16,00%	35,71%	5,56%	13,64%
	Tasa de Error	2,76%	10,33%	15,43%	5,21%	8,75%

Para la Facultad de Ciencias Agrarias, los modelos que obtuvieron el mayor valor de sensibilidad fueron los entrenados con las técnicas AD y SVM con valores de sensibilidad de 67.19% y 43.75% respectivamente. Al igual que en la Facultad de Ciencias, la diferencia es relativamente alta (23.44%), por lo que se elige como mejor modelo el entrenado con AD.

Para la Facultad de Ciencias Humanas y Económicas el SVM y el KNN tuvieron los mayores valores de sensibilidad con 52.94 y 38.89 respectivamente, siendo de estos dos el SVM el que obtuvo una menor tasa de error, por lo que se escogió el modelo entrenado con SVM como el mejor modelo.

Para la Facultad de Minas, los modelos entrenados con AD y SVM obtuvieron los mayores valores de sensibilidad, 66.23% y 40.26% respectivamente. La tasa de error para AD fue de 16.07% y para SVM fue de 14.64%, pero dado que la diferencia entre los dos valores de sensibilidad es de 25.97%, se escoge el modelo de AD como el mejor modelo para predecir la deserción en la Facultad de Minas para el segundo semestre.

4.5.3 Tercer Semestre

Para el tercer semestre la proporción de estudiantes que presentan deserción ya es mucho menor en comparación a los que no la presentan, es por esto que para algunas de las técnicas como la AD (ver tabla 4-19) en la Facultad de Arquitectura y Ciencias Humanas y Económicas, el valor de sensibilidad es de 0, es decir que no hacen una clasificación de estudiantes que presenten deserción.

Para la Facultad de Arquitectura, se observa que los modelos que mayor valor de sensibilidad tienen son el SVM y el KNN con un valor de 66.67 en ambos modelos, pero la tasa de error es menor en el SVM (0.40%) por lo que se escoge éste como mejor modelo para predecir la deserción en la Facultad de Arquitectura en el tercer semestre.

Para la Facultad de Ciencias, el AD y el SVM son los que mayor valor de sensibilidad presentan, con un 64.71% y un 41.18% respectivamente, la tasa de error para éstos modelos fue de 6.79% para el SVM y 9.51% para el AD, pero dado que la diferencia en la sensibilidad entre ambos modelos es relativamente alta (23.53%) se selecciona el modelo con AD como el mejor.

Para la Facultad de Ciencias Agrarias, los modelos con SVM y KNN obtuvieron un valor de sensibilidad de 42.86% y 33.33% respectivamente, y la tasa de error fue de 8.72% para SVM y 7.9% para KNN. Dado que las tasas de error están muy cercanas, se seleccionó el modelo entrenado con SVM como el mejor modelo.
Para la facultad de Ciencias Humanas y Económicas, los modelos entrenados con SVM y KNN son los que obtuvieron los mayores valores de sensibilidad, 50% y 40% respectivamente. La tasa de error es la misma para ambos modelos, por lo que se seleccionó el modelo entrenado con SVM como el mejor modelo.

Para la Facultad de Minas, los modelos entrenados con AD y SVM obtuvieron los mayores valores de Sensibilidad de 66.67% y 33.33% respectivamente. El modelo de SVM obtuvo una tasa de error del 6.07% y el de AD obtuvo una tasa de error del 14.6%; se seleccionó el AD como mejor modelo, dado que la diferencia de la sensibilidad de ambos modelos es relativamente alta (33.34%)

Tabla 4-19 Resumen de resultados de los modelos para el tercer semestre

FACULTAD		ARQUITECTURA	CIENCIAS	CIENCIAS AGRARIAS	CIENCIAS HUMANAS Y ECONOMICAS	MINAS
SVM	Especificidad	100,00%	95,73%	94,22%	95,99%	96,24%
	Sensibilidad	66,67%	41,18%	42,86%	50,00%	33,33%
	Tasa de Error	0,40%	6,79%	8,72%	5,63%	6,07%
KNN	Especificidad	98,39%	97,15%	95,66%	96,35%	96,38%
	Sensibilidad	66,67%	35,29%	33,33%	40,00%	22,81%
	Tasa de Error	1,99%	5,71%	7,90%	5,63%	6,33%
AD	Especificidad	100,00%	91,74%	99,13%	100,00%	86,12%
	Sensibilidad	0,00%	64,71%	23,81%	0,00%	66,67%
	Tasa de Error	1,20%	9,51%	5,18%	3,52%	14,60%
RL	Especificidad	99,28%	99,76%	99,75%	98,65%	100,00%
	Sensibilidad	50,00%	25,00%	7,69%	25,00%	0,00%
	Tasa de Error	1,42%	4,39%	5,83%	4,22%	3,99%

Para la Facultad de Minas, los modelos entrenados con AD y SVM obtuvieron los mayores valores de Sensibilidad de 66.67% y 33.33% respectivamente. El modelo de SVM obtuvo una tasa de error del 6.07% y el de AD obtuvo una tasa de error del 14.6%; se seleccionó el AD como mejor modelo, dado que la diferencia de las sensibilidades de ambos modelos es relativamente alta (33.34%)

En la tabla 4-20 se presenta las técnicas utilizadas para el entrenamiento de los modelos que fueron seleccionados para predecir la posibilidad de deserción en cada uno de los semestres en las cinco Facultades

Tabla 4-20 Resumen de Modelos Seleccionados para cada Facultad y Semestre

FACULTAD	PRIMER SEMESTRE	SEGUNDO SEMESTRE	TERCER SEMESTRE
ARQUITECTURA	KNN	SVM	SVM
CIENCIAS	KNN	AD	AD
CIENCIAS AGRARIAS	KNN	AD	SVM
CIENCIAS HUMANAS Y ECONOMICAS	KNN	SVM	SVM
MINAS	AD	AD	AD

Para los datos del primer semestre, los modelos entrenados con KNN fueron seleccionados como los mejores en cuatro de la cinco Facultades, para la Facultad de Minas el modelo

seleccionado fue el de AD. Para el segundo semestre el AD fue seleccionado como mejor modelo en tres de las cinco Facultades y el SVM en dos Facultades. Para el tercer semestre el mejor modelo fue el SVM en tres de las 5 Facultades, en las otras dos se seleccionó el AD. Para la Facultad de Minas el AD obtuvo los mejores resultados en los tres semestres analizados.

La predicción de las nuevas observaciones se podrá hacer a partir del algoritmo elaborado en Matlab para la técnica de Clasificación Supervisada correspondiente y los datos de entrenamiento empleados. A medida que se incluyan nuevas observaciones, el modelo cambia, por lo que es necesario volver a hacer el entrenamiento del mismo.

5 Conclusiones

Los resultados obtenidos con las técnicas utilizadas, dan una aproximación a la predicción de los estudiantes con posibilidad de deserción, pero podría mejorarse el valor de la sensibilidad (predicción de desertores) adicionándose información de algunas otras variables como son entre otras el interés del estudiante para cursar el programa, ya que, por el proceso de admisión a la universidad, muchos estudiantes eligen un programa que no fue su primera opción. Otra variable que es importante considerar es el nivel de apoyo académico con el que cuentan los estudiantes por parte de los familiares, a través del nivel de escolaridad que tienen las personas que conviven con él. En términos generales con los factores identificados como determinantes de la deserción se podría elaborar una encuesta, la cual podría ser aplicada a los estudiantes desde el momento en que son admitidos y haciendo un seguimiento en cada semestre, con el fin de evaluar posibles cambios en las condiciones presentadas inicialmente.

En la aplicación de las técnicas multivariantes, se observó que la técnica de Regresión Logística no obtuvo los mejores resultados en la predicción de la deserción con las variables y los grupos de datos analizados, ya que para todos los casos fue la técnica que dio los menores resultados de sensibilidad.

Los modelos entrenados con las técnicas multivariantes son dinámicos, es decir, su desempeño varía dependiendo de los grupos de datos y variables que se utilicen para el entrenamiento, esto permitió a través de la iteración del algoritmo con grupos de datos de entrenamiento y validación aleatorios. poder seleccionar el mejor modelo, en ese sentido la generación de una ecuación para predecir la deserción en función de los factores analizados no tendría sentido, ya que como se explicó varía según los datos que se utilicen en el entrenamiento. Por lo anterior, lo más importante es una adecuada identificación de cuáles son aquellos factores que más incidencia tienen en la deserción y que puedan ser intervenidos a través de los programas de apoyo con los que cuenta la universidad.

De los 26 programas de pregrado que se ofrecen en la Sede Medellín, 13 programas presentan valores de deserción al tercer semestre por encima del 30%, y 9 programas presentan deserción con valores entre el 20% y el 30%. En promedio el 64% de la deserción de los tres primeros semestres se da en el primer semestre, esta realidad hace que sea importante buscar mecanismos para prevenir la deserción desde el momento en el que los admitidos ingresan a la Universidad, a través de la vinculación de los estudiantes en los

programas de inducción y de apoyo y el acompañamiento permanente por parte de los tutores asignados.

Dada la limitación de recursos con que cuenta la Universidad se debe hacer una identificación adecuada de las personas que requieren ayuda (con posible deserción) y cuál es el tipo de ayuda que necesitan, con el fin de que sean optimizados los recursos según las necesidades reales de los estudiantes y se pueda beneficiar a los estudiantes que tienen mayores posibilidades de deserción, a partir de la identificación inicial que se realice.

Con los resultados obtenidos se pueden identificar un grupo de estudiantes con posibilidad de deserción, pero es necesario hacer un análisis individual de las características de éstos estudiantes con el fin de identificar aquellas falencias o causas de la posible deserción, por ejemplo, la identificación de si es estudiante tiene dificultades económicas, de adaptación a la vida universitaria, problemas con los métodos de aprendizaje, problemas en el grupo familiar, entre otros.

Se debe evaluar el proceso de selección para los estudiantes y en particular los de la Facultad de Ciencias Agrarias, ya que analizando los datos de deserción, para tres de los cuatro programas de pregrado que se ofrecen en esta Facultad la deserción en el tercer semestre supera el 40%, esto se puede explicar entre otras causas a que gran parte de los estudiantes que eligen éstos programas lo hacen porque no pueden acceder a su primera opción y lo tienen que seleccionar por ser uno de los que tiene cupo disponible al momento de la inscripción.

Al analizar las cifras de deserción de la Sede Medellín y comparándolas con las cifras de la Sede Bogotá, se observa una gran diferencia entre ambas Sedes, esto se podría explicar en parte por la cantidad de aspirantes que se presentan a los programas de ambas Sedes, al tener la Sede Bogotá un mayor número de aspirantes para los programas que ofrece, los aspirantes que son admitidos generalmente tienen mayores puntajes o su nivel académico al ingresar es más alto. Por lo anterior la Sede debe buscar hacer una mayor divulgación de los programas en los diferentes colegios, con el fin de que captar un mayor número de estudiantes de secundaria que se interesen por los programas que ofrece la Sede.

En los resultados del examen de admisión, se evidencia a través de la cantidad de admitidos que tienen que cursar la nivelación en Matemáticas Básicas el bajo nivel académico con el que están saliendo los estudiantes de secundaria. Lo cual se traduce en los valores altos de deserción en la universidad en el primer semestre. La Universidad adopto la estrategia de ofrecer los cursos de nivelación en Matemáticas Básicas con el fin de preparar a estos estudiantes para superar las diferentes asignaturas del área de Matemáticas, pero sería conveniente evaluar el impacto que está teniendo en el rendimiento académico de los estudiantes el tener esta asignatura dentro del cálculo del PAPA, dado que es una asignatura de “Nivelación” y que no hace parte como tal de su plan de estudios.

Para la Facultad de Minas, el programa que presentó un mayor porcentaje de deserción fue el de Ingeniería de Sistemas e informática con una deserción al tercer semestre del 44%, también se pudo observar que para este programa, en promedio el puntaje de admisión

para el componente de Matemáticas esta entre los más bajos de la Sede y es el más bajo de la Facultad de Minas. Por lo anterior se recomienda hacer un análisis individual con los datos de este programa, y poder así identificar cuáles son las causas de deserción.

Bibliografía

Ben-Gal, Irad (2005) “Outlier detection”, Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers," Kluwer Academic Publishers. recuperado de http://www.eng.tau.ac.il/~bengal/outlier.pdf

Carmona Suarez, Enrique (2014) Tutorial sobre Máquinas de Vectores Soporte (SVM) en recuperado de http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf

Castejón Sandoval, Osiris (2011) Diseño y análisis de experimentos con Statistix Fondo Editorial Biblioteca Universidad Rafael Urbaneta recuperado de http://www.uru.edu/fondoeditorial/libros/pdf/manualdestatistix/occompleto.pdf

CIDSE (2011). Análisis de la herramienta SPADIES diseñada por el Ministerio de Educación Nacional y el CEDE Recuperado de http://www.alfaguia.org/alfaguia/files/1320349109_33.pdf

C. M. Cuadras (2014). Nuevos Métodos de Análisis Multivalente CMC Editions Barcelona recuperado de http://www.ub.edu/stat/personal/cuadras/metodos.pdf

Diaz,Christian (2008). Modelo Conceptual Para La Deserción Estudiantil Universitaria Chilena. Estudios Pedagógicos XXXIV, N° 2: 65-86

Fernández, Santiago (2011). Análisis Discriminante. Universidad Autónoma de Madrid. Recuperado de http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/SEGMENTACION/DISCRIMINANTE/analisis-discriminante.pdf

MEN (2014). Determinantes de la Deserción. Bogotá, Colombia Recuperado de en: http://www.mineducacion.gov.co/sistemasdeinformacion/1735/articles-254702_Informe_determinantes_desercion.pdf

MEN (2017). *¿Qué es el SPADIES? - Sistemas información*. Disponible en: http://www.mineducacion.gov.co/sistemasdeinformacion/1735/w3-article-254648.html

OCDE. PISA 2015 Recuperado de http://www.oecd.org/centrodemexico/estadisticas/

OpenCV dev team (2011-2014) Introduction to Support Vector Machines, OpenCV 2.4.13.2 documentation. Recuperado de http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html

Peña, Daniel (2002). Análisis de Datos Multivariados. McGraw-Hill Interamericana de España, S.A.U.

Perez Diez, Jose L. (1987) “Identificación de Outliers en Muestras Multivariantes” (Tesis de Doctorado). Universidad de Sevilla. Sevilla, España. Recuperado de http://fondosdigitales.us.es/tesis/tesis/1411/identificacion-de-outliers-en-muestras-multivariantes/

Rodriguez Rodriguez, Jorge E., Rojas Blanco, Edwar., Franco Camacho, Roger. Clasificación de datos usando el método K-NN. Revista Vinculos, Universidad Distrital. Diciembre de 2007 Volumen 4 número 1 pagina 10. Recuperado de http://revistas.udistrital.edu.co/ojs/index.php/vinculos/article/viewFile/4111/5778

Salvador Figueras, M (2000): "Introducción al Análisis Multivariante", Recuperado de http://www.5campus.com/leccion/anamul

Sancho Caparrini, Fernando (2017) Clasificación Supervisada y No Supervisada Recuperado de http://www.cs.us.es/~fsancho/?e=77

Tinto, Vincent (1989). “Definir la deserción: una cuestión de perspectiva”, Revista de Educación Superior, 71, México. Recuperado de http://publicaciones.anuies.mx/pdfs/revista/Revista71_S1A3ES.pdf

UNESCO (2014) Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura. (2004). Repetition at high cost in Latin America and the Caribbean. IESALC - UNESCO. Recuperado de http://www.unesco.org/fileadmin/MULTIMEDIA/HQ/ED/ED_new/pdf/LAC-GEM-2014-ENG.pdf

Zuluaga, Carlos Mario (2011) “Análisis Estadístico Multivariado: una Herramienta Estratégica para el Control de Procesos y Calidad en la Industria Agroalimentaria” Publicaciones e Investigación (5), P 143 recuperado de: https://academia.unad.edu.co/images/investigacion/hemeroteca/PeI/volumen5_2011/analisis%20estadistico%20multivariado.pdf

A. Anexo: Número de estudiantes analizados por periodo y plan

Tabla A-1 Número de estudiantes analizados por periodo y plan

PLAN DE ESTUDIOS	2009-01	2009-03	2010-01	2010-03	2011-01	2011-03	2012-01	2012-03	2013-01	2013-03	2014-01	2014-03	2015-01	2015-03	2016-01	TOTAL SEDE
ARQUITECTURA	44	43	61	46	56	41	53	53	45	47	44	47	63	53	54	750
ARTES PLÁSTICAS							14			17	17	19	23	18	14	122
CONSTRUCCIÓN	21	25	38	27	25	32	28	28	32	33	28	30	34	36	32	449
FACULTAD DE ARQUITECTURA	**65**	**68**	**99**	**73**	**81**	**73**	**95**	**81**	**77**	**97**	**89**	**96**	**120**	**107**	**100**	**1321**
ESTADÍSTICA	33	37	44	47	54	41	45	45	44	37	36	51	32	46	54	646
INGENIERÍA BIOLÓGICA	65	43	74	58	57	54	56	55	56	45	55	61	67	45	66	857
INGENIERÍA FÍSICA	30	36	58	48	47	51	47	58	54	49	48	57	63	57	52	755
MATEMÁTICAS	18	28	38	16	36	39	41	27	23	21	17	34	39	41	53	471
FACULTAD DE CIENCIAS	**146**	**144**	**214**	**169**	**194**	**185**	**189**	**185**	**177**	**152**	**156**	**203**	**201**	**189**	**225**	**2729**
INGENIERÍA AGRÍCOLA	44	47	69	37	62	65	68	46	53	41	35	59	62	54	53	795
INGENIERÍA AGRONÓMICA	46	46	65	40	56	47	67	35	25	20	23	30	30	35	34	599
INGENIERÍA FORESTAL	40	35	68	60	64	66	55	35	31	47	48	62	60	60	66	797
ZOOTECNIA	55	47	56	47	66	24	65	24	47	16	22	35	46	47	29	626
FACULTAD DE CIENCIAS AGRARIAS	**185**	**175**	**258**	**184**	**248**	**202**	**255**	**140**	**156**	**124**	**128**	**186**	**198**	**196**	**182**	**2817**
CIENCIA POLÍTICA	28	33	45	60	56	37	30	36	32	28	23	34	37	44	42	565
ECONOMÍA	32	46	51	36	68	36	20	27	34	24	31	29	40	45	35	554
HISTORIA	13	27	46	44	54	36	25	37	30	26	20	36	34	41	36	505
FACULTAD DE CIENCIAS HUMANAS Y ECONOMICAS	**73**	**106**	**142**	**140**	**178**	**109**	**75**	**100**	**96**	**78**	**74**	**99**	**111**	**130**	**113**	**1624**
INGENIERÍA ADMINISTRATIVA	69	75	65	76	79	67	66	56	60	74	67	50	51	44	47	946
INGENIERÍA AMBIENTAL				45	52	30	34	28	29	31	29	44	42	40	37	441
INGENIERÍA CIVIL	81	77	71	71	76	69	76	57	74	79	70	84	85	79	80	1129
INGENIERÍA DE CONTROL	41	25	41	40	39	35	37	24	52	46	45	58	53	44	48	628
INGENIERÍA DE MINAS Y METALURGIA	38	31	31	33	39	28	34	28	46	54	49	64	62	36	50	623

PLAN DE ESTUDIOS	2009-01	2009-03	2010-01	2010-03	2011-01	2011-03	2012-01	2012-03	2013-01	2013-03	2014-01	2014-03	2015-01	2015-03	2016-01	TOTAL, SEDE
INGENIERÍA DE PETRÓLEOS	59	55	47	39	53	43	45	36	43	53	45	48	45	52	42	705
INGENIERÍA DE SISTEMAS E INFORMÁTICA	72	77	74	57	74	67	64	63	58	50	58	62	68	52	78	974
INGENIERÍA ELÉCTRICA	35	26	43	41	50	40	35	42	56	49	49	54	46	58	66	690
INGENIERÍA GEOLÓGICA	33	29	41	36	36	32	27	30	42	60	57	61	71	35	56	646
INGENIERÍA INDUSTRIAL	77	73	73	71	74	69	70	59	66	78	65	56	45	46	40	962
INGENIERÍA MECÁNICA	79	71	64	78	68	72	61	45	47	51	54	62	69	68	70	959
INGENIERÍA QUÍMICA	66	66	70	56	62	58	61	56	55	58	66	74	72	70	66	956
FACULTAD DE MINAS	**650**	**605**	**620**	**643**	**702**	**610**	**610**	**524**	**628**	**683**	**654**	**717**	**709**	**624**	**680**	**9659**
TOTAL SEDE	**1119**	**1098**	**1333**	**1209**	**1403**	**1179**	**1224**	**1030**	**1134**	**1134**	**1101**	**1301**	**1339**	**1246**	**1300**	**18150**

B. Anexo: Deserción Acumulada para los tres primeros tres semestres

Tabla B-1 Deserción Acumulada para los primeros tres semestres

PLAN DE ESTUDIOS	TOTAL ESTUDIANTES	SALIERON PRIMER SEMESTRE	SALIERON SEGUNDO SEMESTRE	SALIERON TERCER SEMESTRE	DESER. ACUM. PRIMER SEMESTRE	DESER. ACUM. SEGUNDO SEMESTRE	DESER. ACUM. TERCER SEMESTRE
ARQUITECTURA	750	54	20	14	7%	10%	12%
ARTES PLÁSTICAS	122	9	4	0	7%	11%	11%
CONSTRUCCIÓN	449	61	18	5	14%	18%	19%
TOTAL FACULTAD DE ARQUITECTURA	**1321**	**124**	**42**	**19**	**9%**	**13%**	**14%**
ESTADÍSTICA	646	148	54	31	23%	31%	36%
INGENIERÍA BIOLÓGICA	857	128	59	35	15%	22%	26%
INGENIERÍA FÍSICA	755	179	57	18	24%	31%	34%
MATEMÁTICAS	471	152	48	17	32%	42%	46%
TOTAL FACULTAD DE CIENCIAS	**2729**	**607**	**218**	**101**	**22%**	**30%**	**34%**
INGENIERÍA AGRÍCOLA	795	184	96	40	23%	35%	40%
INGENIERÍA AGRONÓMICA	599	167	59	24	28%	38%	42%
INGENIERÍA FORESTAL	797	146	84	28	18%	29%	32%
ZOOTECNIA	626	230	61	18	37%	46%	49%
TOTAL FACULTAD DE CIENCIAS AGRARIAS	**2817**	**727**	**300**	**110**	**26%**	**36%**	**40%**
CIENCIA POLÍTICA	565	46	21	8	8%	12%	13%
ECONOMÍA	554	61	39	27	11%	18%	23%
HISTORIA	505	92	19	18	18%	22%	26%
TOTAL FACULTAD DE CIENCIAS HUMANAS Y ECONOMICAS	**1624**	**199**	**79**	**53**	**12%**	**17%**	**20%**
INGENIERÍA ADMINISTRATIVA	946	162	50	18	17%	22%	24%
INGENIERÍA AMBIENTAL	441	70	18	9	16%	20%	22%
INGENIERÍA CIVIL	1129	120	71	31	11%	17%	20%
INGENIERÍA DE CONTROL	628	170	49	22	27%	35%	38%
INGENIERÍA DE MINAS Y METALURGIA	623	139	49	17	22%	30%	33%
INGENIERÍA DE PETRÓLEOS	705	102	42	24	14%	20%	24%

PLAN DE ESTUDIOS	TOTAL ESTUDIANTES	SALIERON PRIMER SEMESTRE	SALIERON SEGUNDO SEMESTRE	SALIERON TERCER SEMESTRE	DESER. ACUM. PRIMER SEMESTRE	DESER. ACUM. SEGUNDO SEMESTRE	DESER. ACUM. TERCER SEMESTRE
INGENIERÍA DE SISTEMAS E INFORMÁTICA	974	252	134	42	26%	40%	44%
INGENIERÍA ELÉCTRICA	690	163	55	27	24%	32%	36%
INGENIERÍA GEOLÓGICA	646	145	53	14	22%	31%	33%
INGENIERÍA INDUSTRIAL	962	125	84	40	13%	22%	26%
INGENIERÍA MECÁNICA	959	200	75	28	21%	29%	32%
INGENIERÍA QUÍMICA	956	155	78	19	16%	24%	26%
TOTAL FACULTAD DE MINAS	**9659**	**1803**	**758**	**291**	**19%**	**27%**	**30%**
TOTAL general	**18150**	**3460**	**1397**	**574**	**19%**	**27%**	**30%**

C. Anexo: Codificación de variables dummy

Tabla C-1 Variables dummy para cada estado civil

ESTADO CIVIL	EC_1	EC_2	EC_3	EC_4	EC_5	EC_6
Casado	0	0	0	0	0	0
Soltero	1	0	0	0	0	0
Separado	0	1	0	0	0	0
Unión Libre	0	0	1	0	0	0
Viudo	0	0	0	1	0	0
Divorciado	0	0	0	0	1	0
No definido	0	0	0	0	0	1

Tabla C-2 Variables dummy para tipo colegio

CLASIFICACIÓN DEL COLEGIO	TC_1	TC_2	TC_3	TC_4
A+	0	0	0	0
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

Tabla C-3 Variables dummy para estrato

ESTRATO	E_1	E_2	E_3	E_4	E_5
1	0	0	0	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6 o no informa	0	0	0	0	1

Tabla C-4 Variables dummy para propiedad de la vivienda

PROPIEDAD DE LA VIVIENDA	PV_1	PV_2
Sin propiedad raíz	0	0
Pagando crédito Hipoteca	1	0
Vivienda propia	0	1

Tabla C-5 Variables dummy para número de hijos

NÚMERO DE HIJOS	NH_1	NH_2	NH_3
Siete o más	0	0	0
Cinco o Seis	1	0	0
Tres o Cuatro	0	1	0
Uno o Dos no informa	0	0	1

D. Anexo: Coeficientes estimados para Regresión Logística

Tabla D-1 Estimación Final de coeficientes para Regresión Logística Facultad de Arquitectura Primer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6
    Distribution = Binomial

Estimated Coefficients:
                    Estimate      SE        tStat       pValue
                    ________    _______    _______    __________

    (Intercept)     -3.6543      0.29313  -12.466     1.1392e-35
    Género           0.49984     0.2222     2.2495    0.024478
    TC1              1.0418      0.2977     3.4995    0.00046607
    TC2              1.0025      0.31131    3.2203    0.0012807
    TC3              1.3446      0.37716    3.5651    0.00036372
    TC4              1.3351      0.38167    3.498     0.0004688
    Mat. Basicas     0.53774     0.19567    2.7482    0.0059918
```

Tabla D-2 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias Primer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
x12 + x13 + x14
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    _______    _______    __________

    (Intercept)       8.1998      1.3921     5.8901    3.8592e-09
    P_MAT            -0.34948    0.071409   -4.8941    9.8771e-07
    P_CIENC          -0.29028    0.060038    -4.835    1.3318e-06
    P_SOC            -0.15295       0.061   -2.5074      0.012161
    P_TEX            -0.22598    0.071191   -3.1743     0.0015019
    Género            0.59877    0.10624     5.6359    1.7412e-08
    Car_col            0.2691     0.1159     2.3217      0.020247
    TC1               0.47732    0.15738      3.033     0.0024216
    TC2               0.64312    0.16792     3.8299    0.00012821
    TC3               0.74897    0.19314      3.878    0.00010533
    TC4               0.67783     0.2132     3.1794      0.001476
    PV_2              0.22954    0.11296     2.0321      0.042141
    Edad             0.043702    0.014339    3.0477     0.0023063
    MB               -0.37944    0.17305    -2.1926      0.028333
    LE               -0.86442    0.16045    -5.3874    7.1503e-08


2729 observations, 2714 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 204, p-value = 1.02e-35
```

Tabla D-3 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias Agrarias Primer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    _______    __________

    (Intercept)       6.5946      1.1911     5.5363    3.0887e-08
    P_MAT            -0.39328    0.063207    -6.2221    4.9061e-10
    P_CIENC          -0.26982    0.060848    -4.4343     9.236e-06
    P_IMG            -0.22446    0.058603    -3.8302    0.00012803
    GÉNERO            0.66505    0.095789     6.9428    3.8429e-12
    TC2               0.26181     0.10011     2.6153     0.0089138
    TC3               0.26833     0.13288     2.0193      0.043454
    LRES              0.29801    0.093436     3.1895     0.0014252
    EDAD             0.051697    0.015187      3.404    0.00066409
    LE               -0.31833     0.11164    -2.8515     0.0043511


2817 observations, 2807 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 167, p-value = 2.81e-31
```

Tabla D-4 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias Humanas y Económicas Primer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    _______    __________

    (Intercept)      -4.3408     0.37019    -11.726    9.4047e-32
    x1               0.35608     0.16995     2.0952      0.036156
    x2               0.39621     0.21215     1.8676      0.061822
    x3               0.10602    0.017695     5.9917    2.0766e-09


1624 observations, 1620 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 48.4, p-value = 1.77e-10
```

Tabla D-5 Estimación Final de coeficientes para Regresión Logística Facultad de Minas Primer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 +
x12 + x13 + x14
    Distribution = Binomial

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                    _________    _________    _______    __________

    (Intercept)        5.6097      0.68758     8.1587     3.387e-16
    P_mat             -0.46633     0.041108    -11.344    7.9473e-30
    P_cienc           -0.28728     0.034134    -8.4162    3.8881e-17
    P_soc             -0.08331     0.034367    -2.4241      0.015346
    Género             0.64081     0.070388     9.1039    8.7123e-20
    E3                 -1.1118      0.42591    -2.6104     0.0090443
    TC1                0.28065      0.07932     3.5382    0.00040283
    TC2                0.45848     0.082218     5.5765    2.4544e-08
    TC3                0.52138      0.10058     5.1836    2.1762e-07
    TC4                0.49741      0.10836     4.5905    4.4222e-06
    Ingresos         0.0029695    0.0010405      2.854     0.0043169
    Lugar_res          0.16851     0.057611     2.9249     0.0034456
    Edad              0.075052     0.009045     8.2976    1.0623e-16
    Mat. Basicas      -0.24437     0.085689    -2.8518     0.0043466
    Lectoescritura    -0.51725     0.087003    -5.9452    2.7607e-09


9659 observations, 9644 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 627, p-value = 7.59e-125
```

Tabla D-6 Variables significativas para el modelo de Regresión Logística correspondiente a los datos del primer semestre

N°	VARIABLE	ARQUITECTURA	CIENCIAS	CIENCIAS AGRARIAS	CIENCIAS HUMANAS	MINAS
1	HSTMATEMATICAS		X	X		X
2	HSTCIENCIAS		X	X		X
3	HSTSOCIALES		X			X
4	HSTTEXTUAL		X			
5	HSTIMAGEN			X		
6	GÉNERO	X	X	X	X	X
7	E1					
8	E2					
9	E3					X
10	E4					
11	E5					
12	CAR_COL		X			
13	EC1					
14	EC2					
15	EC3					
16	EC4					
17	EC5					
18	EC6					
19	TC1	X	X			X
20	TC2	X	X	X		X
21	TC3	X	X	X	X	X
22	TC4	X	X			X
23	VRPE					
24	INGR					X
25	LRES			X		X
26	PV_1					
27	PV_2					
28	NH_1					
29	NH_2					
30	NH_3					
31	EDAD		X	X	X	X
32	MB	X	X			X
33	LE		X	X		X

Tabla D-7 Estimación Final de coeficientes para Regresión Logística Facultad de Arquitectura segundo semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x25 + x31 + x38
    Distribution = Binomial

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)     5.1952       2.1879     2.3745      0.017571
    LRES            1.0127       0.5055     2.0035      0.045128
    EDAD            0.1219     0.046579     2.6171     0.0088689
    PROMEDIO       -3.3157      0.59256    -5.5956    2.1981e-08


1160 observations, 1156 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 48.1, p-value = 2.02e-10
```

Tabla D-8 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias segundo semestre

```
mdl =


Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6
    Distribution = Binomial

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    __________

    (Intercept)      10.333      1.4642      7.057    1.7015e-12
    HSTTEXTUAL     -0.24763    0.095843    -2.5837     0.0097755
    EC_6             2.4512      1.0438     2.3483       0.01886
    LRES             0.5045      0.1739     2.9012     0.0037176
    PMB             0.55139      0.1811     3.0447     0.0023294
    EFICIENCIA     -0.71508     0.34347     -2.082      0.037345
    PROMEDIO        -2.8044      0.2981    -9.4076    5.0742e-21


1938 observations, 1931 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 278, p-value = 3.32e-57
```

Tabla D-9 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias Agrarias segundo semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE       tStat       pValue
                    ________    ________    _______    __________

    (Intercept)      10.126       1.4121    7.1706    7.4694e-13
    HSTTEXTUAL       -0.212     0.092256    -2.298      0.021562
    PMB             0.71797      0.17846    4.0232    5.7411e-05
    PLE              1.2474      0.57073    2.1856      0.028847
    PROMEDIO         -2.996      0.28344    -10.57     4.097e-26


1998 observations, 1993 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 270, p-value = 3.05e-57
```

Tabla D-10 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias Humanas y Económicas segundo semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    _______    __________

    (Intercept)       7.8599      1.4543     5.4045    6.4976e-08
    EDAD            0.073693    0.030551     2.4121      0.015861
    PROMEDIO         -3.2801     0.37971    -8.6384    5.7006e-18


1306 observations, 1303 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 98.4, p-value = 4.32e-22
```

Tabla D-11 Estimación Final de coeficientes para Regresión Logística Facultad de Minas segundo semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    _______    __________

    (Intercept)       7.2221     0.78797     9.1654    4.9363e-20
    HSTIMAGEN        0.09524    0.046266     2.0585      0.039538
    TIPO_COLEGIO     -0.3357    0.090438    -3.7119    0.00020569
    LRES             0.28622    0.089057     3.2139     0.0013096
    PMB              0.92488     0.10145     9.1163    7.7726e-20
    EFICIENCIA       -1.1455     0.17869    -6.4104    1.4511e-10
    PERDIDAS        -0.20318    0.070911    -2.8652      0.004167
    PROMEDIO          -2.823     0.17664    -15.982    1.7108e-57


7632 observations, 7624 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 1.03e+03, p-value = 9.39e-218
```

Tabla D-12 Variables significativas para el modelo de Regresión Logística correspondiente a los datos del segundo semestre

N°	VARIABLE	ARQUITECTURA	CIENCIAS	CIENCIAS AGRARIAS	CIENCIAS HUMANAS	MINAS
1	HSTMATEMATICAS					
2	HSTCIENCIAS					
3	HSTSOCIALES					
4	HSTTEXTUAL		X	X		
5	HSTIMAGEN					X
6	GÉNERO					
7	E_1					
8	E_2					
9	E_3					
10	E_4					
11	E_5					
12	TIPO_COLEGIO					X
13	EC_1					
14	EC_2					

N°	VARIABLE	ARQUITECTURA	CIENCIAS	CIENCIAS AGRARIAS	CIENCIAS HUMANAS	MINAS
15	EC_3					
16	EC_4					
17	EC_5					
18	EC_6		X			
19	TC_1					
20	TC_2					
21	TC_3					
22	TC_4					
23	VRPE					
24	INGR					
25	LRES	X	X			X
26	PV_1					
27	PV_2					
28	NH_1					
29	NH_2					
30	NH_3					
31	EDAD	X			X	
32	MB					
33	LE					X
34	PMB		X	X		
35	PLE			X		
36	EFICIENCIA		X			X
37	PERDIDAS					X
38	PROMEDIO	X	X	X	X	X

Tabla D-13 Estimación Final de coeficientes para Regresión Logística Facultad de Arquitectura tercer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x1 + x37 + x38
    Distribution = Binomial

Estimated Coefficients:
                    Estimate      SE        tStat       pValue
                    ________    _______    _______    __________

    (Intercept)       24.268     5.3264     4.5562    5.2096e-06
    x1               -0.94206    0.32377    -2.9096     0.0036184
    x37               -1.2343    0.50914    -2.4243      0.015338
    x38               -4.9185     1.1098    -4.4318     9.347e-06


1127 observations, 1123 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 37.7, p-value = 3.21e-08
```

Tabla D-14 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias tercer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x4 + x26 + x30 + x31 + x34 + x36 + x38
    Distribution = Binomial

Estimated Coefficients:
                    Estimate      SE        tStat       pValue
                    ________    _______    _______    __________

    (Intercept)       4.8629     2.0038     2.4269       0.01523
    x4               -0.35605    0.13048    -2.7289     0.0063554
    x30                2.3832    0.88699     2.6869      0.007212
    x31               0.16145    0.02979     5.4198     5.966e-08
    x34               0.62201    0.25307     2.4578      0.013979
    x36              -0.98472    0.47082    -2.0915      0.036482
    x38               -1.7849     0.3721    -4.7969    1.6113e-06


1735 observations, 1727 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 146, p-value = 2.98e-28
```

Tabla D-15 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias Agrarias tercer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x35 + x38
    Distribution = Binomial

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)     7.0519      1.2447     5.6653    1.4677e-08
    x35               2.093     0.70037     2.9885     0.0028037
    x38            -2.8421     0.37203    -7.6393    2.1841e-14


1718 observations, 1715 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 84.8, p-value = 3.92e-19
```

Tabla D-16 Estimación Final de coeficientes para Regresión Logística Facultad de Ciencias Humanas y Económicas tercer semestre

```
Generalized linear regression model:
    logit(y) ~ 1 + x12 + x24 + x25 + x36 + x38
    Distribution = Binomial

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   _________    _________    _______    __________

    (Intercept)       2.0351       1.4905     1.3654       0.17214
    x12             -0.71284       0.3599    -1.9806      0.047631
    x24            -0.014329    0.0086292    -1.6605      0.096816
    x25               0.8374      0.35257     2.3751      0.017542
    x36              -2.2616      0.59562     -3.797    0.00014644
    x38             -0.85625      0.41332    -2.0716      0.038298


1233 observations, 1227 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 51.7, p-value = 6.14e-10
```

Tabla D-17 Estimación Final de coeficientes para Regresión Logística Facultad de Minas tercer semestre

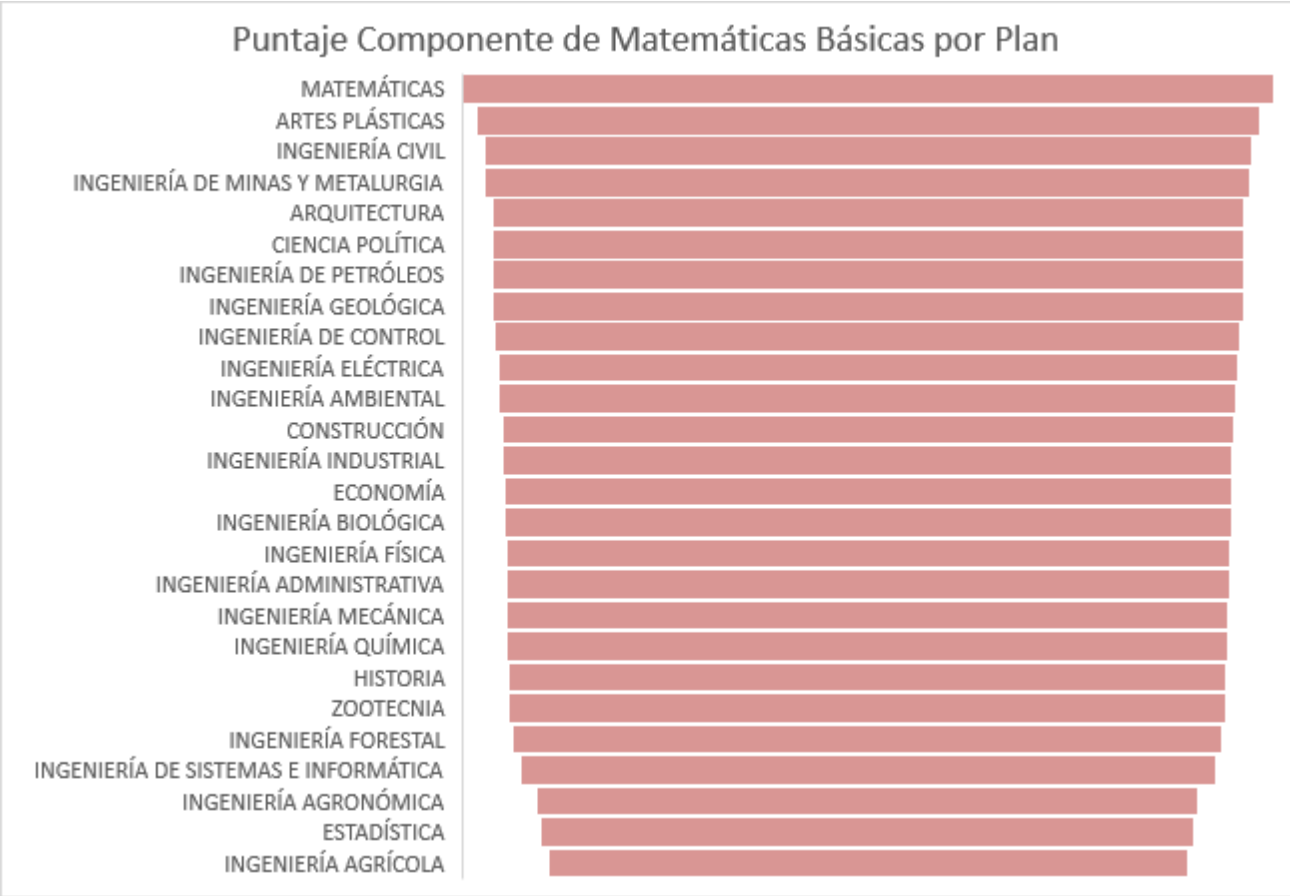
```
Generalized linear regression model:
    logit(y) ~ 1 + x7 + x9 + x25 + x31 + x33 + x34 + x36 + x37 + x38
    Distribution = Binomial

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    _______    __________

    (Intercept)       3.8754     0.95917     4.0403    5.3381e-05
    x7               0.31841      0.1308     2.4343      0.014921
    x9               -0.6597     0.30932    -2.1328      0.032945
    x25              0.39759     0.13324     2.9841     0.0028445
    x31             0.067326    0.018519     3.6355    0.00027742
    x33               0.3675     0.16247     2.2619      0.023703
    x34              0.59232     0.16064     3.6873    0.00022662
    x36              -1.1239     0.25206    -4.4591    8.2315e-06
    x37             -0.27479     0.10777    -2.5499      0.010775
    x38              -2.1816     0.24306    -8.9756    2.8181e-19


6927 observations, 6917 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 305, p-value = 1.79e-60
```

E. Anexo: Puntaje promedio para el componente de Matemáticas por plan



F. Anexo: Puntajes de admisión por programa para el primer semestre de 2017

PROGRAMA	PUNTAJE ESTÁNDAR				Número de admitidos
	ÚLTIMO ADMITIDO POR GRUPO DE CLASIFICACIÓN			ÚLTIMO ADMITIDO	
	G1	G2	G3		
Arquitectura	651,4678			651,4678	60
Ciencia Política	625,2425	582,5847		582,5847	60
Construcción	625,3595	606,1204		606,1204	42
Economía	626,0676	601,3474		601,3474	58
Estadística	629,2099	575,2391	573,8531	573,8531	85
Historia	625,1624	575,6592	570,7043	570,7043	65
Ingeniería Administrativa	634,5134			634,5134	50
Ingeniería Agrícola	659,5238	575,7891	576,6510	575,7891	90
Ingeniería Agronómica	629,5823	575,0896	583,5528	575,0896	82
Ingeniería Ambiental	629,0691			629,0691	50
Ingeniería Biológica	625,4201	616,3993		616,3993	68
Ingeniería Civil	659,8226			659,8226	84
Ingeniería de Control	625,3551	611,8310		611,8310	69
Ingeniería de Minas y Metalurgia	625,2403	603,8378		603,8378	55
Ingeniería de Petróleos	627,3605	608,9096		608,9096	57
Ingeniería de Sistemas e Informática	638,0780			638,0780	88
Ingeniería Eléctrica	626,3166	621,4506		621,4506	70
Ingeniería Física	625,5923	604,4542		604,4542	80
Ingeniería Forestal	626,3556	588,9407		588,9407	70
Ingeniería Geológica	631,9862			631,9862	52
Ingeniería Industrial	627,2500			627,2500	47
Ingeniería Mecánica	663,0128			663,0128	73
Ingeniería Química	625,2468	629,0020		625,2468	63
Matemáticas	632,1767	575,4968	591,2211	575,4968	75
Zootecnia	643,7908	575,2759	551,9265	551,9265	80

G. Anexo: Gráfico del número de estudiantes con deserción al tercer semestre