IDENTIFICACIÓN DE CARACTERÍSTICAS DE RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE PREGRADO DE LA UNIVERSIDAD DE NARIÑO ENTRE LOS AÑOS 2010 Y 2014 USANDO MINERÍA DE DATOS

SANDRA VIVIANA ESCOBAR MADROÑERO
JAIME HARVEY ENRÍQUEZ TULCÁN

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍA INDUSTRIAL
EN CONVENIO CON LA UNIVERSIDAD DE NARIÑO
SAN JUAN DE PASTO
2018

IDENTIFICACIÓN DE CARACTERÍSTICAS DE RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE PREGRADO DE LA UNIVERSIDAD DE NARIÑO ENTRE LOS AÑOS 2010 Y 2014 USANDO MINERÍA DE DATOS

SANDRA VIVIANA ESCOBAR MADROÑERO
JAIME HARVEY ENRÍQUEZ TULCÁN

Trabajo de grado presentado como requisito para optar el título de magister en Investigación de Operaciones y Estadística

Director
Hernán Abdón García
Magister en estadística

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍA INDUSTRIAL
EN CONVENIO CON LA UNIVERSIDAD DE NARIÑO
SAN JUAN DE PASTO
2018

Nota de aceptación

_____
_____
_____
_____
_____
_____

_____
Presidente del jurado

_____
Jurado

_____
Jurado

San Juan de Pasto, 09 de Septiembre de 2018

AGRADECIMIENTOS

Agradecemos profundamente a la universidad Tecnológica de Pereira y sus profesores de la maestría en investigación operativa y estadística por permitir que hayamos podido ejecutar este proyecto de investigación.

Agradecemos a la Universidad de Nariño y los profesores vinculados con esta maestría porque sin ellos no habríamos podido lograr este objetivo tan anhelado.

De manera especial, agradecemos a nuestro asesor, el magister en estadística y profesor, Hernán García por su orientación y acertados consejos.

Y en general, agradecemos a todas las personas presentes en todo este proceso, profesores, compañeros, colegas por su incondicional apoyo.

CONTENIDO

pág.

LISTA DE TABLAS

pág.

LISTA DE FIGURAS

pág.

INTRODUCCIÓN

La Minería de Datos es un campo interdisciplinario, que en términos generales; organiza, procesa, analiza y genera reportes de grandes volúmenes de datos con el objetivo de descubrir información útil en la forma de relaciones, tendencias y patrones significativos y nuevos que pueden ayudar en la toma de decisiones y mejoramiento de procesos para organizaciones o empresas de diferente tipo.

(Pérez López y Santín Gonzalez, 2007) mencionan respecto de la minería de datos “Las técnicas de minería de datos persiguen el descubrimiento automático del conocimiento contenido en la información almacenada de modo ordenado en grandes bases de datos. Estas técnicas tienen como objetivo descubrir patrones, perfiles, y tendencias a través del análisis de los datos utilizando tecnologías de reconocimiento de patrones, redes neuronales, lógica difusa, algoritmos genéticos y otras técnicas avanzadas de análisis de datos”, a esta lista de técnicas se puede agregar: clustering, predicción y otras técnicas de análisis multivariado.

En (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004) "La información histórica es útil para explicar el pasado, entender el presente y predecir la información futura”, es precisamente el objetivo principal de la Minería de Datos a través del uso de las diferentes técnicas que la componen.

Las líneas de trabajo en el ámbito de la minería de datos tienen sus orígenes en tres desarrollos teóricos importantes:

- Estadística Clásica
- Inteligencia Artificial
- Aprendizaje Automático

Y producen cinco tipos de información:

- Asociaciones
- Secuencias
- Clasificaciones
- Agrupamientos
- Pronósticos

Entonces se podría dividir la minería de datos en:

- Técnicas predictivas: en estas las variables pueden clasificarse inicialmente en dependiente e independiente y además estiman valores futuros o desconocidos de variables de interés.

- Técnicas descriptivas: explican o resumen los datos, es decir, exploran las propiedades de los datos pero no para predecir.

Pero también se puede dividir las técnicas de minería de datos en:

- Técnicas de aprendizaje supervisado (orientado): aquellas que permiten la clasificación o predicción a partir de ejemplos existentes (Labeled data). Ejemplos de estas técnicas son: Bayes Naïve, Arboles de Decisión, Regresión, Vecinos Más Cercanos, Máquinas de Vectores de Soporte, Bosques Aleatorios, etc (Perez Marqués, 2015).

- Técnicas de aprendizaje no supervisado (no orientado): aquellas que agrupan (clustering), crean reglas de asociación, crean mapas autoorganizados (Hastie, Tibshirani, y Friedman, 2009), etc., con los datos de acuerdo a las características y relaciones inherentes entre ellos (Unlabeled data). Ejemplos de estas técnicas son: Agrupación Jerárquica (Hierarchical clustering), Redes Neuronales, Componentes Principales, Escala Multidimensional, y los algoritmos K-means, DBSCAN, etc, (Pérez Marques, 2013).

En los últimos años, ha habido un interés creciente en utilizar la Minería de Datos en la educación, en colegios y especialmente en universidades; para analizar el rendimiento académico estudiantil y en función de los resultados tomar decisiones que beneficien a los estudiantes y a las instituciones.

Los antecedentes también nos permitirán establecer que en la actualidad existe una elevada preocupación en torno al bajo nivel de aprovechamiento estudiantil reflejado en altos índices de mortalidad académica, deserción, entre otros y los factores que pueden influir en este, por lo que varios centros educativos muestran especial interés en investigar este tema con el fin de establecer políticas institucionales que permitan actuar de manera preventiva frente a situaciones que puedan afectar el rendimiento académico de un estudiante y no de una manera recuperadora como sucede actualmente.

El rendimiento académico es un factor que se ve influenciado tanto por las condiciones internas de las instituciones educativas, como por la labor docente llevada a cabo por cada uno de sus maestros, sin embargo no se puede desconocer cómo las características propias de los estudiantes influyen de manera positiva o negativa en el rendimiento académico; esto se puede decir por lo abordado en los diferentes documentos relacionados aquí y también por observación natural de estas situaciones. Por lo tanto existen muchos factores que intervienen o condicionan este rendimiento, de ahí que este trabajo centre su interés en determinar las características relacionadas entre las condiciones de vida, situación socioeconómica, características demográficas, antecedentes

escolares y el rendimiento académico de los estudiantes de pregrado de la Universidad de Nariño, definiendo en esta investigación, el rendimiento académico como el promedio acumulado resultado de las valoraciones cuantitativas emitidas por semestre académico, la razón de que se utilice el promedio acumulado como variable para medir el rendimiento académico es porque en la base de datos de la Universidad de Nariño es la única que da cuenta del rendimiento de manera global.

Para orientar este trabajo de minería de datos, se emplea una metodología llamada CRISP-DM (Cross Industry Standard Process for Data Mining). Nació en 1996 por la necesidad de organizar los procesos de minería para ser utilizados de forma industrial y no tanto al ámbito académico. Este proyecto recibió financiación de la Unión Europea en 1997 y fue creado por las empresas SPSS, Teradata, Daimler AG, NCR Corporation y OHRA (Gallardo Arancibia, 2009). La razón por la que se utiliza esta metodología se debe a que CRISP-DM es la más conocida y es la que más se referencia en los artículos consignados en los **ANTECEDENTES DEL PROBLEMA**.

La idea general de la metodología CRISP-DM es estandarizar todo el proceso de minería de datos a través de niveles y etapas que permiten la diferenciación de un paso o momento de otro en la minería.

Este informe de trabajo de grado ha sido estructurado en 6 partes, la primera es la definición del problema; la segunda es la justificación, tercera los objetivos general y específicos, la cuarta es el diseño metodológico en el que se describen aspectos como la población y muestra, tipo de diseño y variables del problema, la quinta es el marco de referencia en el que se describe los procesos de minería de datos que se siguieron para resolver el problema y se acompaña con la parte teórica que sustenta este trabajo, la sexta es el análisis de los resultados, conclusiones, recomendaciones, la séptima es una bibliografía y finalmente la sección de anexos.

1. DEFINICIÓN DEL PROBLEMA

1.1 ANTECEDENTES DEL PROBLEMA

1.1.1 Antecedentes Internacionales

Título: Academic Performance of University Students and its Relation with Employment

Autores: Laura Lanzarini, María Emilia Charbelli, Javier Díaz

Fecha: 2015

Fuente: (Lanzarini, Charnelli, y Javier, 2015)

Objetivo:

Identificar las características más relevantes en relación al rendimiento académico de los estudiantes en la Escuela de Ciencias Computacionales de la Universidad Nacional de la Plata.

Resumen:

La Minería de Datos Académica recoge varios métodos que permiten extraer información útil y novedosa de grandes volúmenes de datos en el contexto educativo. Este artículo describe el proceso usado para, a través de técnicas de minería de datos; identificar las características más relevantes en relación al rendimiento académico en la escuela de Ciencias Computacionales de la Universidad Nacional de la Plata. Los resultados obtenidos usando los métodos propuestos para procesar la información relacionada con estudiantes regulares y no regulares en la UNLP permitió el establecimiento de relaciones interesantes en función del rendimiento académico del estudiante. Basándose en los modelos obtenidos se puede decir que el hecho de que el estudiante trabaje no significa que su rendimiento académico se reduzca y que los estudiantes que toman varios años para unirse a la facultad tienen mejor rendimiento si ellos manifiestan un interés por conseguir trabajo.

En la tabla 1 se presenta el análisis crítico del artículo Academic Performance of University Students and its Relation with Employment.

Tabla 1: Antecedente uno

Modelo presentado en este artículo	Modelo presentado en este trabajo de maestría
Su objeto de estudio son los 5268 estudiantes de la Facultad de Informática de la UNLP comprendidos en el periodo 2002 y 2012.	El objeto de estudio en este trabajo son los 10199 estudiantes de pregrado de la Universidad de Nariño entre los años 2010 y 2014
Reducir la dimensión de la información a considerar, identificando los atributos que mejor caracterizan el avance académico de un alumno.	En este trabajo se trata de conservar las dimensiones de la base de datos ya que interesa conocer qué variables explican mejor el rendimiento académico
En este trabajo se trata de identificar las características más relevantes en relación al rendimiento académico de los estudiantes en la Escuela de Ciencias Computacionales de la Universidad Nacional de la Plata.	En este trabajo se trata de identificar las características que inciden en el rendimiento académico
La información se toma a partir de datos personales como laborales recolectados del sistema SIU-Guaraní.	La información se toma a partir de los datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la Universidad cuando el estudiante se matricula al primer semestre
La técnica que emplea este trabajo para la selección de características son los métodos de filtros y los métodos wrappers cuyo clasificador se basó en una Máquina de Vectores. Para establecer los rankings entre los atributos utiliza el método Chi-cuadrado. Una vez seleccionados los atributos utilizaron tres modelos diferentes para que permitan clasificar alumnos, estos fueron C4.5, PART y un multiperceptron entrenado con el algoritmo de backpropagation. Los dos primeros se utilizan para obtener un modelo descriptivo, mientras que el tercero permite verificar la precisión obtenida.	En este trabajo, se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCOPUS
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”

- Dirección (URL): http://ieeexplore.ieee.org.ezproxy.utp.edu.co/document/7360017/

Título: A Data Mining Approach to Guide Students Through The Enrollment Process Based on Academic Performance

Autores: Cesar Vialardi, Jorge Chue, Juan Pablo Peche, Gustavo Alvarado, Bruno Vinatea, Jhonny Estrella, Álvaro Ortigosa

Fecha: 2011

Fuente: (Vialardi et al., 2011)

Objetivo: Proponer una metodología para desarrollar un sistema de recomendación (CRSs) y usar motores de recomendación basados en minería de datos para sugerir, de manera inteligente; acciones a los estudiantes basados en las decisiones de otras personas anteriores con características académicas, demográficas y personales similares en la universidad de Lima (Perú).

Resumen:

El rendimiento académico estudiantil en universidades es crucial para los sistemas de administración académicos. Muchas acciones y decisiones son hechas basadas en un sistema como tal, específicamente en el proceso de matrícula. Durante el proceso de matrícula, los estudiantes tienen que decidir a cuáles materias se desean inscribir. Esta investigación introduce la lógica detrás del diseño de un sistema de recomendación para dar soporte en el proceso de matrícula usando los registros del rendimiento académico de los estudiantes. Para construir este sistema, la metodología CRISP-DM fue aplicada a los datos de los estudiantes del departamento de ingeniería de sistemas de la universidad de Lima, Perú. Una de las principales contribuciones de este trabajo es el uso de dos atributos sintéticos para mejorar la relevancia de las recomendaciones hechas. El primer atributo estima la dificultad inherente de un curso dado. El segundo atributo, llamado Potencial; es la medida de la aptitud de un estudiante para un curso en particular basado en las notas obtenidas en cursos relacionados. Los datos fueron minados usando los algoritmos C4.5, KNN (K-nearest neighbour), Naïve Bayes, Bagging y Boosting, y se desarrolló un conjunto de experimientos para determinar el mejor algoritmo para el dominio de esta aplicación. Los resultados muestran que Bagging es el mejor método con respecto a la precisión de la predicción. Basándose en estos resultados, el sistema de Recomendación del Rendimiento Estudiantil (SPRS (Student Performance Recommender System)) fue desarrollado, incluyendo un motor de aprendizaje. El SPRS se probó con una

muestra de 39 estudiantes durante el proceso de matrícula. Los resultados muestran que el sistema tuvo un buen rendimiento bajo condiciones reales.

En la tabla 2 se presenta el análisis crítico del artículo A Data Mining Approach to Guide Students Through The Enrollment Process Based on Academic Performance.

Tabla 2: Antecedente dos

Modelo presentado en este artículo	**Modelo propuesto en el trabajo de maestría**
Contempla la construcción de un sistema de recomendación basado en minería de datos para ser usado en el departamento de ingeniería de sistemas de la universidad de Lima	Al final del trabajo, se plantea una serie de sugerencias y recomendaciones a la universidad y ella es responsable de aplicarlas
El trabajo investigativo se enfocó en los estudiantes del departamento de ingeniería de sistemas de la universidad de Lima (Perú)	Se enfoca en los 10199 estudiantes de pregrado de la Universidad de Nariño entre los años 2010 y 2014
Sugiere que el diseño del sistema de recomendación puede ser construido para estudiantes con clases presenciales y para estudiantes con clases virtuales	Se enfoca únicamente en estudiantes de jornada presencial del pregrado
En la base de datos se utiliza las siguientes variables: nombre del curso, número de intentos, promedio académico acumulado, número de créditos por curso, número de créditos y dos variables sintéticas: dificultad en cada curso y el potencial de cada estudiante para cursarlo.	En la base de datos del modelo propuesto se pretende utilizar 44 variables personales, socioeconómicas, académicas e institucionales que se recogen cuando el estudiante se matricula al primer semestre
Se contempla el uso de las siguientes técnicas de clasificación en minería de datos: KNN (K-Nearest neighbour), Bayes Naïve, C4.5, Boosting, Bagging. Al final Bagging es seleccionada como la mejor luego de una serie de experimentos	En este trabajo, se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCOPUS
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”

- Dirección (URL)
  http://link.springer.com.ezproxy.utp.edu.co/article/10.1007%2Fs11257-011-9098-4

Título: Predicting Student Success Based on Prior Performance

Autores: Ahmad Slim, Gregory Heileman, Jarred Kozlick, Abdallah Chaouki

Fecha: 2014

Fuente: (Slim et al., 2015)

Objetivo: Proponer un modelo gráfico probabilístico que permita razonar sobre el rendimiento y progreso de un estudiante usando redes de creencia Bayesianas (Bayesian Belief Network (BBN))

Resumen:

Las instituciones de Educación superior están interesadas cada vez más en llevar un control del progreso de sus estudiantes mientras monitorean y trabajan para mejorar las tasas de retención y graduación. Idealmente indicadores tempranos del progreso estudiantil o en el caso de que no existan se pueden usar para proveer intervenciones apropiadas que incrementen la probabilidad de éxito de los estudiantes. En este documento presentamos un modelo de trabajo que usa aprendizaje automático y en particular la técnica Bayesian Belief Network (BBN), para predecir el rendimiento de estudiantes en la etapa temprana de sus carreras académicas. Los resultados obtenidos muestran que el modelo de trabajo propuesto puede predecir el progreso estudiantil, específicamente el promedio de la nota (GPA) dentro de la carrera, con error mínimo después de haber observado un solo semestre de rendimiento. Además a medida en que más rendimiento adicional es observado, la nota predicha (GPA) se convierte en incrementalmente precisa en los semestres subsiguientes, para proveer consejo a los estudiantes respecto a la probabilidad de éxito en una etapa temprana de sus carreras académicas.

En la tabla 3 se presenta el análisis crítico del artículo Predicting Student Success Based on Prior Performance.

Tabla 3: Antecedente tres

Modelo presentado en este artículo	**Modelo presentado en este trabajo de maestría**

Su objeto de estudio son los estudiantes de diferentes programas de la Universidad de Nuevo México.	El objeto de estudio en este trabajo son los 10199 estudiantes de pregrado de la Universidad de Nariño entre los años 2010 y 2014
Este documento aplica la técnica BBN para crear un modelo gráfico probabilístico que permita razonar sobre el rendimiento y progreso de un estudiante	En este trabajo para determinar las características, se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación
Este trabajo contempla el rendimiento académico medido de la siguiente forma investigar las propiedades estructurales de cada curriculum teniendo en cuenta el grado que los cursos particulares de un currículo dado puedan impactar en el progreso del estudiante y también las notas que los estudiantes obtienen. De esta forma se mide el rendimiento académico.	Este trabajo mide el rendimiento académico del estudiante utilizando para ello el promedio acumulado hasta el segundo semestre del año 2014 de cohortes completas
El artículo utiliza notas anteriores de los estudiantes para inferir algunas características respecto al rendimiento futuro de los estudiantes.	Este trabajo tiene en cuenta el puntaje ICFES al momento del ingreso a la universidad

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCOPUS
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”
- Dirección (URL) http://ieeexplore.ieee.org.ezproxy.utp.edu.co/document/7008697/

Título: Comments Data Mining for Evaluating Student’s Performance

Autores: Shaymaa E Sorour, Tsunenori Mine, Kazumasa Goda, Sachio Hirokawax

Fecha: 2014

Fuente: (Sorour et al., 2014)

Objetivo:

Proveer retroalimentación individual a los estudiantes para mejorar sus actividades de aprendizaje.

Resumen:

El presente estudio propone métodos de predicción de la nota de un estudiante basado en datos de comentarios. Los estudiantes describen sus actitudes de aprendizaje, tendencias y comportamientos escribiendo sus comentarios libremente después de cada clase. La principal dificultad de esta investigación es predecir el rendimiento de los estudiantes usando separadamente datos de comentario en dos clases. Aunque los estudiantes aprenden la misma materia, existen diferencias entre los comentarios de las dos clases. Los métodos propuestos básicamente emplean análisis semántico latente (LSA Latent Semantic Analysis) y dos técnicas de aprendizaje automático Máquinas de Soporte Vectorial y Redes Neuronales Artificiales para predecir los resultados finales de los estudiantes en cuatro notas S, A, B y C. Adicionalmente un método de solapamiento se propuso para mejorar la precisión de los resultados predichos, el método permite aceptar dos calificaciones para una nota para obtener la relación correcta entre los resultados LSA y las notas de los estudiantes. Los métodos propuestos logran 50.7% y 48.7% de precisión en la predicción de las notas de los estudiantes para Máquinas de Soporte Vectorial y Redes Neuronales respectivamente. Para este propósito los resultados de este estudio reportaron modelos de rendimiento académico de los estudiantes como predictores que son recursos útiles para el entendimiento del comportamiento de los estudiantes y proveen retroalimentación a ellos de forma que podemos mejorar sus actividades de aprendizaje.

En la tabla 4 se presenta el análisis crítico del artículo Comments Data Mining for Evaluating Student's Performance.

Tabla 4: Antecedente cuatro

Modelo presentado en este artículo	**Modelo presentado en este trabajo de maestría**
El trabajo contempla usar minería de texto, para predecir las notas de los estudiantes usando los comentarios tipo C del método PCN. En el método PCN el comentario tipo C se refiere al entendimiento y actitudes de aprendizaje de la clase.	Este trabajo pretende usar técnicas de minería de datos para identificar las características que inciden en el rendimiento académico
En el trabajo se usan datos de dos	La información se toma a partir de los

grupos, clase A(60 estudiantes) y clase B(63 estudiantes), usando una clase como entrenamiento y la otra como prueba. Los estudiantes aprenden lo mismo en ambas clases pero hay diferencias en los comentarios para cada clase.	datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la Universidad cuando el estudiante se matricula al primer semestre
El trabajo usa las técnicas de redes neuronales, máquinas de soporte vectoriales para predecir la nota del estudiante.	En este trabajo para determinar las características, se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación
El método propone una técnica de solapamiento que toma las notas adyacentes y conduce experimentos para validar ese método calculando el estadístico F y la precisión de la nota final calculada.	Este trabajo utiliza técnicas de minería de datos existentes

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCOPUS
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”
- Dirección (URL) http://ieeexplore.ieee.org.ezproxy.utp.edu.co/document/6913261/

Título: To What Extend Can We Predict Students’ Performance? A case Study in Colleges in South Africa

Autores: Norman Poh, Ian Smythe

Fecha: 2014

Fuente: (Poh y Smythe, 2015)

Objetivo:

En este estudio nos gustaría investigar la posibilidad de predecir el rendimiento de un estudiante en el contexto del uso de una plataforma de evaluación online.

Resumen:

El rendimiento estudiantil depende de factores además de la simple habilidad, como el ambiente, estatus socioeconómico, contexto familiar y personal. Capturar estos patrones de influencia puede habilitar a un educador a mejorar algunos de estos factores o para los gobiernos ajustar su política social acordemente. Para entender estos factores, hemos ejecutado el ejercicio de predecir el rendimiento estudiantil, usando una cohorte de aproximadamente 8000 estudiantes de educación superior Sud Africanos. Todos tomaron una prueba de inglés y matemáticas. Mostramos que es posible predecir los resultados de las pruebas de comprensión en inglés (1) de otros resultados de la prueba; (2) de co-variables sobre autoeficacia, estatus socioeconómico y dificultades específicas de aprendizaje hay 100 preguntas en total; (3) de otros resultados + co-variables (combinación de (1) y (2)); y de (4) un modelo más avanzado similar al (3) excepto que las co-variables son sujeto de reducción de dimensionalidad a través de PCA. Los modelos de 1-4 pueden predecir el rendimiento estudiantil hasta un error estándar del 13 al 15%. En comparación a una suposición aleatoria tendría un error del 17%. En pocas palabras es posible predecir condicionalmente el rendimiento del estudiante basado en la autoeficacia, entorno socioeconómico, dificultades de aprendizaje y resultados de pruebas académicas relacionadas.

En la tabla 5 se presenta el análisis crítico del artículo To What Extend Can We Predict Students' Performance? A case Study in Colleges in South Africa.

Tabla 5: Antecedente cinco

Modelo presentado en este artículo	**Modelo presentado en este trabajo de Maestría**
Su objeto de estudio son los estudiantes de una cohorte de aproximadamente 8,000 estudiantes universitarios sudafricanos.	El objeto de estudio en este trabajo son los 10199 estudiantes de pregrado de la Universidad de Nariño de los años 2010 y 2014
Los autores investigan la posibilidad de predecir el rendimiento de un estudiante en el contexto del uso de una plataforma de evaluación online.	Este trabajo pretende identificar las características que inciden en el rendimiento académico a partir de datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la universidad cuando el estudiante se matricula al primer semestre
En el trabajo se contempla el uso variables predictoras que son autoeficacia, optimismo.	En la base de datos se pretende utilizar 44 variables personales, socioeconómicas, académicas e institucionales

En este trabajo se plantea el uso de modelos de regresión con regularización para un data set de 100 variables.	Se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación
Este trabajo utiliza cuatro modelos de regresión, los dos primeros usan algunas de las 100 variables, los modelos 3 y 4 las usan conjuntamente. El modelo 4 se le aplica reducción de dimensionalidad a través del análisis de componentes principales.	En este trabajo se pretende utilizar todas las variables con técnicas predictivas y clustering
En el trabajo se usan una serie de 8 pruebas con respuesta de tipo selección múltiple para poder tener datos extra de los estudiantes.	En este trabajo solo se cuenta con la información que se tomara a partir de los datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la Universidad

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: IEEE Explore
- Frase Lógica utilizada: Predictive models AND student performance
- Dirección (URL): http://ieeexplore.ieee.org/document/7008698/?part=1

Título: Extracting Relationships Between Students' Academic Performance and Their Area of Interest Using Data Mining Techniques

Autores: Ana Ktona, Denada Xhaja, Ilia Ninka

Fecha: 2014

Fuente: (Ktona, Xhaja, y Ninka, 2014)

Objetivo: Encontrar reglas de clasificación entre el rendimiento académico del estudiante y el programa al que ellos desean asistir, así como también particionar los estudiantes en grupos de acuerdo a sus características como es el rendimiento académico.

Resumen: La minería de datos es un campo de las ciencias computacionales que combina herramientas de inteligencia artificial y estadística con el manejo de las bases de datos. La minería de datos puede ser usada en varios campos de la vida real y una de las áreas donde es aplicada y presentada en este documento es en

la educación. Los hallazgos provistos por el uso de minería de datos en educación pueden ayudar en la mejora de la calidad de la educación. En este estudio aplicamos técnicas de minería de datos para encontrar reglas de clasificación entre el rendimiento académico del estudiante y el programa al que ellos quieren asistir, así como también particionar estudiantes en grupos de acuerdo a sus características como es el rendimiento académico. La Extracción de reglas de clasificación y el agrupamiento son ejecutados usando los algoritmos C4.5 para árboles de decisión y k-means. Los resultados de ambas técnicas sugieren ayudar a los estudiantes a enfocarse en el área en la que ellos están interesados.

En la tabla 6 se presente el análisis crítico del artículo Extracting Relationships Between Students' Academic Performance and Their Area of Interest Using Data Mining Techniques.

Tabla 6: Antecedente seis

Modelo presentado en este artículo	**Modelo presentado en este trabajo de Maestría.**
El objeto de estudio en este trabajo son los 277 estudiantes que cursan el segundo y tercer año de las ramas de Informática y Tecnologías de la Información y la Comunicación de la Facultad de Ciencias Naturales	El objeto de estudio en este trabajo son los 10199 estudiantes de pregrado de la Universidad de Nariño de los años 2010 y 2014
Este trabajo utiliza las técnicas k-means y el algoritmo C4.5 de árboles de decisión.	Se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación
Los datos utilizados para los fines de este estudio se recogen mediante un cuestionario cerrado de 25 preguntas	La información se toma a partir de los datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la Universidad cuando el estudiante se matricula al primer semestre
En este estudio se aplicó la minería de datos sobre los datos de los estudiantes con dos propósitos: extraer reglas de clasificación entre el rendimiento académico de los estudiantes y el programa al que ellos quieren asistir.	Este trabajo pretende identificar las características que inciden en el rendimiento académico

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCOPUS
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”
- Dirección (URL): http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7059136

Título: Mapping Student’s Performance Based on Data Mining Approach (A Case Study)

Autores: Harwati, Ardita Permata Alfiani, Febriana Ayu Wulandari

Fecha: 2014

Fuente: (Harwati, Alfiani, y Wulandari, 2015)

Objetivo: El propósito de este estudio es examinar el patrón del rendimiento estudiantil en el Departamento de Ingeniería Industrial de la Universidad de Indonesia usando la técnica de agrupamiento k-means.

Resumen: El mejoramiento del rendimiento estudiantil es un foco importante en la administración de instituciones de educación superior. El mapeo de la condición real de un estudiante es el requerimiento que se debe hacer antes de diseñar el programa de mejoramiento académico. Este estudio se enfoca en mapear los estudiantes usando los algoritmos de agrupación k-means para revelar patrones escondidos y clasificar estudiantes basados en sus datos demográficos (género, origen, GPA, Grado de conocimiento de materias), y el promedio de la asistencia al curso. Se cubrió datos de cerca de 300 estudiantes. Desde el cálculo usando SPSS 16, se encuentra que hay cuatro grupos formados basados en seis variables: estudiantes inteligentes (45.74%), estudiantes estándar (33.33%) y 20.92% pertenecen a los estudiantes de bajo rendimiento.

En la tabla 7 se presenta el análisis crítico del artículo Mapping Student’s Performance Based on Data Mining Approach (A Case Study).

Tabla 7: Antecedente siete

Modelo presentado en este artículo	**Modelo presentado en este trabajo de Maestría.**
El objeto de estudio en este trabajo son cerca de 300	El objeto de estudio en este trabajo son los 10199 estudiantes de pregrado de la

estudiantes del departamento de Tecnología Industrial de la Universidad de Indonesia.	Universidad de Nariño de los años 2010 y 2014
Este trabajo revela patrones escondidos y clasifica estudiantes basados en sus datos demográficos (género, origen, GPA, grado de conocimiento de materias), y el promedio de la asistencia al curso	Este trabajo pretende identificar las características que inciden en el rendimiento académico a partir de los datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la universidad cuando el estudiante se matricula al primer semestre
Este trabajo utiliza los algoritmos de agrupación k-means.	Se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación
El trabajo contempla la agrupación de estudiantes de acuerdo a sus características en las diferentes variables y el descubrimiento de patrones.	Este trabajo contempla la agrupación de estudiantes de acuerdo a las características de las diferentes variables disponibles y el señalamiento de las variables que influyen en el rendimiento académico.

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCOPUS
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”
- Dirección (URL): http://dx.doi.org/10.1016/j.aaspro.2015.01.034

Título: Predicción del rendimiento académico de alumnos de primer año de la FACENA(UNNE) en función de su caracterización socioeducativa

Autores: Eduardo Adolfo Porcel, Gladys Noemí Dapozo, María Victoria López

Fecha: 2010

Fuente: (Porcel, Eduardo Adolfo; Dapozo y Lopéz, 2010)

Objetivo:

Predecir el rendimiento académico de los alumnos de primer año de la FACENA (UNNE), en función de sus características socioeducativas empleando la técnica de regresión logística.

Resumen:

En este trabajo se analiza la relación del rendimiento académico de los alumnos ingresantes a la Facultad de Ciencias Exactas y Naturales y Agrimensura de la Universidad Nacional del Nordeste (FACENA-UNNE) en Corrientes, Argentina, durante el primer año de carrera con las características socioeducativas de los mismos. El rendimiento fue medido por la aprobación de los exámenes parciales o finales de la primera materia de Matemática que los alumnos cursan. Se ajustó un modelo de regresión Logística binaria, el cual clasifico adecuadamente el 75% de los datos. Entre las variables más relevantes para explicar el rendimiento académico se encuentran el título secundario, la carrera elegida y el nivel educacional alcanzado por la madre.

En la tabla 8 se presenta el análisis crítico del artículo Predicción del rendimiento académico de alumnos de primer año de la FACENA(UNNE) en función de su caracterización socioeducativa.

Tabla 8: Antecedente ocho

Modelo presentado en este artículo	**Modelo presentado en este trabajo de Maestría**
Su objeto de estudio estuvo integrada por los alumnos que ingresaron a la Facultad en los años 2004 y 2005.	El objeto de estudio en este trabajo son los estudiantes de pregrado de la Universidad de Nariño de los años 2010 y 2014
Este trabajo pretende predecir el rendimiento académico de los alumnos de primer año de la FACENA, en función de sus características socioeducativas, registradas en el formulario de ingreso a la universidad.	Este trabajo pretende identificar las características que inciden en el rendimiento académico a partir de datos socioeconómicos, académicos e institucionales que se recogen en la base de datos de la universidad cuando el estudiante se matricula al primer semestre
En el trabajo se contempla el uso 8 variables cualitativas predictoras.	En este trabajo se pretende utilizar tanto variables cuantitativas como cualitativas, que se encuentran en la base de datos, para la predicción y la clasificación.
En este trabajo se formuló un modelo de regresión logística binaria de efectos principales con un nivel de significancia $\alpha = 0.05$, debido a que la variable dependiente es de tipo cualitativa y asume dos categorías.	Se pretende emplear el mejor modelo que se ajuste a los datos tanto cualitativos como cuantitativos con los que cuenta la base de datos para la técnica predictiva y además clustering para la clasificación.
La información sobre las características	En este trabajo solo se cuenta con la

socioeducativas y el desempeño académico fue obtenida del sistema informático de gestión de alumnos de la unidad académica. Los datos sobre las características socioeconómicas se recogen en un formulario al ingreso a la universidad.	información que se tomara a partir de los datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la Universidad, cuando el estudiante se matricula a primer semestre.
Los datos relacionados con el rendimiento académico provienen de los informes que los docentes de cada asignatura realizan al finalizar cada dictado y solo se registra como la aprobación o no de los exámenes parciales o finales. En este estudio solo se tiene en cuenta el rendimiento académico asociado con la asignatura de Matemática, ya que todas las carreras tienen en el primer cuatrimestre del primer año esta materia.	En cuanto a los datos relacionados con el rendimiento académico para este caso se toma el promedio acumulado, que se registra hasta el momento de la generación de los datos.

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCIELO
- Frase Lógica utilizada: “Predicción del rendimiento académico”
- Dirección (URL): https://redie.uabc.mx/redie/article/view/264/730

Título: Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos

Autores: R. Alcover, J. Benlloch, P. Blesa, M. A. Claduch, M. Celma, C. Ferri, J. Hernandez-Orallo, L. Iniesta, J. Más, M. J. Ramirez-Quintana, A. Robles, J. M. Valiente, M. J. Vicent, L. R. Zúnica

Fecha: 2007

Fuente: (Alcover et al., s. f.)

Objetivo:

Aplicar técnicas de minería de datos para analizar la influencia de los parámetros (socioeconómicos, características personales, nota de entrada …) más relevantes sobre el rendimiento académico de un alumno de primer curso en las titulaciones de informática de la UPV.

Resumen:

En este trabajo presentamos un análisis del rendimiento académico de los alumnos de nuevo ingreso en la titulación de Ingeniería Técnica en Informática de Sistemas de la Universidad Politécnica de Valencia (UPV) a lo largo de tres cursos, aunque también se ha trabajado con las titulaciones de Ingeniería Técnica en Informática de Gestión y de Ingeniería de Informática.
Este análisis relaciona el rendimiento con las características socioeconómicas y académicas de los alumnos, que se obtienen en el momento de la matrícula, y que se recogen en la base de datos de la universidad. Hemos definido un indicador del rendimiento para cada alumno, teniendo en cuenta las calificaciones obtenidas y las convocatorias utilizadas.
Para el estudio utilizamos técnicas de minería de datos, que pretenden determinar qué nivel de condicionamiento existe entre dicho rendimiento y características como el nivel de conocimientos de entrada del alumno, su contexto geográfico y sociocultural, etc… esto proporciona una herramienta importante para la acción tutorial, que puede apoyarse en las predicciones de los modelos que se obtienen para encauzar sus recomendaciones y encuadrar las expectativas y el esfuerzo necesario para cada alumno, lógicamente dentro de la cautela habitual a la hora de tratar modelos inferidos a partir de datos.

En la tabla 9 se presenta el análisis crítico del artículo Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos.

Tabla 9: Antecedente nueve

Modelo presentado en este artículo	**Modelo presentado en este trabajo de Maestría**
Su objeto de estudio estuvo conformado por todos los alumnos de nuevo ingreso en cualquiera de las tres titulaciones de informática de la UPV, 569 alumnos de Ingeniería en Informática, 646 alumnos Ingeniería Técnica en Informática de Gestión y 572 alumnos de Técnica en Informática de Sistemas.	El objeto de estudio en este trabajo son los estudiantes de pregrado de la Universidad de Nariño de los años 2010 y 2014
Este trabajo pretende predecir el rendimiento académico disponiendo únicamente de la información aportada por el alumno en el momento de su	Este trabajo pretende identificar las características que inciden en el rendimiento académico a partir de datos socioeconómicos, académicos e

matrícula.	institucionales que se recogen en la base de datos de la universidad cuando el estudiante se matricula al primer semestre
La información se toma a partir de la base de datos con los que cuenta la universidad, y que relaciona únicamente datos anteriores a su entrada a esta.	La información se toma a partir de los datos socioeconómicos, académicos e institucionales que se recogen en la base de datos de la Universidad cuando el estudiante se matricula al primer semestre
Para generar los modelos predictivos del rendimiento académico los investigadores utilizaron árboles de decisión y la regresión multivariante.	En este trabajo, se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la agrupación.
En este trabajo el rendimiento académico se calcula teniendo en cuenta entre otras variables, la convocatoria en que el alumno supera la asignatura, la calificación numérica que obtuvo el alumno en una asignatura j, el número de créditos que figura en la asignatura j; determinándose por tanto el rendimiento como un atributo que toma valores entre 0 y 100.	En cuanto a los datos relacionados con el rendimiento académico para este caso se toma el promedio acumulado, que se registra hasta el momento de la generación de los datos.

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: bioinfo
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”
- Dirección (URL): http://bioinfo.uib.es/~joemiro/aenui/procJenui/Jen2007/alanal.pdf

1.1.2 Antecedentes Nacionales

Título: Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance

Autores: S. M. Merchan, J. A. Duarte

Fecha: 2016

Fuente: (Merchán y Duarte, 2016)

Objetivo: Generar un modelo predictivo de desempeño académico basado en los datos académicos y demográficos de los estudiantes.

Resumen:

Este documento presenta y analiza la experiencia de aplicar ciertos métodos y técnicas de Minería de Datos en la base de datos de los estudiantes de Ingeniería de Sistemas de la Universidad del Bosque Colombia. Esto con el fin de construir un modelo predictivo del rendimiento académico de los estudiantes. Trabajos previos fueron tomados en cuenta relacionados con construcción de modelos predictivos en el ámbito de ambientes académicos usando arboles de decisión, redes neuronales artificiales y otras técnicas de clasificación. Dado que esto es un proceso iterativo de descubrimiento y aprendizaje, la experiencia es analizada de acuerdo a los resultados obtenidos en cada iteración del proceso. Cada resultado obtenido es evaluado sin importar los resultados que se esperan, la caracterización de los datos de entrada y salida, aquello que la teoría dicta y la importancia del modelo obtenido en términos de la precisión de la predicción. La importancia mencionada se evalúa teniendo en cuenta detalles particulares de la población en estudio y las necesidades específicas manifestadas por la institución, por ejemplo el acompañamiento de los estudiantes a lo largo de su proceso de aprendizaje, y la toma de decisiones basadas en el tiempo para prevenir deserción y riesgo académico. Finalmente algunas recomendaciones y pensamientos son puestos para un desarrollo futuro de este trabajo y para otros investigadores trabajando en estudios similares.

En la tabla 10 se presenta el análisis crítico del artículo Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance.

Tabla 10: Antecedente diez

Modelo presentado en este artículo	**Modelo presentado en este trabajo de maestría**
Su objeto de estudio son 932 estudiantes de Ingeniería de sistemas de la Universidad del Bosque.	El objeto de estudio en este trabajo son los 10199 estudiantes de pregrado de la Universidad de Nariño entre los años 2010 y 2014
Los investigadores plantean generar un modelo predictivo de desempeño académico que actué preventivamente y que sea capaz de identificar las causa de riesgo académico y actuar sobre estas causas antes de que el riesgo	Este trabajo trata de detectar las características que inciden en el rendimiento académico usando las variables personales, socioeconómicas, académicas e institucionales utilizando modelos

ocurra.	descriptivos y clustering, y elaborar recomendaciones sobre los posibles usos de los resultados obtenidos que podrían servir como estrategias para la gestión académica de la Universidad.
La información se toma a partir de datos académicos y demográficos de los estudiantes.	La información se toma a partir de los datos socioeconómicos, académicos e institucionales que se recogen en la base de datos de la Universidad cuando el estudiante se matricula al primer semestre
Para el trabajo de minería de datos estos investigadores utilizaron los algoritmos J48, PART y Ridor de WEKA debido a su similitud en el propósito de inducción de reglas de clasificación.	En este trabajo, se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: SCOPUS
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”
- Dirección (URL) http://ieeexplore.ieee.org.ezproxy.utp.edu.co/document/7555255/

Título: Modelos de predicción del rendimiento académico en matemáticas I en la Universidad Tecnológica de Pereira

Autores: Patricia Carvajal Olaya, Julio Cesar Mosquera, Irina Artamonava

Fecha: 2009

Fuente: (Carvajal Olaya, Mosquera, y Artamonova, 2009)

Objetivo:

Determinar los factores que influyen de manera significativa sobre el rendimiento académico de los estudiantes de ingeniería y tecnologías de la Universidad Tecnológica de Pereira en la asignatura Matemáticas I.

Resumen:

En este trabajo se presentan los resultados sobre el estudio de los factores que influyen de manera significativa sobre el rendimiento académico de los estudiantes de ingeniería y tecnologías de la Universidad Tecnológica de Pereira en la asignatura Matemáticas I. Se propone un modelo de regresión logística múltiple que emplea las variables más relevantes halladas durante la investigación y que afectan el rendimiento de los estudiantes en la asignatura Matemáticas I. Se demuestra que, a partir de esta información, es posible predecir con una aceptable confiabilidad el rendimiento de un alumno dado. Como factores determinantes del rendimiento de los nuevos alumnos en la asignatura Matemáticas I se hallaron: El puntaje del examen ICFES, el nivel de lectura literal y el nivel de razonamiento lógico abstracto.

En la tabla 11 se presenta el análisis crítico del artículo Modelos de predicción del rendimiento académico en matemáticas I en la Universidad Tecnológica de Pereira.

Tabla 11: Antecedente once

Modelo presentado en este artículo	**Modelo presentado en este trabajo de Maestría.**
El objeto de estudio de este trabajo estuvo constituido por estudiantes tanto de ingenierías como de tecnologías de la Universidad Tecnológica de Pereira, que cursaron la asignatura de Matemática I en el primer semestre del año 2008.	El objeto de estudio en este trabajo son los estudiantes de pregrado de la Universidad de Nariño de los años 2010 y 2014
Este trabajo pretende determinar los factores que influyen de manera significativa sobre el rendimiento académico de los estudiantes de ingeniería y tecnologías de la Universidad Tecnológica de Pereira en la asignatura Matemáticas I, a partir de factores personales, socioeconómicos, académicos, institucionales y de riesgo que se encuentran en la base de datos de la universidad.	Este trabajo pretende identificar las características que inciden en el rendimiento académico a partir de datos socioeconómicos, académicos e institucionales que se recogen en la base de datos de la universidad cuando el estudiante se matricula al primer semestre
Este estudio empleo un modelo de regresión logística múltiple con una muestra aleatoria de 582 estudiantes que cursaron la asignatura de matemática I en el primer semestre del año 2008, con el fin de valorar los posibles factores que	Se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación

causaron diferencias entre la aprobación de esta asignatura en las diferentes tecnologías e ingenierías. Además con una muestra de 630 estudiantes y las mismas variables independientes los autores crean un árbol de clasificación para el rendimiento en matemáticas I utilizando el algoritmo CHAID.	
Este trabajo toma el rendimiento académico de la asignatura Matemáticas I como aprobada o reprobada.	En cuanto a los datos relacionados con el rendimiento académico para este caso se toma el promedio acumulado, que se registra hasta el momento de la generación de los datos.

Fuente: este trabajo

Criterio de Búsqueda

- Metabuscador: Scientia et technica
- Frase Lógica utilizada: “Academic Performance” AND “Data Mining”
- Dirección (URL): http://revistas.utp.edu.co/index.php/revistaciencia/article/download/2323/1237

1.1.3 Antecedentes Regionales

Título: Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos

Autores: Ricardo Timarán Pereira

Fecha: 2009

Fuente: (Pereira, 2009)

Objetivo:

Determinar en la Universidad de Nariño perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de descubrimiento de conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años.

Resumen:

En este artículo se presentan los resultados de la investigación realizada en la Universidad de Nariño (Colombia) cuyo objetivo fue determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de descubrimiento de conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años. Este proceso se apoyó con TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios de DCBD del Departamento de Ingeniería.

En la tabla 12 se presenta el análisis crítico del artículo Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos.

Tabla 12: Antecedente doce

Modelo presentado en este artículo	**Modelo presentado en este trabajo de Maestría.**
El objeto de estudio de este trabajo son los estudiantes de pregrado de la Universidad de Nariño en un periodo de 15 años.	El objeto de estudio en este trabajo son los estudiantes de pregrado de la Universidad de Nariño de los años 2010 y 2014
En este trabajo se pretende identificar perfiles de bajo rendimiento académico y deserción estudiantil en la Universidad de Nariño. La información se obtuvo de una fuente interna: la base de datos histórica de la Universidad de Nariño, compuesta por información personal y académica y una información externa: información de los colegios de educación secundaria del país.	Este trabajo pretende identificar las características que inciden en el rendimiento académico a partir de datos socioeconómicos, académicos e institucionales que se recogen en la base de datos de la universidad cuando el estudiante se matricula al primer semestre
En este trabajo se utilizan técnicas de clasificación y reglas de asociación, utilizando para ello los algoritmos C4.5 y EquipAsso.	Se pretende emplear el mejor modelo que se ajuste a los datos para la técnica predictiva y además clustering para la clasificación
Para predecir los perfiles de bajo rendimiento académico, se discretiza el atributo clase promedio el cual indica el rendimiento académico del estudiante basado en el promedio acumulado de las notas hasta el semestre cursado.	En cuanto a los datos relacionados con el rendimiento académico para este caso se toma el promedio acumulado, que se registra hasta el momento de la generación de los datos.
En este trabajo se tienen en cuenta los	En este trabajo se trata de conservar las

atributos más relevantes para la investigación además de aquellos que no contenían valores nulos.	dimensiones de la base de datos ya que interesa conocer qué variables explican mejor el rendimiento académico.

Fuente: este trabajo

Criterio de Búsqueda

- Dirección (URL): http://www.iiis.org/cds2008/cd2009cSc/CISCI2009/PapersPdf/C692YV.pdf

1.2 FORMULACIÓN Y DESCRIPCIÓN

Como se pudo evidenciar en los antecedentes, el problema de medir el rendimiento académico es un asunto de interés para las instituciones educativas en general, ya que es una forma de medir la calidad de la educación en las aulas en las que se imparte, entonces identificar las características que inciden en el rendimiento académico es indispensable pues a partir de este se pueden generar políticas educativas que permitan hacer frente de la mejor manera a los retos que la educación actual tiene.

Siendo este un tema de interés, existen diferentes investigaciones que lo abordan de manera variable; en algunos casos sólo para una materia al tiempo o para una sola carrera y un cierto grupo de materias, además en algunos de esos estudios, se utiliza una muestra de la población total de los datos; también se plantea este problema desde el punto de vista de la clasificación usando diferentes técnicas para tal caso, por ejemplo, árboles de clasificación, redes neuronales, vecino más cercano e incluso máquinas de vectores de soporte.

Adicionalmente, el rendimiento académico se define de forma diferente para cada trabajo revisado aquí, para unos es el número de materias aprobadas al final del periodo académico, para otros es tener en cuenta las notas de periodos pasados y para otros es medir este factor para estudiantes de plataformas online.

Debido a la situación expuesta anteriormente, surge la necesidad de ejecutar un trabajo de investigación que no sesgue el uso de una técnica de minería de datos desde el inicio y que tenga en cuenta toda la población de datos sin hacer ningún tipo de muestreo y que además defina el rendimiento académico sólo analizando el promedio acumulado hasta la fecha de corte, para el caso particular de este trabajo, no se pudo tener acceso a información adicional a la que se dispuso en la base de datos de la universidad, porque el promedio acumulado es la única variable que da cuenta del rendimiento académico de manera global en ese conjunto de datos.

2. JUSTIFICACIÓN

En los últimos años ha habido un interés creciente en el uso de técnicas avanzadas para el análisis de datos para abordar problemas o situaciones de cualquier tipo, siempre que estos provean una buena cantidad de datos; y de esta manera poder analizar aspectos particulares de cada asunto. Así, la educación universitaria también ha sido objeto de estos análisis con técnicas como la minería de datos, los resultados de estos análisis permiten establecer relaciones de acuerdo al objetivo establecido. Particularmente para este trabajo de investigación, resulta interesante determinar las características que se asocian al rendimiento académico de los estudiantes matriculados a pregrado de la Universidad de Nariño entre los años 2010 y 2014 usando los mencionados métodos.

Ya que en cualquier trabajo de investigación debe estar el factor novedad, ya sea cómo se plantea y aborda el problema, analiza o soluciona; este trabajo emprende el problema del rendimiento académico desde una perspectiva global para todas las carreras disponibles entre esos años y no se utiliza una muestra de la población total de los registros disponibles, esto difiere de los trabajos de investigación relacionados y consignados aquí.

Trabajos de investigación como este son pertinentes para el análisis y mejoramiento continuo de la situación educacional, porque a cualquier institución educativa superior le interesaría saber cómo se desempeñan sus estudiantes y qué es aquello que les afecta, este conocimiento abre las puertas a la mejora ya mencionada; también interesaría que, desde una perspectiva investigativa; los partícipes de este estudio puedan aportar sugerencias y recomendaciones que ayuden a la institución a gestionar los datos de los estudiantes de mejor manera, así se garantiza que estudios futuros centren más sus esfuerzos en el análisis del problema y no tanto en la limpieza y organización de los datos.

Específicamente para este trabajo de investigación, la viabilidad es posible ya que la universidad dispone de un sistema informático para gestionar los datos personales, académicos y socioeconómicos en una sola base de datos de todos sus estudiantes entonces con el permiso adecuado se puede acceder a esta información para aplicar técnicas de minería de datos.

Finalmente, en este trabajo de investigación se hace una revisión de las técnicas predictivas para regresión que podrían trabajar con este problema y desde el punto de vista práctico, se hace un buen uso de las técnicas de minería de datos para un caso real de investigación. Al final, la institución adquiere un análisis detallado de la situación, además de sugerencias generales alineadas a los resultados obtenidos aquí.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Identificar las características que inciden en el rendimiento académico de los estudiantes de pregrado de la Universidad de Nariño a partir de los datos personales, socioeconómicos, académicos e institucionales que se recogen en la base de datos de la universidad a través de técnicas de minería de datos.

3.2 OBJETIVOS ESPECÍFICOS

- Determinar las características de los estudiantes en relación a las variables socioeconómicas, académicas e institucionales.

- Identificar entre las variables personales, socioeconómicas, académicas e institucionales las que mejor predicen el rendimiento académico utilizando modelos predictivos.

- Analizar y comparar los modelos de agrupación en clústeres que identifiquen grupos de registros con características similares en torno al rendimiento académico.

- Elaborar recomendaciones sobre los posibles usos de los resultados obtenidos y las características de las fuentes de información que podrían servir como estrategias para la gestión académica.

4. DISEÑO METODOLÓGICO

4.1 POBLACIÓN Y MUESTRA

La población objeto de estudio de este trabajo de investigación son los estudiantes matriculados de los programas presenciales de pregrado de la Universidad de Nariño durante los años 2010 y 2014.

No se tiene en cuenta una base de datos más actual porque en el 2010 y 2014 la forma de evaluación de las pruebas ICFES[*] era diferente que las del año 2015 y 2016. En el 2015, se evaluó en las áreas: Lectura, Matemáticas, Sociales y Ciudadanía, Ciencias Naturales, Inglés, Razonamiento Cuantitativo, y

[*] Por motivos de uniformidad en el trabajo, las pruebas ICFES mencionadas en este documento se conocen así hasta el año 2013, desde el 2014 se conocen como Pruebas Saber 11. https://www.mineducacion.gov.co/1759/w3-article-244735.html

Competencias Ciudadanas; y en el 2016 las áreas evaluadas se modificaron nuevamente a Lectura Crítica, Matemáticas, Sociales y Ciudadanía, Ciencias Naturales e Inglés, por lo tanto se cree conveniente que para este trabajo de investigación se analice la base de datos entre los años 2010 y 2014.

Debido a que este es un trabajo de minería de datos y se dispone de toda la población objetivo entonces se decide trabajar con todos los registros y no tener muestra.

4.2 TIPO DE DISEÑO

Este estudio es de tipo descriptivo y predictivo.

4.3 VARIABLES

Las variables independientes que se estudian en este trabajo se organizan en las siguientes categorías:

- Personales: Fecha de Nacimiento, edad, sexo.

- Demográficas: ciudad nacimiento, ciudad permanencia, ciudad donde la universidad tiene sede, barrio.

- Socioeconómicas: estrato, jefe familia, más de una persona a cargo, tipo de residencia, vive con familia, número de hermanos universitarios, ingresos de la familia, año de ingreso, valor matricula colegio, año de pago colegio, pago de contado, puntaje.

- Antecedentes escolares: tipo de colegio, puntaje total ICFES, puntaje ICFES biología, puntaje ICFES matemáticas, puntaje ICFES física, puntaje ICFES filosofía, puntaje ICFES química, puntaje ICFES lenguaje, puntaje ICFES idiomas, puntaje ICFES sociales

- Universitarias: Ingreso a la universidad con cupo especial, estado actual, periodo egreso, fecha de grado, periodo grado, vigente actualmente, semestre actual, año ingreso, código, código carrera, código facultad, nombre de la carrera, nombre facultad, está vigente, periodo académico.

La definición de cada una de las variables se encuentra en el capítulo 5, sección **5.1.11.1 Comprensión de los datos**.

5. MARCO REFERENCIAL

5.1 MARCO TEÓRICO

5.1.1 Rendimiento Académico

La educación tanto en Colombia como en el mundo tiene por objeto la formación integral de las personas, se trata de una educación que supera la mera instrucción y que permite el desarrollo de las potencialidades, competencias y habilidades tanto sociales como intelectuales y afectivas de los individuos. Por lo tanto, la sociedad requiere de una educación que permita a las personas enfrentarse a los retos que propone un mundo en continuo cambio, una educación que promueva la creatividad, el pensamiento crítico, la autonomía, el trabajo en equipo, el cuidado del ambiente, y que forme no solo individuos capaces de adaptarse a las innovaciones que se presentan, sino que sean capaces de optimizar y crear nuevas herramientas tecnológicas, nuevos enfoques, teorías e ideas, que permitan ofrecer a la humanidad una mejor calidad de vida.

El rendimiento académico se constituye en una de las herramientas que determina los avances y progresos de los alumnos en su formación, además de ser una medida generalmente utilizada en la mayoría de los centros de enseñanza no solo para determinar el progreso de sus estudiantes sino también como un factor de medición de la calidad educativa que se imparte en estos centros, por lo que son numerosas las investigaciones que giran en torno a esta dimensión, centrándose algunas de ellas en determinar cuáles factores influyen en el rendimiento académico; con el propósito de establecer en la mayoría de los casos políticas educativas enfocadas a brindar ayuda a los estudiantes con mayor riesgo, es decir aquellos alumnos con dificultades académicas debidas en algunos casos a su situación económica, social, afectiva, institucional o personal; y no políticas recuperadoras como las que actualmente se aplican en varias de las instituciones educativas que tratan de apoyar a los alumnos que presentan bajo rendimiento académico, una vez este haya ocurrido.

Aunque existen varios factores tanto intrínsecos como extrínsecos que influyen en el rendimiento académico, en este trabajo se pretende analizar aquellos que el estudiante registró al momento de la matrícula al primer semestre de la universidad durante los años 2010 y 2014, además se utiliza el promedio acumulado que resulta de las valoraciones particulares que cada docente emite en su asignatura en el semestre correspondiente y por estudiante; y determinar con esta base de datos qué variables influyen en el rendimiento académico, todo esto utilizando técnicas de minería de datos.

Como se mencionó anteriormente, en este trabajo de investigación se toma al rendimiento académico como lo define la Universidad de Nariño en su estatuto estudiantil título IV, capítulo 2, Evaluación Académica (Nariño, 1998), artículo 106: "El promedio general acumulado y el semestral o anual de calificaciones de un estudiante, será el que resulte de calcular el promedio aritmético de todas las notas registradas, tomado en unidades, décimas y centésimas", este promedio acumulado es el resultado de una evaluación integral, permanente, sistemática, acumulativa, objetiva, formativa y consecuente, que todos los docentes de la Universidad de Nariño aplican a sus estudiantes como se menciona en el estatuto estudiantil (Nariño, 1998) en su artículo 90, sin embargo esta definición no demerita o contradice otras posiciones o puntos de vista respecto del rendimiento académico, ya que existen diferentes enfoques que permiten abordar el tema relacionado con los factores que influyen en este, por lo tanto, aquí se menciona unos trabajos de investigación centrados en esta medida.

El trabajo de (Erazo, 2012) muestra conceptualmente las relaciones y complejidades que se atraviesan para dar como resultado la nota y el promedio académico del estudiante; el autor hace una recopilación de las definiciones del rendimiento académico referenciando a una entidad gubernamental como el Ministerio de Educación (MEN) y otros autores, concluye que el rendimiento académico presenta características físicas y objetivas como la representación de la nota, las instancias políticas que la estructuran y los sistemas de evaluación que la justifican como elemento educativo y de evaluación en casi todos los países del mundo pero también resalta que es el resultado de los recursos y capacidades individuales del estudiante convirtiéndolo en una condición subjetiva y social.

Se puede ver de la recopilación hecha por el autor sobre el rendimiento académico que este es un tema complejo y no sólo relaciona al estudiante o al docente y su interacción sino que el rendimiento académico resulta de múltiples variables de tipo personal y social, además que este tema se convierte en un objeto de investigación importante para la educación, las ciencias sociales y la psicología educativa.

El español (Navarro, 2003) destaca que el rendimiento académico es una intrincada red de articulaciones cognitivas generadas por el hombre y sintetiza las variables de cantidad y calidad como factores de medición y predicción de la experiencia educativa, señala también que no es suficiente reducir el rendimiento académico a sólo un indicador de desempeño escolar sino que lo considera como una constelación dinámica de atributos cuyas características distinguen los resultados de cualquier proceso de enseñanza y aprendizaje.

En ese trabajo también se hace una reflexión personal respecto del rendimiento académico y lo relaciona con las habilidades sociales y el autocontrol, también propone que una investigación sobre el rendimiento académico debe ser una

comprensión integrada de manera inductiva y deductiva a través de una perspectiva holística.

La directora de la Escuela de Administración Educativa de La Universidad de Costa Rica en San José, Costa Rica (Garbanzo, 2007), realiza una revisión de los hallazgos de investigación consignados en la literatura que señalan los posibles factores asociados al rendimiento académico en estudiantes universitarios y su vinculación con la calidad de la educación superior pública en general. La autora habla de determinantes personales, sociales e institucionales y muestra la preocupación del Estado que invierte en la educación pública y busca mejorar su calidad, resalta que esta situación encadena estudios sobre rendimiento académico que permiten encontrar obstáculos y facilitadores vinculados a este. Para el Estado, señala la autora, el rendimiento académico es de gran utilidad para la toma de decisiones en aras de un sistema educativo más justo.

Finalmente, la directora concluye que el análisis del rendimiento académico se constituye en un factor imprescindible en torno a la búsqueda de la calidad de la educación superior, permite una aproximación a la realidad educativa y conocer los posibles factores que inciden en el rendimiento académico para predecir posibles resultados y hacer un análisis sobre su influencia en la calidad de la educación superior además que se convierte en una herramienta para la toma de decisiones.

5.1.2 Minería de datos

La Minería de Datos existe por la necesidad de analizar los grandes volúmenes de información almacenados en bases de datos de diferentes sistemas alrededor del mundo. Tan pronto como existieron las bases de datos y sistemas de administración de las mismas, las actividades humanas ejecutadas a través de estos sistemas empezaron a generar datos que se almacenaban en sistemas centralizados. Antes y debido a las limitaciones en las capacidades físicas de estos sistemas y a los altos costos de los mismos estos registros de datos no eran numerosos con respecto a los actuales; por ejemplo, en los años ochenta decir que se disponía de mil registros era considerado un gran volumen. Con el pasar del tiempo, el mejoramiento de estos sistemas y abaratamiento de los costos permitieron un mayor crecimiento de registros.

En los años noventa y con la posibilidad de tener computadores personales en casa, cada persona empezó a generar, casi involuntariamente más datos. Con la entrada a la era del internet y de la conexión simultánea de muchos sistemas informáticos el crecimiento de los datos es ahora exponencial. Por ejemplo, las instituciones bancarias, hospitales, instituciones gubernamentales, etc. ahora dependen de grandes sistemas de información y estos a su vez se apoyan en

grandes bases de datos, tanto así que decir que ahora se tiene mil registros se considera poco.

En la actualidad, además de las computadoras existen otros dispositivos que generan datos por ejemplo teléfonos móviles, sensores términos, sensores sismográficos, sensores meteorológicos, otros tipos de sensores, electrodomésticos conectados a internet, televisores inteligentes, dispositivos de posicionamiento global, dispositivos del Internet de las Cosas, entre otros; el volumen de los datos es cada vez mayor y exigen una creciente demanda de almacenamiento para guardar los mismos.

Dada la situación anterior, un ejemplo de minería de datos podría ser que en una cadena de supermercados se generan cientos o quizá miles de transacciones por día de las compras realizadas por sus clientes, estos datos yacen casi "inmóviles" en el sistema, sin que la administración se dé cuenta del potencial oculto que tiene a sus manos. Con esto se quiere decir que si el gerente tiene idea de que actualmente se puede analizar y procesar estos datos con el fin de generar algún tipo de reporte útil, esto puede representar un ahorro o ganancias traducidas en dinero para su organización.

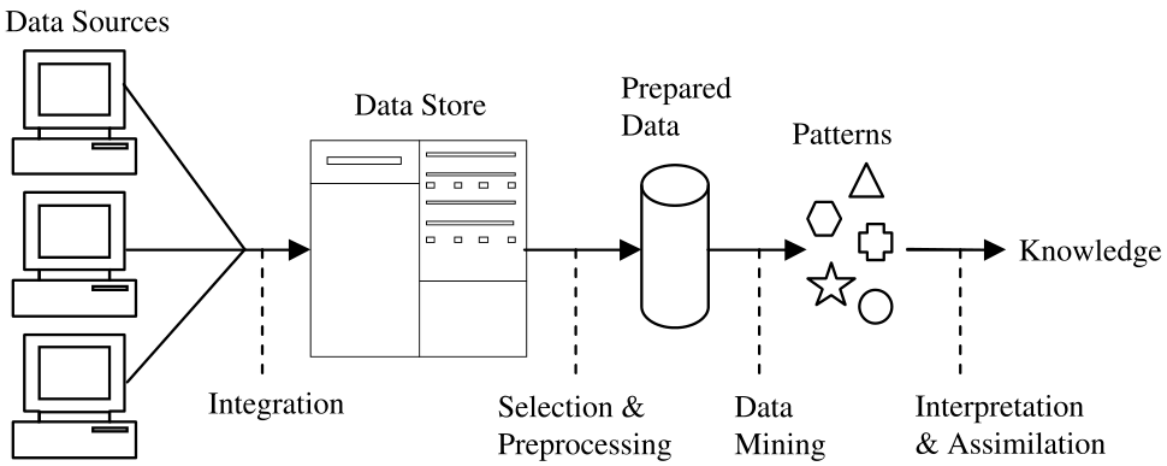
Otro ejemplo, un banco tiene a su disposición una base de datos con millones de registros de las transacciones de sus clientes, al cuerpo directivo le podría interesar analizar estos datos y ofrecer productos más personalizados a ciertos segmentos de su clientela con el fin de retener clientes. Esto se lograría a través de un estudio concienzudo con minería de datos.

Se puede ver de los ejemplos anteriores que al aplicar minería de datos se puede obtener o descubrir un conocimiento oculto que sirve para la toma de decisiones o mejora de procesos. Esto recibe el nombre de Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Data bases, por sus siglas en inglés).

(Bramer, 2007) define la minería de datos como "la extracción no trivial de información implícita previamente desconocida y potencialmente útil de los datos. Y en este proceso la minería de datos solo forma una parte, quizá la central".

Se puede ver en la figura 1 que en el Proceso de descubrimiento de Conocimiento la minería de datos es sólo un paso como menciona (Bramer, 2007) y que antes se debe disponer de los datos, integrar, seleccionar y pre-procesarlos, aplicar minería de datos y luego de interpretar y asimilar los resultados de todo este proceso entonces ya se puede considerar como conocimiento nuevo.

Figura 1. Proceso de descubrimiento de conocimiento



Fuente: (Bramer, 2007)

Por la versatilidad y flexibilidad de la minería de datos, esta se puede aplicar a diferentes áreas del conocimiento, aquí se nombran algunas:

- Finanzas
- Predicciones del clima
- Análisis de riesgo de créditos
- Evaluación de campañas publicitarias
- Selección o captación de estudiantes
- Detección de abandonos y de fracaso
- Identificación de patologías
- Gestión hospitalaria y asistencial
- Análisis de secuencias de genes
- Modelos de calidad de aguas, indicadores ecológicos
- Modelo de carga de redes
- Recursos Humanos: Selección de empleados
- Detección de fraude
- Detección de piezas con trabas. Modelos de calidad
- Hacienda: Detección de evasión fiscal

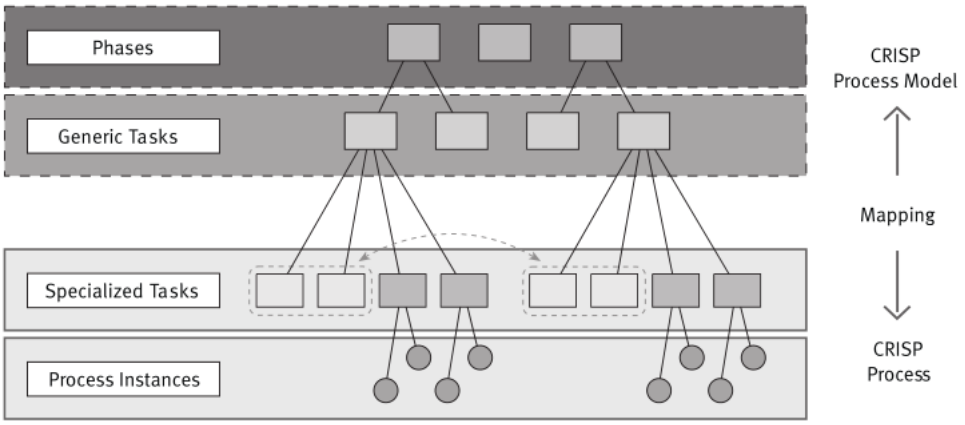
5.1.3 Metodología CRISP-DM

CRISP-DM es un modelo que describe una serie de pasos y procesos comúnmente usados por expertos en minería de datos para abordar los problemas concernientes a este campo.

Este trabajo de investigación se adhirió a esta metodología para el abordaje, desarrollo y evaluación del problema.

Este modelo organiza los diferentes pasos y procesos en forma jerárquica, como se muestra en la siguiente figura 2.

Figura 2. Jerarquía CRISP-DM



Fuente: (Chapman et al., 2000)

Las fases que comprende la metodología CRISP-DM son:

5.1.3.1 Comprensión del Negocio

En (Chapman et al., 2000) se establece que la Comprensión del Negocio es una fase en la que se busca comprender los objetivos del proyecto y requerimiento desde un punto de vista del negocio, luego de esto se establece un problema de minería de datos y un plan preliminar para lograr los objetivos.

Teniendo en cuenta la definición anterior, se examinó el rendimiento académico desde un punto de vista de la institucionalidad en la sección **5.1.1. Rendimiento Académico** y de igual manera, los objetivos se pueden detallar en el capítulo **3. OBJETIVOS**.

5.1.3.2 Preparación de los datos

En (Chapman et al., 2000), la fase de preparación de los datos cubre todas las actividades necesarias para construir el dataset final (los datos que serán el insumo de las herramientas de modelado) de los datos iniciales. Es probable que las tareas de preparación de datos se ejecuten varias veces y no en orden

preestablecido. Estas tareas pueden incluir selección de atributos, registros, tablas, así como también la transformación y limpieza de datos para las herramientas de modelado.

5.1.3.3 Modelado

En (Chapman et al., 2000), “en esta fase, se selecciona y aplica varias técnicas de modelado y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas requieren un formato específico en los datos. Así, regresarse a la fase de preparación de datos es a menudo necesario”.

5.1.3.4 Evaluación

En (Chapman et al., 2000), “En esta etapa del proyecto, se ha construido un modelo o modelos que aparentemente tienen alta calidad desde el punto de vista del análisis de datos. Antes de continuar con el despliegue final del modelo es importante evaluarlo completamente y revisar los pasos ejecutados para crearlo, para tener la certeza de que el modelo logra apropiadamente los objetivos del negocio. Un objetivo clave es determinar si hay algún problema de negocios que no ha sido suficientemente considerado. Al final de esta fase, se tiene que tomar una decisión respecto al uso de los resultados de la minería de datos”.

5.1.3.5 Despliegue

En (Chapman et al., 2000), “Generalmente el fin del Proyecto no es la creación del modelo. Aún si el propósito del modelo es incrementar el conocimiento de los datos, el conocimiento adquirido necesitará ser organizado y presentado en una forma que el cliente pueda usarla. A menudo requiere involucrar modelos reales dentro del proceso de toma de decisiones de la organización, por ejemplo, personalización en tiempo real de páginas web o calificación repetida en bases de datos de mercadeo. Dependiendo de los requerimientos, la fase de despliegue puede ser tan simple como generar un reporte o tan compleja como implementar un proceso de minería de datos repetible a lo largo de la organización. En muchos casos es el cliente, no el analista de datos; quien lleva a cabo los pasos de despliegue. Sin embargo, aun si el analista ejecutara el despliegue es importante que el cliente entienda de antemano que acciones son necesarias de llevar a cabo para hacer uso real de los modelos creados”.

5.1.4 Dataset, variables y observaciones.

En minería de datos, se entiende por variables a los atributos del conjunto de datos que se quiere estudiar. Las observaciones son los registros del conjunto de datos, las observaciones también son llamadas instancias. El dataset corresponde al conjunto de variables y observaciones bajo estudio.

5.1.5 Técnicas Descriptivas y Predictivas

Respecto a las Técnicas Predictivas, (Pérez López y Santín Gonzalez, 2007) mencionan: “Las técnicas predictivas especifican el modelo para los datos con base en un conocimiento teórico previo. El modelo supuesto para los datos debe contrastarse después del proceso de minería de datos antes de aceptarlo como válido. Formalmente, la aplicación de todo modelo debe superar las fases de identificación objetiva (a partir de los datos se aplican reglas que permitan identificar el mejor modelo posible que ajuste los datos), estimación (proceso de cálculo de los parámetros del modelo elegido para los datos en la fase de identificación), diagnosis (proceso de contraste de la validez del modelo estimado) y predicción (proceso de utilización del modelo identificado, estimado y validado para predecir valores futuros de las variables dependientes)”.

Algunas técnicas predictivas son:

- Regresión
- Series temporales
- Métodos bayesianos
- Algoritmos genéticos
- Técnicas de clasificación ad hoc: análisis discriminante, árboles de decisión y redes neuronales
- Máquinas de vectores de Soporte

Cabe resaltar que en los modelos predictivos, la variable a predecir (dependiente) puede ser numérica o categórica, esto conduce a una distinción de estos modelos así:

- Clasificación (cuando la variable dependiente es categórica): estos modelos utilizan el aprendizaje supervisado en el que se provee la respuesta deseada al algoritmo para que luego este sea capaz de generalizar una entrada nueva y decir a qué clasificación pertenece; las técnicas usadas, entre otras son: árboles de decisión, regresión logística, máquinas de vectores de soporte, redes neuronales.

- Predicción (cuando la variable dependiente es numérica): estos modelos tienen como objetivo predecir el valor de una variable numérica continua. Algunas técnicas usadas son: Ancova Simple, Ancova Múltiple, Regresión Múltiple con Variables Ficticias, Árboles QUEST, Árboles CART, Máquinas de Vectores de Soporte para Regresión, Redes Neuronales, Regresión Lineal Generalizada.

Con respecto a las Técnicas Descriptivas, (Pérez López y Santín Gonzalez, 2007) se refieren a estas: “En las técnicas descriptivas no se asigna ningún papel predeterminado a las variables. No se supone la existencia de variables independientes ni dependientes y tampoco se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente del reconocimiento de patrones”

Un listado de técnicas de aprendizaje no supervisado (no orientado) es:

- Reglas de asociación
- Dependencia
- Reducción de la dimensión
- Análisis exploratorio
- Escalamiento multidimensional
- Técnicas de clasificación post hoc: clustering y segmentación

5.1.6 Aprendizaje Supervisado

Las Técnicas Predictivas también se conocen como Aprendizaje Supervisado porque se infiere una función objetivo a partir de datos de entrenamiento etiquetados (labeled training data), es decir, el analista proporciona al algoritmo la variable objetivo (target variable) de la que se desea la respuesta.

(Hastie, Tibshirani, y Friedman, 2009) presentan el Aprendizaje Supervisado como una metáfora bajo la que el aprendizaje se realiza por un estudiante y es supervisada por un profesor. El estudiante presenta una respuesta Y por cada registro X en el conjunto de datos y el supervisor (profesor) provee una respuesta correcta o error de acuerdo a la respuesta del estudiante.

Ya que en este trabajo de investigación se desea predecir el rendimiento académico de los estudiantes de pregrado de la Universidad de Nariño entre los años 2010 y 2014 valiéndose de la variable Promedio Acumulado y sabiendo a priori que esta es numérica (continúa) y que las variables independientes son de naturaleza numérica o categórica (mixtas), estos son los algoritmos predictivos elegibles para este trabajo:

- Modelo lineal generalizado

- Árboles CART
- Árboles CHAID
- Máquina de vectores de soporte
- Redes Neuronales

Para este trabajo y los algoritmos predictivos antes mencionados, la variable objetivo o dependiente (Promedio Acumulado) no se discretiza o transforma, siempre se utiliza como es, es decir, en su naturaleza continua.

5.1.6.1 Máquinas de vectores de soporte (SVM)

Las máquinas de vectores de soporte es una técnica de minería de datos que se ha consolidado como una de las más usadas a la hora de ejecutar este tipo de proyectos porque proveen resultados con una precisión alta y requieren pocos datos para el entrenamiento.

(Deng, Tian, y Zhang, 2013) señala que se han hecho avances teórico-prácticos en las Máquinas de Vectores de Soporte y que se han implementado con éxito en muchos campos como la generalización de texto, reconocimiento de imagen y voz, seguridad de la información, series de tiempo, etc.

La teoría de las Máquinas de Vectores de Soporte tiene sus inicios en la Teoría del Aprendizaje Estadístico realizado por Vapnik a finales de los años 70 (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004) pero la forma en como se conoce actualmente a este algoritmo fue presentada por el mismo Vapnik y otros autores en los años 90 (Deng, Tian, y Zhang, 2013). Desde entonces el interés por esta técnica ha crecido por su alta eficiencia, el hecho de que se pueda evitar problemas como el sobreajuste, la alta dimensionalidad de los datos, entre otras.

Esta técnica es un algoritmo de clasificación lineal que traza un plano o hiperplano en espacios de características de alta dimensionalidad para clasificarlos en grupos. Aunque inicialmente las SVM se desarrollaron para problemas de clasificación también han sido extendidas para el problema de regresión.

Un hiperplano es un espacio de D-dimensiones $\Re D$, que se puede expresar como $h(x) =< \omega, x > +b$, donde $\omega \in \Re D$ y es el vector ortogonal al hiperplano, $b \in \Re$, y $<.,.>$ es el producto escalar habitual en $\Re D$. (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004)

El producto escalar habitual se define como:

$$< \omega, x > = \sum_{i=1}^{l} \omega_i x_i$$

Para un clasificador binario, la clasificación se puede expresar como $f(x) = signo(x)$, donde: (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004)

$$signo(x) = \begin{cases} +1 \ \ si\ x \geq 0 \\ -1 \ \ si\ x < 0 \end{cases}$$

Aquí, las $x \in \Re D$ son los ejemplos (registros) de los datos representados como vectores y que tienen una componente real por cada atributo, ω es el vector de pesos e indica la importancia en la clasificación por cada atributo, b es el sesgo y define el umbral de clasificación.

- **El problema del margen máximo**

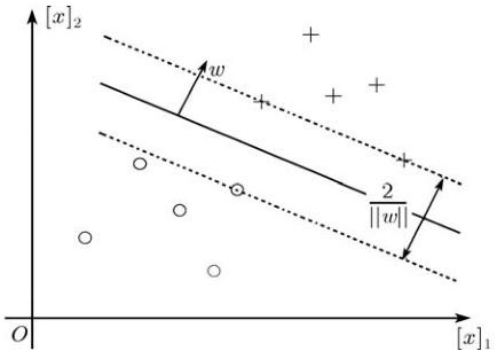
Si se tiene una tarea de clasificación binaria con un conjunto de datos linealmente separables entonces este problema es uno de optimización en el que se desea trazar un hiperplano de separación justo en medio de los datos teniendo como puntos de referencia los ejemplos más cercanos de cada clase.

Entiéndase aquí que *linealmente separables* quiere decir que los datos se pueden separar por un hiperplano correctamente (Deng, Tian, y Zhang, 2013).

Este problema se llama formalmente *Optimización Del Margen Máximo*. Para lograr este propósito sólo se tiene en cuenta los puntos de las clases que están en la frontera de la región de decisión y a estos se les llama *Vectores de Soporte*.

Mírese la gráfica de la figura 3. Los datos son una tarea de clasificación binaria y son linealmente separables. Las líneas no continuas son las Líneas de Soporte para cada clase en el conjunto de datos. La distancia entre estas dos líneas se llama margen, trazar cualquier hiperplano entre estas dos líneas de soporte es perfectamente posible y clasificarían correctamente a cada clase, no obstante, se desea que el hiperplano de separación se ubique justo en el medio del margen de forma que este hiperplano estará tan lejos del ejemplo más cercano de cualquier clase.

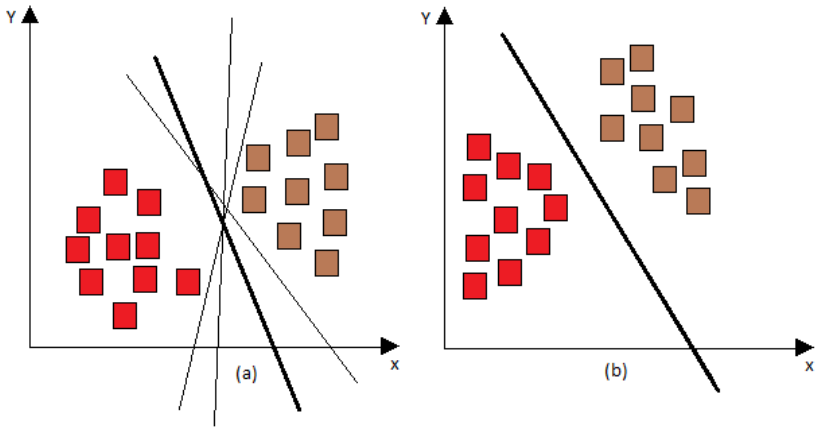
Figura 3. Gráfica del margen máximo para una SVM binaria



Fuente: (Deng, Tian, y Zhang, 2013)

En la figura 4, se puede apreciar lo que se dijo anteriormente respecto de que se puede trazar muchos hiperplanos de separación entre las 2 clases, gráfico (a), pero sólo luego de haber maximizado este hiperplano entonces se ubicará justo en el medio como se ve en la gráfica (b).

Figura 4. Gráfica del trabajo de una máquina de vectores de soporte



Fuente: este trabajo

El hecho de que el hiperplano de separación esté en el medio de los datos tiene dos razones importantes además del requerimiento matemático y es que primero, el algoritmo generalizará correctamente una entrada nueva y desconocida, y segundo, evitará un problema de sobreajuste ya que sin haber optimizado este margen máximo entonces el hiperplano se trazaría más cerca de la clase con mayor número de registros en el conjunto de datos.

- **Clasificación con máquinas de vectores de soporte**

Se describirá cómo las máquinas de vectores de soporte se utilizan para la clasificación binaria.

Considérese el siguiente problema de clasificación binaria con el siguiente conjunto N de datos: $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ donde cada x_i pertenece al espacio de entrada X y cada y_i señala la clase a la que pertenece x_i. Cada y_i puede valer $\{-1, +1\}$ y cada x_i es un vector (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004).

La máquina de vectores más básica es aquella que maximiza el margen (Margen Máximo) que hay entre el hiperplano de separación y los vectores más cercanos de cada clase. Este modelo asume de entrada que los datos son linealmente separables, es decir que no se necesita transformarlos para poder ubicar en medio de ellos un hiperplano que clasifique correctamente un ejemplo de otro. (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004)

Así, se resuelve el siguiente problema de maximización:

$$Maximizar \ \frac{1}{||\omega||}$$

$$Sujeto\ a\colon\ y_i(<\omega, x_i> + b) \geq 1$$
$$1 \leq i \leq N$$

La formulación más habitual de la SVM lineal con margen máximo y que es equivalente a la anterior es:

$$Minimizar \ \frac{1}{2} <\omega, \omega>$$

$$Sujeto\ a\colon\ y_i(<\omega, x_i> + b) \geq 1 \qquad 1 \leq i \leq N$$

(Orallo, Ramirez Quintana, y Ferri Ramírez, 2004) se puede ver que este es un problema de optimización convexa que consiste en minimizar una función cuadrática bajo restricciones en forma de desigualdad lineal. En esta formulación, el hiperplano queda caracterizado con un vector de pesos ω con una componente por cada atributo e indica la importancia relativa de cada atributo en el hiperplano solución.

Sin embargo, el problema de optimización bajo restricciones puede ser transformado con el uso de multiplicadores de Lagrange (Kantardzic, 2011).

$$L(\omega, b, \alpha) = \frac{||\omega||^2}{2} - \sum_{i=1}^{N} \alpha_i \{(< \omega, x^i > + b) y^i - 1\}$$

Donde α_i son los multiplicadores de Lagrange, uno para cada punto. El primer término es el mismo que en la función objetivo original y el segundo término captura las restricciones de desigualdad.

Derivando con respecto las variables originales, los Langragianos tienen que ser minimizados en función de ω y b.

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=0}^{N} \alpha_i y^i = 0$$

$$\frac{\partial L}{\partial \omega} = 0 \Rightarrow \omega_0 = \sum_{i=0}^{N} y^i \alpha_i x^i = 0$$

Sustituyendo en la función $L(\omega, b, \alpha)$ conduce a la formulación dual del problema de optimización que tiene que ser maximizado con respecto a las restricciones $\alpha_i \geq 0$.

$$D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^i y^j (x^i . x^j)$$

La anterior formulación requiere ser optimizada con técnicas de programación cuadrática (QP) que converge en un óptimo global. Cuando los parámetros α_i hayan sido encontrados entonces es necesario calcular los valores para ω y b que determinarán el hiperplano de clasificación final.

Para calcular ω_0:

$$\omega_0 = \sum_{i \in SVs} y^i \alpha_i^0 x^i$$

Para calcular b_0:

$$b_0 = -\frac{1}{2} \omega_0 . [x^r + x^s]$$

Donde x^r y x^s son cualquier vector de soporte de cada clase.

Entonces el clasificador se puede definir así:

$$f(x) = signo(< \omega_0 + x > + b_0) = signo\left(\sum_{i \in SVs} y^i \alpha_i^0 (x^i . x) + b_0\right)$$

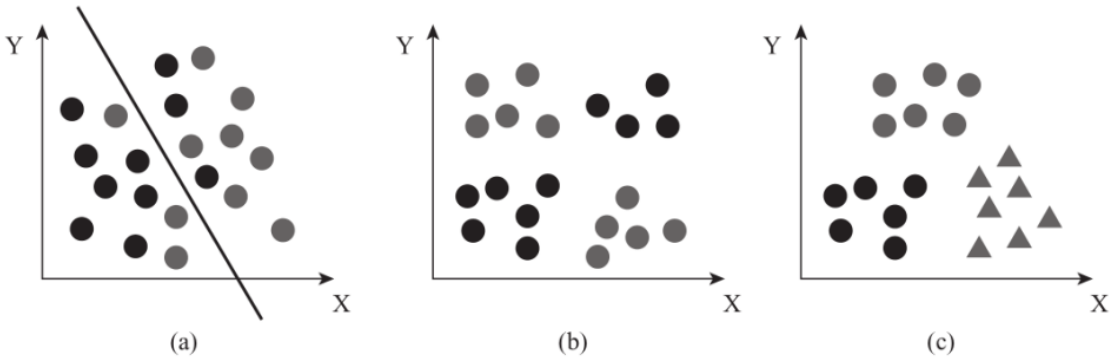
Sólo aquellos x^i que tengan multiplicadores de Lagrange diferentes de cero, $\alpha^0 \neq 0$; son llamados SVs (vectores de soporte). Si los datos son linealmente separables entonces todos los SVs estarán en la frontera de la región de decisión (Kantardzic, 2011).

Además se cumple que si los demás vectores (ejemplos) fueran eliminados del conjunto de entrenamiento, la SVM lineal de margen máximo sería la misma. (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004)

- **Máquina de vectores de soporte con margen suave**

En el mundo real, casi todos los problemas no son linealmente separables y aunque lo fueran, sería difícil trazar un hiperplano de separación entre ellos si el conjunto de datos tiene ruido. Sin mencionar que hay aplicaciones en las que hay más de dos categorías de clasificación, ver figura 5.

Figura 5. Máquina de vectores con problemas reales, (a) los subconjuntos no se pueden separar completamente, (b) separación no lineal, (c) problema multiclase



Fuente: (Kantardzic, 2011)

Para enfrentar este problema, se emplea una técnica llamada Soft Margin (Margen Suave) en la que se permite que algunos ejemplos del conjunto de entrenamiento resulten mal clasificados, a cambio se gana que el clasificador pueda generalizar mejor los ejemplos nuevos. (Kantardzic, 2011). Ver la figura 6.

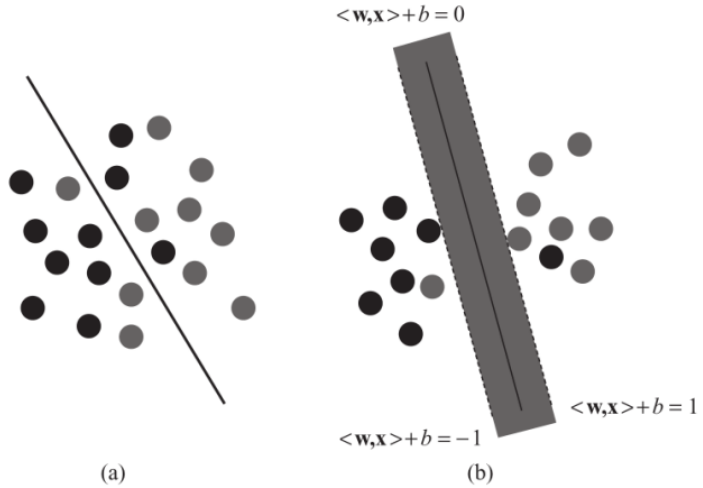
Modificando el problema de optimización para incluir el factor de violación para aquellos ejemplos que no cumplen con las restricciones.

$$Minimizar\ \frac{1}{2} < \omega, \omega > + C \sum_{i=1}^{N} \xi_i$$

Bajo las restricciones:

$$Sujeto\ a\colon\ y_i(< \omega, x_i > + b) \geq 1 - \xi_i \qquad 1 \leq i \leq N$$

Dónde:
C es el parámetro que representa el costo de violar las restricciones.
ξ_i son las distancias de los ejemplos que violan las restricciones.
Figura 6. Máquina de vectores con margen suave, (a) hiperplano de separación suave, (b) puntos de error con sus distancias



Fuente: (Kantardzic, 2011)

El parámetro C permite crear el margen suave (Soft Margin) en las máquinas de vectores de soporte que permite errores en la clasificación, C controla la compensación entre permitir errores en el entrenamiento y forzar un margen rígido, entonces si el valor de C es muy bajo, no habrá mucho estrés en el margen y por lo tanto se permite más errores, si el valor de C es muy alto entonces el costo por errores de clasificación se incrementa y por ende se crea un modelo que es más preciso pero que luego tendrá problemas de generalización.

El proceso de optimización es el mismo, sólo que ahora se cuenta con un límite superior C para todos los parámetros α_i. Este límite establece que tan grande se desea el margen en lugar de cuántos y por cuánto los ejemplos de entrenamiento violan este margen. (Kantardzic, 2011)

El proceso de optimización es como antes: cálculo de multiplicadores de Lagrange, optimización de los parámetros α_i, cálculo de los valores para ω y b

para el hiperplano de clasificación. El problema dual es el mismo sólo que se agrega la restricción: $0 \leq \alpha_i \leq C$.

- **Tipos de máquinas de vectores de soporte**

- **Clasificación con vectores de soporte (C-SVC)**

(SPSS Modeler, 2016) Dados los siguientes vectores de entrenamiento $x_i \in \Re^l$, $i = 1 \dots l$, en 2 clases y un vector $y \in \Re^l$ de forma que $y_i \in \{-1, 1\}$, entonces C-SVC resuelve el siguiente problema dual: (SPSS Modeler, 2016)

$$\min f(\alpha) = \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$

Donde e es el vector con todos los elementos igual a 1.

Teniendo en cuenta que $0 \leq \alpha_i \leq C, i = 1 \dots, l$ y $y^T\alpha = 0$, donde $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_l)^T$ y Q es una matriz de $l\ x\ l$, $Q(x_i.x_j) = y_i y_j K(x_i.x_j)$.

La función de decisión es:

$$Signo\left(\sum_{i=1}^{l} y_i\alpha_i K(x_i, x) + b\right)$$

Donde b es un término constante.

- **Regresión con Vectores de Soporte-ε (ε-SVR)**

(SPSS Modeler, 2016) En modelos de regresión, se estima la dependencia funcional de la variable objetivo $y \in \Re$ en un vector x de n dimensiones. Dado el conjunto de datos $\{(x_1, z_1), \dots, (x_l, z_l)\}$, de forma que $x_I \in \Re^n$ es una entrada y $z_i \in \Re^1$ es una salida objetivo, la forma dual de los vectores de regresión de soporte-ε es: (SPSS Modeler, 2016).

$$\min f(\alpha, \alpha^*) = \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + z_i\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)$$

De forma que $0 \leq \alpha_i$ y $\alpha_i^* \leq C$ para $i = 1, \dots, l$ y

$$\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0$$

Donde $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_l)^T$, $\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_l^*)^T$ y Q es una matriz de $l\,x\,l$, $Q_{ij} = K(x_i . x_j)$

La función de aproximación es:

$$\sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)\,K\left(x_i . x_j\right) + b$$

Donde b es un término constante.

- **Funciones Kernel**

Una SVM lineal asume, de entrada; que los datos son linealmente separables, entonces como se mencionó antes, se crea un hiperplano en el conjunto datos (Espacio de Entrada - *Input Space*) que sirve como clasificador.

Cuando el problema de clasificación provee un conjunto de datos que no son linealmente separables en el Espacio de Entrada, se hace necesario poder representar estos datos en otro espacio de mayor dimensionalidad, llamado Espacio de Características (*Feature Space*) en el que la separación lineal si es posible y por ende se pueda hacer una clasificación correcta de nuevos ejemplos.

Esta transformación del *Espacio de Entrada* al *Espacio de Características* es computacionalmente costosa pero ya que en el conjunto de datos sólo algunos ejemplos son en realidad Vectores de Soporte entonces las SVM pueden definir una función que calcula este mapeo de un espacio al otro y son llamadas funciones Kernel (K).

$$K(x, y) = < \Phi(\mathrm{x}), \Phi(\mathrm{y}) >$$

Como se ve, una función Kernel actúa como el producto escalar en el espacio de características, además, de esta manera se evita el costoso proceso computacional de representar explícitamente todos los datos de origen en vectores de mayor dimensionalidad. (Kantardzic, 2011)

Algunas funciones núcleo (Kernel) de propósito general en $\Re^D$ son: (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004)

- Polinómica: $(< x, y > +1)^d$
- Gaussiana: $\exp\left(\frac{-||x-y||^2}{\sigma^2}\right)$
- Sigmoidal: $\tanh(s < x, y > +r)$ $s, r \in \Re$
- Multicuadrática inversa: $\frac{1}{\sqrt{||x-y||^2+c^2}}$ $c \geq 0$

Aunque de ellas, las más usadas son las funciones Polinómica y Gaussiana (Kantardzic, 2011).

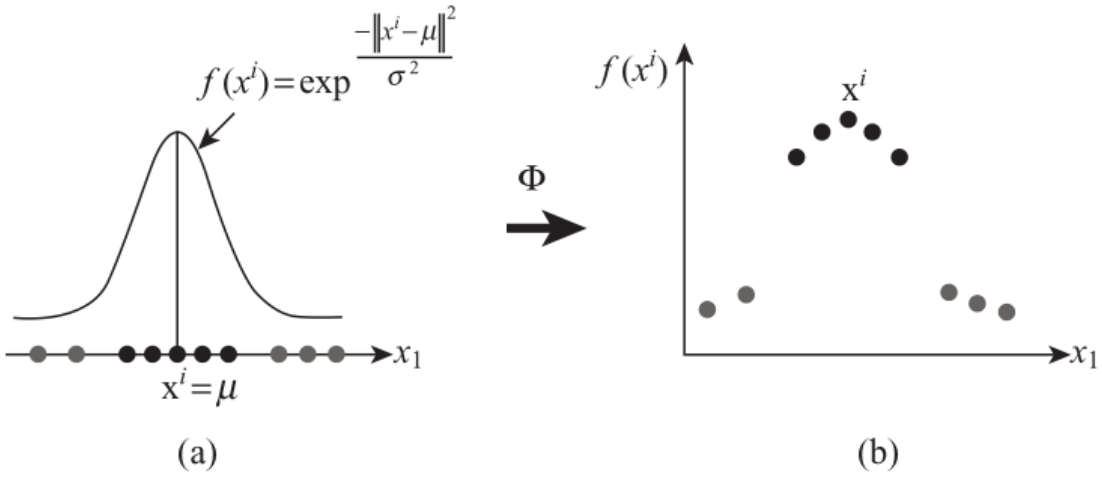
La función Gaussiana pertenece a un grupo de funciones Kernel llamado Funciones de Base Radial (Radial Basis Function - RBF) en la que la única dependencia es su distancia geométrica entre un punto x y otro y y sólo es válida cuando la anchura σ es diferente de cero. (Kantardzic, 2011)

La elección de una función Kernel apropiada con base en la naturaleza de los datos es muy importante al momento de ejecutar un proceso de clasificación con SVM porque la función Kernel define el Espacio de Características (Feature Space) sobre el que trabajará el algoritmo SVM y por ende se podrá hacer una separación correcta. (Kantardzic, 2011), (Tobergte y Curtis, 2013), ver figura 7.

El concepto de funciones de mapeo Kernel es muy poderoso, esto permite que las SVMs puedan realizar clasificaciones correctas aun cuando el problema de clasificación sea muy complejo. (Kantardzic, 2011)

Como punto de inicio para problemas de clasificación no lineal con Máquinas de Vectores de Soporte, la literatura recomienda el uso del Kernel RBF Gaussiano porque trabaja bien en la mayoría de los tipos de datos (Wendler y Gröttrup, 2016), sólo tiene un parámetro para ser ajustado σ y el tiempo de afinamiento del modelo es menor (Kantardzic, 2011), (Wendler y Gröttrup, 2016) en comparación con otras funciones Kernel.

Figura 7. Ejemplo de un mapeo Φ a un espacio de características en el que los datos si son linealmente separables. (a) espacio de entrada de una dimensión, (b) espacio de características de dos dimensiones



Fuente: (Kantardzic, 2011)

Finalmente, las Máquinas de Vectores de Soporte se están convirtiendo en una técnica muy usada en el área de minería de datos, este algoritmo está incluido en paquetes de software cada vez más profesionales y amigables para el usuario, esto ha permitido aplicarlo a problemas de la vida real que tienen datasets muy grandes.

Con respecto a problemas de clasificación multiclase, hay aportes de investigación en los que se ha modificado la función objetivo del problema de optimización para obtener un clasificador multiclase proveyendo resultados experimentales prometedores frente a otras técnicas que abordan este problema en varios subproblemas de binarización.

Otra rama de investigación que tiene en cuenta las Máquinas de Vectores de Soporte es el aprendizaje no supervisado o clustering, hay aportes en los que se usa un núcleo gaussiano para encontrar la mínima esfera que incluya todos los puntos en el espacio de características entonces el proceso de clustering se puede controlar modificando los parámetros como la anchura σ del núcleo y la constante del margen blando (Orallo, Ramirez Quintana, y Ferri Ramírez, 2004).

5.1.7 Criterios para comparar modelos predictivos

Aunque hay técnicas estándar para comparar el rendimiento de modelos de minería de datos, como el error de predicción o las curvas ROC (Receiver Operating Characteristic), estas se aplican cuando el modelo es de clasificación (Kantardzic, 2011).

Ciertamente, un modelo de clasificación puede predecir a cuál categoría pertenece un ejemplo, teniendo como variable objetivo una categórica, sólo en estos casos

es necesario medir el rendimiento del modelo utilizando alguna técnica como las mencionadas en el párrafo anterior. Un ejemplo claro de un problema de clasificación es uno de diagnóstico médico para establecer si un paciente está enfermo o no está enfermo de una cierta dolencia.

Vale recordar que aunque uno de los objetivos de este trabajo es predecir, esto se lleva a cabo con una variable dependiente continua (Promedio Acumulado) con la que no se hace una tarea de clasificación sino de regresión utilizando Máquinas de Vectores de Soporte, por esa razón para comparar modelos predictivos cuya variable dependiente es numérica se utiliza los siguientes criterios:

5.1.7.1 Correlación

La correlación es un coeficiente estadístico que permite medir la asociación lineal entre dos variables. El valor de este coeficiente está entre -1 y 1. Cuando el valor de la correlación es -1 significa que la asociación lineal entre una variable x y otra y es opuesta o negativa, cuando es 1 su significado es que la asociación lineal entre esas 2 variables es positiva o semejante. Si el valor es 0 (cero) entonces no hay asociación lineal entre las 2 variables. (Kantardzic, 2011)

En este contexto, se utilizó la correlación de Pearson, entonces para medir qué modelo es mejor que otro, el valor de la correlación entre la variable dependiente (observada) y la variable que predice el modelo debe ser positivo y tan alta (cercano a 1) como sea posible, en este caso la correlación entre el rendimiento académico observado y el rendimiento académico que predice el modelo.

5.1.7.2 Error relativo

En (IBM, 2016) el error relativo se define como: “El error relativo es el cociente de la varianza de los valores observados de aquellos pronosticados por el modelo a la varianza de los valores observados de la media. En la práctica, compara el buen rendimiento del modelo con respecto a un modelo nulo o de intersección que simplemente devuelve el valor medio del campo objetivo como el pronóstico. En un buen modelo, este valor debe ser inferior a 1, lo que indica que el modelo es más preciso que el modelo nulo. Un modelo con un error relativo superior a 1 es menos preciso que el modelo nulo y por lo tanto no es útil. En el caso de modelos de regresión lineal, el error relativo es igual al cuadrado de la correlación y no añade información nueva. En el caso de modelos no lineales, el error relativo no está relacionado con la correlación y proporciona una medida adicional para valorar el rendimiento del modelo”.

5.1.8 Importancia de las variables

La importancia se define como aquellas variables independientes que retienen mayor información sobre la variable dependiente, en este caso las variables más importantes son aquellas que retienen mayor información sobre el rendimiento académico (**promedio_acumulado** – variable dependiente).

Formalmente, la importancia de las variables se define así:

5.1.8.1 Definición formal de la importancia de las variables predictoras

Se refiere a qué tan importante es una variable predictora en la construcción del modelo predictivo. La importancia de la variable puede ser determinada calculando la reducción de la varianza de la variable respuesta atribuible a cada predictor, a través de un análisis de sensibilidad, este mismo método es utilizado en los algoritmos: C5.0, CART, QUEST, CHAID, Regresión Logística, Análisis Discriminante, GenLin, SVM, y Redes Bayesianas, estos modelos sólo se limitan a inferir qué factores o variables predictoras inciden y con qué peso a la hora de pronosticar pero no indican la dirección puesto que las relaciones por lo general no son lineales y sería muy difícil su interpretación.

(A. Saltelli et al., 2004) proponen esta notación para entender mejor el procedimiento:

Y – Variable objetivo
x_j – Predictor, donde $j = 1, \dots, k$
k – El número de predictores
$Y = f(X_1, X_2, \dots, X_k)$ – Modelo para Y basado en los predictores X_1 hasta X_k

Entonces utilizando el método basado en la varianza (A. Saltelli et al., 2004) lo desarrollan como:

Los predictores son priorizados de acuerdo a la medida de sensibilidad definida como:

$$S_i = \frac{vi}{V(Y)} = \frac{V\big(E(Y|X_i)\big)}{V(Y)}$$

Donde $V(Y)$ es la varianza incondicional. En el numerador, el operador de la Esperanza E requiere una integral sobre X_{-i}; esto es, sobre todos los predictores excepto X_i, entonces el operador de varianza V implica una integral adicional sobre X_i.

Por ende la importancia de la variable predictora se calcula como la sensibilidad normalizada

$$VI_i = S_i / \sum_{j=1}^{k} S_j$$

(A. Saltelli et al., 2004) muestran que (S_i) es una medida apropiada de sensibilidad para priorizar los predictores en orden de importancia para cualquier combinación de interacción y no ortogonalidad entre los predictores.

La medida de importancia S_{i} es la medida de sensibilidad de primer orden, la cual es exacta si el conjunto de factores de entrada (predictores) $(X_1, X_2, X_3, \dots, X_k)$ es ortogonal/independiente (una propiedad de los factores), y el modelo es aditivo; esto es, el modelo no incluye interacciones (una propiedad del modelo) entre factores de entrada. Para cualquier combinación de interacciones y no ortogonalidad entre factores, (A. Saltelli et al., 2004) señalan que S_{i} es aún la medida de sensibilidad apropiada para priorizar los factores de entrada en orden de importancia, pero existe un riesgo de imprecisión debido a la presencia de interacción o no ortogonalidad. Para una mejor estimación de S_{i}, el tamaño de la base de datos deber ser de unos cientos al menos. De lo contrario, S_{i} puede ser altamente sesgada. En este caso, la medida de importancia apropiada puede ser utilizando bootstrapping.

- **Computación**

En el caso ortogonal, es sencillo estimar la varianza condicional V_i calculando las integrales multidimensionales en el espacio de los factores de entrada (predictores) a través de métodos de Monte Carlo como el siguiente:

Hay que iniciar con 2 matrices M_1 y M_2, cada una con dimensión $N * k$:

$$M_1 = \begin{matrix} x_1^{(1)} & x_2^{(1)} & \dots & x_k^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_k^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(N)} & x_2^{(N)} & \dots & x_k^{(N)} \end{matrix}$$

y

$$M_2 = \begin{matrix} x_1^{(1\prime)} & x_2^{(1\prime)} & \dots & x_k^{(1\prime)} \\ x_1^{(2\prime)} & x_2^{(2\prime)} & \dots & x_k^{(2\prime)} \\ \dots & \dots & \dots & \dots \\ x_1^{(N\prime)} & x_2^{(N\prime)} & \dots & x_k^{(N\prime)} \end{matrix}$$

Donde N es el tamaño de la estimación de Monte Carlo la cual puede variar entre unos cientos a unos miles. Cada registro es un ejemplo de entrada. De M_1 y M_2 se puede construir una tercera matriz N_j:

$$N_j = \begin{matrix} x_1^{(1\prime)} & x_2^{(1\prime)} & \dots & x_j^{(1)} & \dots & x_k^{(1\prime)} \\ x_1^{(2\prime)} & x_2^{(2\prime)} & \dots & x_j^{(2)} & \dots & x_k^{(2\prime)} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_1^{(N\prime)} & x_2^{(N\prime)} & \dots & x_j^{(N)} & \dots & x_k^{(N\prime)} \end{matrix}$$

Se puede pensar que M_1 es la matriz de muestra, M_2 es la matriz de re-muestra, y N_j como la matriz donde todos los factores excepto X_j son re-muestrados. Las siguientes ecuaciones describen cómo obtener las varianzas (a Saltelli, 2002). El carácter ^ denota la estimación numérica:

$$\hat{V}(Y) = \frac{1}{N-1}\sum_{r=1}^{N} f^2(x_1^{(r)}, x_2^{(r)}, \dots, x_k^{(r)}) - \hat{E}^2(Y)$$

Donde

$$\hat{E}^2(Y) = \left[\frac{1}{N}\sum_{r=1}^{N} f(x_1^{(r)}, x_2^{(r)}, \dots, x_k^{(r)})\right]^2$$

$$\hat{V}\left[E(Y|X_j)\right] = \hat{U}_j - \hat{E}^2(Y)$$

Donde

$$\hat{U}_j = \frac{1}{N-1}\sum_{r=1}^{N} f\left(x_1^{(r)}, x_2^{(r)}, \dots, x_k^{(r)}\right) f(x_1^{(r')}, x_2^{(r')}, \dots, x_{J-1}^{(r')}, x_J^{(r)}, x_{(J+1)}^{(r')}, \dots, x_k^{(r')})$$

$$\hat{E}^2(Y) = \frac{1}{N}\sum_{r=1}^{N} f\left(x_1^{(r)}, x_2^{(r)}, \dots, x_k^{(r)}\right) f\left(x_1^{(r')}, x_2^{(r')}, \dots, x_k^{(r')}\right)$$

Cuando la variable objetivo es continua, simplemente se sigue los pasos de acumulación de la varianza y esperanza. Para una variable categórica, los pasos de acumulación son para categoría de Y. Para cada predictor, S_{i} es un vector con un elemento para cada categoría de Y. El promedio de los elementos de S_{i} se usa como la estimación de la importancia del ith predictor en Y.

- **Convergencia**

Para mejorar la escalabilidad, se usa un subconjunto de los ejemplos y predictores cuando se verifica la convergencia. Específicamente, la convergencia se juzga con el siguiente criterio:

$$i \overset{\cap}{\in} I \;\; \frac{\mathrm{i}}{D} \sum_{j=t-D+1}^{t} \frac{|S_i(j) - \overline{S_i}|}{\overline{S_i}} < \epsilon$$

Donde $I = \{1 | S_i(t) > 1/num\}$, $D = 100$ y significa la anchura del interés

$$\overline{S_i} = \frac{1}{D} \sum_{j=t-D+1}^{t} S_i(j)$$

Y $\epsilon = 0.005$ define el promedio del error relativo deseado.

- **Orden de los registros**

Este método de cálculo de la importancia del predictor es deseable porque escala bien en bases de datos grandes, pero los resultados dependen del orden de los registros en la base de datos. Sin embargo, con bases de datos grandes y ordenadas aleatoriamente, se puede esperar que la importancia del predictor sea consistente.

5.1.9 Aprendizaje No Supervisado

Usando la misma idea en la metáfora de (Hastie, Tibshirani, y Friedman, 2009) para el Aprendizaje Supervisado, los autores muestran al Aprendizaje No Supervisado como el hecho de aprender sin la necesidad de un profesor o supervisor. (Hastie, Tibshirani, y Friedman, 2009) dicen al respecto “en el aprendizaje no supervisado, se dispone de un conjunto de N observaciones de un vector aleatorio P que tiene $Pr(X)$ de probabilidad, el objetivo es inferir las propiedades de esa probabilidad $Pr(X)$ sin la necesidad de un profesor o supervisor proveyendo respuestas correctas o grados de error para cada observación”.

Adicionalmente, en el Aprendizaje No Supervisado, no hay una variable objetivo definida y tampoco se devuelve una predicción. Aquí están las técnicas como el clustering que tratan de capturar relaciones interesantes de los datos y proveer descripciones útiles de las agrupaciones (IBM, 2016).

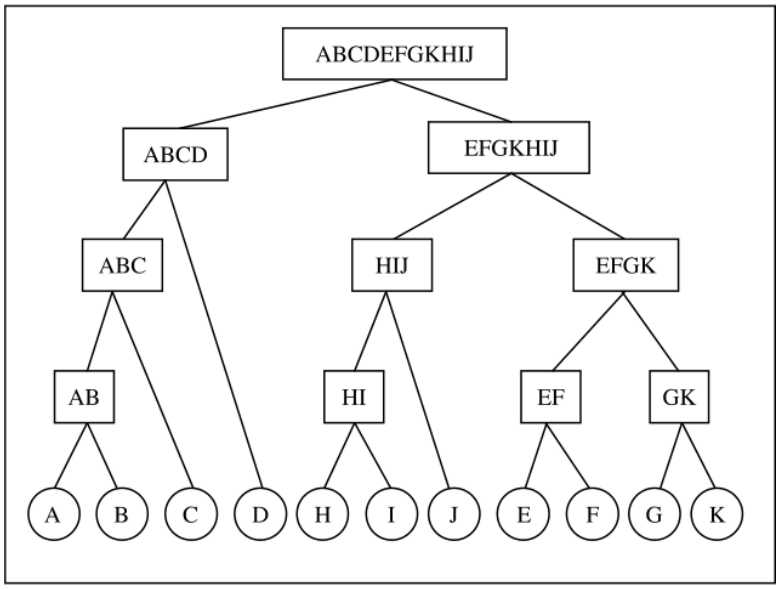
5.1.9.1 Clustering

Al respecto, (Tuffery, 2011) provee una definición muy completa, el autor menciona que el clustering es la operación estadística en la que se agrupan objetos (individuos o variables) en un número limitado de grupos conocidos como clústeres (ó segmentos), que tienen 2 propiedades. Por una parte, ellos no son definidos de manera anticipada por el analista, sino que son descubiertos durante el proceso, a diferencia de las clases usadas en la *clasificación*. Por otra parte, los clústeres son combinaciones de objetos que tienen características similares, que están separados de objetos que tienen características diferentes (resultando en homogeneidad interna y heterogeneidad externa). Esto se puede medir por un criterio como la suma de los cuadrados. Como en la clasificación, la esencia del clustering es la distribución de objetos en grupos. Sin embargo, esta clasificación no se lleva a cabo en la base de un criterio predefinido, y no está intencionado en combinar los objetos que tengan el mismo valor para tal criterio. En otras palabras, el cluster al que pertence el objeto no se conoce con anterioridad, en contraste con el proceso de clasificación. Incluso el número de clústeres no es definido anteriormente. Esto es porque no hay una variable que sea dependiente: el clustering es descriptivo, no predictivo. Se usa ampliamente en el mercadeo, la medicina, la sociología, y otros campos similares. En mercado, se suele referenciar al clustering como “segmentación” ó “análisis tipológico”. En medicina, el término usado es “nosología”. En biología y zoología, se habla de “taxonomía numérica”.

5.1.9.2 Clustering Jerárquico

Brevemente, el clustering jerárquico (Hierarchical Clustering en Inglés), es el proceso de agrupar puntos de un conjunto de datos en diferentes niveles de forma anidada. El resultado se puede representar en un gráfico llamado Dendrograma.

Figura 8. Dendrograma, representación gráfica del resultado de un proceso de clustering jerárquico



Fuente: (Bramer, 2007)

En la figura 8, se puede ver que el primer nivel (nodos hojas) corresponde a todos y cada uno de los elementos del conjunto de datos como un clúster de un solo miembro, en el segundo nivel, hay clústeres formados por 2 elementos y de esa forma hasta el primer nodo (nodo raíz) en el que están todos los elementos contenidos en un solo clúster (Bramer, 2007).

5.1.9.3 K-Medias (K-Means)

Es un algoritmo de clustering no jerárquico pues crea clústeres en un solo nivel que a su vez son similares internamente pero disimilares externamente.

Cuando el algoritmo termina su proceso, este no se puede representar gráficamente con un dendrograma como se apreció anteriormente pues hay solo un nivel para todos los clústeres, por esto y por la forma como el algoritmo crea los clústeres se le llama no jerárquico.

K-Means es ideal para grandes volúmenes de datos y también se puede utilizar para la detección de datos aberrantes (Perez Marqués, 2015), es muy popular en problemas de segmentación porque es fácil de implementar y su complejidad es relativamente baja. Lamentablemente una gran desventaja de este algoritmo es que se tiene que especificar desde el inicio el número de clústeres deseados (Perez Marqués, 2015).

(Kantardzic, 2011) indica que básicamente el algoritmo inicia con una partición aleatoria, procede con la reasignación de ejemplos a los clústeres, esta reasignación se basa en la similitud entre la instancia y el clúster; hasta que se

cumple un criterio de parada que puede ser: un número máximo de iteraciones o que ya no haya reasignaciones para hacer que reduzca el error cuadrado total.

Los pasos básicos del algoritmo son (Kantardzic, 2011):

1. Seleccionar una partición inicial con K clústeres que contienen instancias aleatorias, calcular los centroídes de los clústeres.
2. Generar una nueva partición asignando cada ejemplo al centro del clúster más cercano.
3. Calcular los centros de los clústeres como centroídes.
4. Repetir los pasos 2 y 3 hasta que un valor óptimo de la función de criterio se haya encontrado o hasta que la pertenencia a un clúster se haya estabilizado.

En el punto 4 del algoritmo anterior, se menciona una función de criterio que puede ser el Error Cuadrado Medio de la distancia Euclídea (criterio global) o la Distancia del Vecino Mutua (criterio local) (Kantardzic, 2011), entonces la idea central de esta función de criterio es encontrar el valor óptimo y cuando se haya encontrado el algoritmo termina y devuelve los clústeres creados hasta ese momento.

La herramienta utilizada para la ejecución del algoritmo K-Means y la creación de los clústeres se mencionarán posteriormente en el documento; No obstante, se menciona que dicho software ejecuta los siguientes cálculos cuando va a crear el modelo de clustering, todo para compensar diferencias en las medidas, escala y tipo de las variables.

- **Codificación de campos**

Las variables de entrada son codificadas por el software con anterioridad antes de que sus valores sean el insumo para el algoritmo K-Means.

Escala de campos de rango

Para evitar que a un campo le sea dada más importancia porque sus valores pueden oscilar entre varias posibilidades, por ejemplo: edad de una persona; los campos de rango se transforman (reescalan) para que tengan la misma escala con valores entre 0 y 1 (SPSS Modeler, 2016).

La transformación usada es

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Donde x_i' es el valor reescalado para el campo de entrada x para el registro i, x_i es el valor original de x para el registro i, x_{min} es el menor valor de x para todos los registros y x_{max} es el valor máximo de x para todos los registros.

Codificación numérica de campos categóricos

Para algoritmos que basan sus cálculos como diferencias numéricas entre los registros, los campos categóricos imponente una dificultad (SPSS Modeler, 2016).

Una solución común y la solución usada por la herramienta es recodificar las categorías del campo original como un grupo de campos numéricos con un campo numérico para cada categoría del campo original. Para cada registro, el valor del campo resultante correspondiente a la categoría de ese registro se asigna con el 1 y los otros campos resultantes se asignan con el cero (0).

Ejemplo, para los siguientes datos en los que x es una variable categórica con valores posibles A, B, y C.

Registro	$\boldsymbol{x}$	$\boldsymbol{x_1'}$	$\boldsymbol{x_2'}$	$\boldsymbol{x_3'}$
1	B	0	1	0
2	A	1	0	0
3	C	0	0	1

Para los datos anteriores, el campo original x se codifica a los campos derivados: $x_1^{'}$, $x_2^{'}$ y $x_3^{'}$, las categorías A, B y C corresponden a cada campo nuevo respectivamente.

Aplicando el valor codificado

Luego de codificar los campos como se mostró anteriormente, el algoritmo puede calcular diferencias numéricas para el campo original tomando la diferencia en los k campos derivados (donde k es el número original de categorías). Sin embargo, hay un problema. Para algoritmos que usan la distancia Euclídea para medir diferencias entre los registros, la diferencia entre dos registros con diferentes valores i y j es (SPSS Modeler, 2016):

$$\sqrt{\sum_{k=1}^{J}(x_{k1} - x_{k2})^2}$$

Donde J es el número de categorías, y x_{kn} es el valor del indicador derivado para la categoría k para el registro n. Pero los valores serán diferentes en dos de los indicadores derivados, x_i y x_j. Así, la suma será:

$$\sqrt{(1-0)^2+(0-1)^2} = \sqrt{2} \approx 1.414$$

Que es mayor que 1.0. Eso significa que basándose en esta codificación, los campos categóricos tendrán más peso en el modelo que campos de rango que se escalan entre 0 - 1.

Para tener en cuenta este sesgo, el K-Means aplica un valor de escala a los campos derivados, esa diferencia entre los valores del campo categórico produce una distancia Euclídea de 1.0. El valor por defecto del factor de escala es $\sqrt{1/2} \approx 0.707$. Este valor se inserta en la fórmula de la distancia.

$$\sqrt{\left(\sqrt{\frac{1}{2}}-0\right)^2+\left(0-\sqrt{\frac{1}{2}}\right)^2} = \sqrt{\frac{1}{2}+\frac{1}{2}} = 1$$

Codificación numérica de campos binarios

Los campos binarios son un caso especial de los campos categóricos. Sin embargo, ya que ellos sólo tienen dos valores posibles, estos se pueden lidiar de una forma más eficiente que los campos categóricos. Los campos binarios se representan con un solo campo numérico que toma el valor de uno (1) para la categoría “Verdadero”, y el cero (0) para la categoría “Falso”. Los vacíos para campos binarios se asignan el valor 0.5.

5.1.10 Criterios para comparar modelos de segmentación

5.1.10.1 El coeficiente de la silueta

Es una medida de clustering que usa promedios de proximidades y que es útil cuando se busca clústeres compactos (favoreciendo modelos que tienen clústeres altamente cohesionados) y separados (favoreciendo modelos que tienen clústeres altamente separados) (Rousseeuw, 1987), en pocas palabras, permite medir la calidad de todos y cada uno de los clústeres o de todo el clustering en general, entonces entre más grande sea la media de la silueta, mejor será la calidad de un clúster o de todo el clustering.

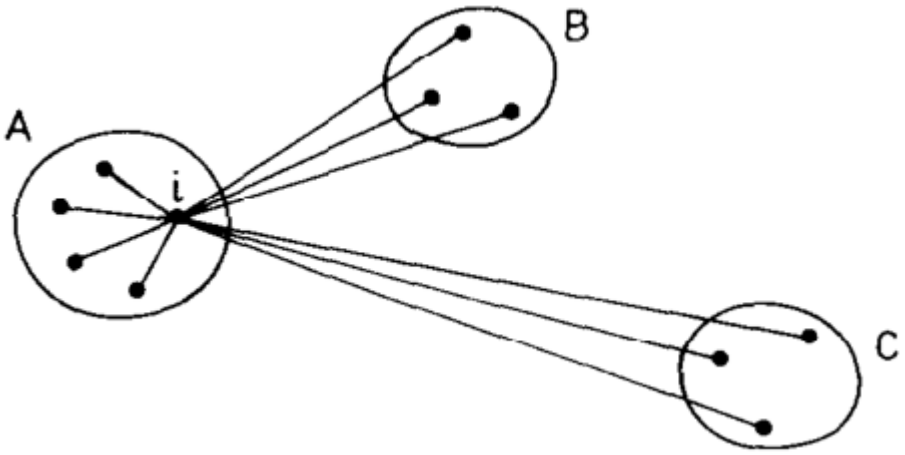
Para poder calcular este coeficiente, es necesario disponer de los valores de proximidades entre los elementos de la base de datos. Tales proximidades pueden ser 2: disimilaridades (que mide cuán lejos están 2 objetos de si mismos) y similaridades (que mide en cuánto se parecen 2 objetos), por esto, las proximidades deben estar en escala de razón, por ejemplo, la distancia Euclídea (Rousseeuw, 1987).

- **Construcción de la silueta**

(Rousseeuw, 1987) menciona que para calcular el coeficiente de la silueta, se necesita 2 cosas: la partición resultante del proceso de clustering y la colección de proximidades entre los objetos. Para un objeto i se crea un valor $s(i)$(silueta) y luego todos estos valores se combinan en un gráfico.

Para entender mejor cómo se calcula el valor $s(i)$, hay que tener en cuenta la figura 9.

Figura 9. Distancias de elementos en clúster y entre clústeres



Fuente: (Rousseeuw, 1987)
De la figura anterior, se puede calcular los siguientes valores para la disimilaridades (Rousseeuw, 1987):

De la base de datos tomar cualquier objeto i, denotar como A el clúster al que ha sido asignado el objeto i.

- $a(i)$ = promedio de la disimilaridad de i hacia todos los objetos de A, es decir, el promedio de todas las líneas en A.

Considerar cualquier clúster C que es diferente de A.

- $d(i, C)$ = promedio de la disimilaridad de i hacia todos los objetos de C, es decir, el promedio de todas las líneas que van de i hasta C. Después de calcular $d(i, C)$ para todos los clústeres $A \neq C$, se selecciona el menor de los valores y se denota como:

$$b(i) = \underset{C \neq A}{minimo\ d(i, C)}$$

Para este caso, $b(i)$ es para el clúster B pues está más cercano, es decir $d(i,B) = b(i)$, a este valor se le llama $Vecino$ que sería la segunda opción para i en el caso de que no se pueda clasificar en A.

Con lo mencionado anteriormente, es necesario calcular estos valores para todos los elementos en la base de datos siempre que el número k de clústeres sea mayor que 1.

El número $s(i)$ se obtiene combinando los valores $a(i)$ y $b(i)$ así:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & si\ a(i) < b(i) \\ 0 & si\ a(i) = b(i) \\ b(i)/a(i) - 1 & si\ a(i) > b(i) \end{cases}$$

En una sola fórmula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Para tener una mejor perspectiva del significado de $s(i)$, considérese las siguientes situaciones:

Cuando $s(i)$ está muy alto (esto es, $s(i)$ cerca de 1) implica que la disimilaridad en el clúster $a(i)$ es menor que la disimilaridad entre clústeres $b(i)$ más pequeña. Por ende, se puede decir que i está bien clasificado, la segunda opción para i que sería el clúster B, no está tan cerca como el clúster actual A (Rousseeuw, 1987).

Cuando $s(i)$ es casi cero (0). Entonces $a(i)$ y $b(i)$ son casi iguales, y no es claro del todo si i debió haber sido clasificado a A o B. El objeto i yace equitativamente de los 2 clústeres, entonces se debería considerar como un caso intermedio (Rousseeuw, 1987).

Cuando $s(i)$ es casi -1. Entonces $a(i)$ es más grande que $b(i)$, es decir que en promedio i yace más cerca de B que de A. Por ende, habría sido más natural asignar el objeto i al clúster B, entonces casi que se puede concluir que este objeto ha sido mal clasificado (Rousseeuw, 1987).

En conclusión, $s(i)$ mide que tan bien un objeto ha sido clasificado en un clúster (Rousseeuw, 1987).

Para el caso de las similaridades, se hace unas pequeñas modificaciones así: se define $a'(i)$ y $d'(i)$ como las correspondientes similaridades promedio y se define $b'(i)$:

$$b'(i) = \underset{C \neq A}{maximo\ d'(i, C)}$$

Los números $s(i)$ son dados por:

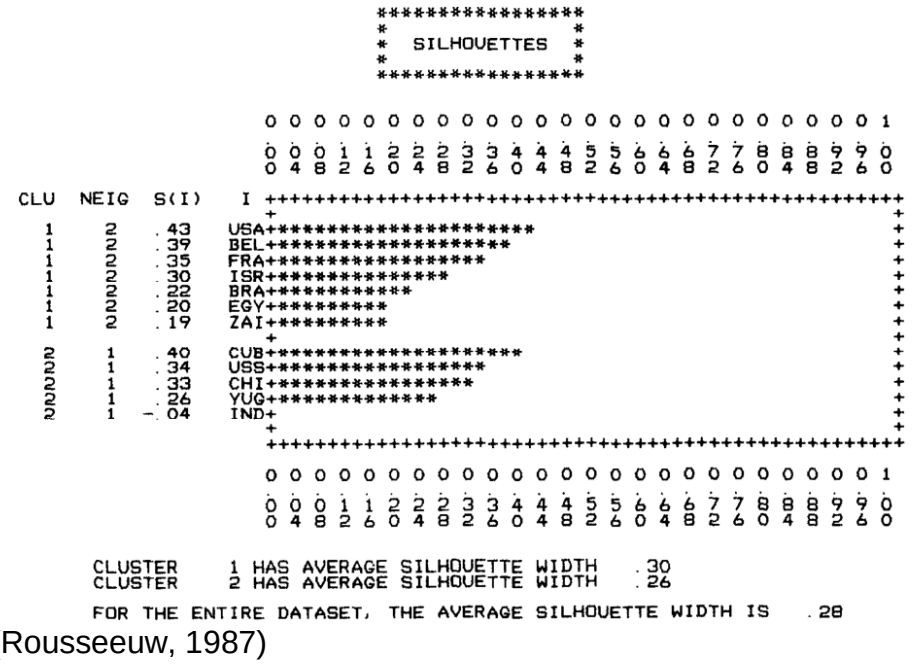
$$s(i) = \begin{cases} 1 - b'(i)/a'(i) & si\ a'(i) > b'(i) \\ 0 & si\ a'(i) = b'(i) \\ a'(i)/b(i) - 1 & si\ a'(i) < b'(i) \end{cases}$$

Y se interpretan de la misma manera que como antes.

Luego de que se han calculado todos los valores $s(i)$, entonces se pueden representar en un gráfico.

A continuación se muestra un ejemplo de una gráfica del coeficiente de la silueta que permitirá entender mejor este proceso, ver figura 10.

Figura 10. Ejemplo1 de una gráfica con el coeficiente de la silueta



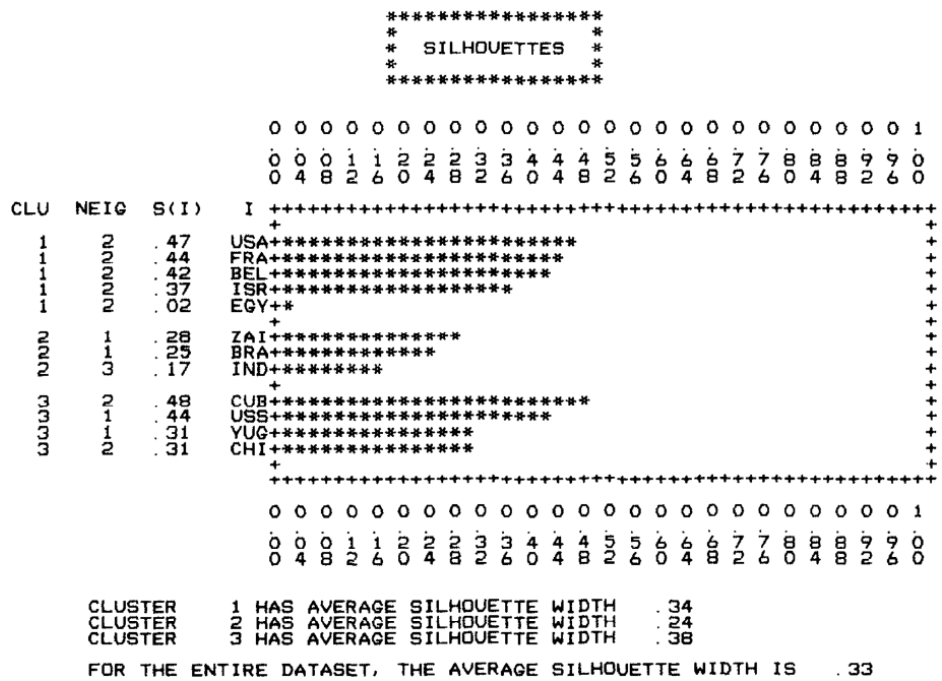
Fuente: (Rousseeuw, 1987)

En la gráfica se puede ver 2 clústeres de una base de datos de unos países. Horizontalmente, el valor del coeficiente de la silueta se mide desde 0 hasta 1, entre más grande sea este valor, más calidad del elemento en un clúster, también se puede ver en la parte inferior el coeficiente de la silueta para cada clúster y para todo el clustering en general.

La primera columna CLU, muestra el clúster al que pertenece el elemento, la segunda NEIG, muestra el vecino más cercado al elemento, la tercera S(I), es el valor del coeficiente de la silueta para el elemento, y la cuarta muestra la etiqueta del elemento.

La interpretación de la figura anterior es la siguiente, en general, la silueta es baja ($s = 0.28$), lo cual indica una clasificación pobre. El elemento USA (Estados Unidos) se clasificó sin duda en el clúster 1, que se podría decir son países occidentales industrializados, los 3 últimos elementos del clúster 1: BRA (Brasil), EGY (Egipto) y ZAI (Zaire, ahora parte de la República del Congo) son países en desarrollo que están al final del clúster 1 pues su valor de la silueta es menor. Todo esto indica que se puede intentar ejecutar el clústering con 3 grupos y mirar si el coeficiente de la silueta aumenta. Mírese la siguiente figura.

Figura 11. Ejemplo2 de una gráfica con el coeficiente de la silueta



Fuente: (Rousseeuw, 1987)

La figura 11 muestra el clustering en 3 grupos de la base de datos mencionada anteriormente. Se puede notar que el valor de la silueta en general es 0.33 lo cual es mejor que el anterior aunque no tan alto como se desearía. Los clústeres 1 y 3 están claramente definidos, el clúster 2 parece oscilar entre los clústeres 1 y 3.

De esta forma, variando el número de clústeres (k) en el proceso de segmentación y analizando el coeficiente de la silueta, puede ayudar a obtener resultados con mejor calidad.

Finalmente, para determinar la calidad de este coeficiente, se utiliza la tabla 13:

Tabla 13: Rango de valores coeficiente Silueta

Coeficiente Silueta	
Valor coeficiente	**Significado calidad**
-1 hasta 0.2	Mala
0.3 hasta 0.5	Aceptable
Mayor que 0.5 hasta 1	Buena

Fuente: este trabajo

5.1.11 Desarrollo Usando la Metodología CRISP-DM

5.1.11.1 Compresión del negocio

Mirar la sección **5.1.3.1 Comprensión del Negocio**.

5.1.11.2 Compresión de los datos

Inicialmente y gracias al repositorio de Datos de la Universidad de Nariño, se dispone de un conjunto de datos con 15504 registros que se extienden desde el primer semestre del año 2010 (2010-A) hasta el segundo semestre del año 2016 (2016-B). Se selecciona aquellos registros que sólo están entre los periodos 2010-A hasta 2014-B que es el periodo de tiempo que interesa a este trabajo de investigación, ver la sección **4.1 POBLACIÓN Y MUESTRA**. El resultado es un conjunto de datos con 10199 registros.

La tabla 14 es un resumen de frecuencias por facultad del repositorio anterior.

Tabla 14: Frecuencias de los estudiantes de pregrado matriculados modalidad presencial de la Universidad de Nariño (2010-A en adelante) entre los años 2010 y 2014

Facultad	**Frecuencia**	
	Cantidad	**Porcentaje**

Facultad	Frecuencia	
	Cantidad	Porcentaje
ARTES	1432	14.04059
CIENCIAS AGRICOLAS	793	7.77527
CIENCIAS DE LA SALUD	490	4.80439
CIENCIAS ECONOMICAS Y ADMINISTRATIVAS	1391	13.63859
CIENCIAS EXACTAS Y NATURALES	1110	10.88342
CIENCIAS HUMANAS	1602	15.70742
CIENCIAS PECUARIAS	662	6.49083
DERECHO	547	5.36327
EDUCACION	514	5.03971
INGENIERIA	1283	12.57966
INGENIERIA AGROINDUSTRIAL	375	3.67683
Total	**10199**	**100**

Fuente: este trabajo

- **Variables Independientes o Explicativas**

Seguidamente, el equipo de trabajo se familiariza con las variables explicativas del dataset, para esto se crea el siguiente diccionario de datos que describe las diferentes variables y sus posibles valores. Ver tabla 15.

Las variables finales que se utilizarán en el análisis están listadas más adelante en la sección **5.1.11.2 Preparación de los datos/Obtención de la vista minable**.

Tabla 15: Diccionario de datos de las variables explicativas

N°	Variables		Dominio
	Tipo	Nombre	
1	Personales	fecha_nacimiento	Indica la fecha de nacimiento del estudiante.
2		anios	Edad de la persona
3		sexo	Sexo del estudiante.
4	Demográficas	ciudad_nacimiento	Ciudad de Nacimiento
5		ciudad_per	Ciudad de donde proviene el estudiante
6		Ciudad	Ciudad Sede de la Universidad
7		barrio	Barrio donde el estudiante vive al momento de matricularse a la Universidad.
8	Socioeconómicas	Estrato	Estrato socioeconómico
9		jefe_familia	Jefe de Familia
10		mas_una_cargo	Más de una persona a cargo
11		tipo_residencia	Tipo de residencia
12		vive_con_familia	Vive con familia
13		no_hermano_univers	Hermanos que se encuentran en la Universidad
14		ingresos_familiare	Ingresos Familiares
15		ano_ingresos	Año en el que se reporta el ingreso
16		valor_matric_coleg	Valor que pagó el estudiante en su último año de

			colegio.
17		ano_pago_colegio	Año en el que se realizó el último pago de matrícula del colegio.
18		pago_contado	Si el pago de la matricula a la Universidad se hizo de contado
19		Puntaje	Puntaje que la universidad calcula para determinar el costo de la matrícula
20	Antecedentes Escolares	p_p_total	Puntaje total ICFES con el que ingreso a la Universidad
21		tipo_de_colegio	Tipo de colegio de donde proviene el estudiante
22		biologia	Puntaje obtenido en la prueba ICFES saber once en el área de Biología
23		matematicas	Puntaje obtenido en la prueba ICFES saber once en el área de Matemáticas
24		filosofia	Puntaje obtenido en la prueba ICFES saber once en el área de Filosofía
25		fisica	Puntaje obtenido en la prueba ICFES saber once en el área de Física
26		quimica	Puntaje obtenido en la prueba ICFES saber once en el área de Química
27		lenguaje	Puntaje obtenido en la prueba ICFES saber once en el área de Lenguaje
28		idiomas	Puntaje obtenido en la prueba ICFES saber once en el área de Idiomas
29		sociales	Puntaje obtenido en la prueba ICFES saber once en el área de Sociales
30	Universitarias	detalle_especial1	Ingreso a la Universidad con cupo especial
31		estado_actual	Estado Actual del Estudiante
32		periodo_egreso	Periodo en el que el estudiante egresa de la Universidad. Comprende desde el semestre A del 2012 sin incluir el semestre B del 2012 hasta el semestre A del 2017
33		fecha_grado	Fecha de Grado
34		periodo_grado	Periodo en el cual el estudiante se graduó
35		vigente_actualmente	Indica si el estudiante se encuentra vigente en la universidad
36		semestre_actual	Último semestre cursado
37		anios_ingreso	Indica el año en que el estudiante ingresa a la Universidad.
38		código	De 1 hasta el 10199
39		cod_carrera	Código de la Carrera. Identificador
40		cod_facultad	Código de la Facultad. Identificador
41		nombre_carrera_cor	Nombre de la carrera
42		nombre_facultad_cor	Nombre la Facultad
43		esta_vigente	Está Vigente
44		periodo_academico	Periodo Académico. Desde 2010A hasta 2014B

Fuente: Repositorio de Datos de la Universidad de Nariño

Ver los valores posibles para las diferentes variables en el **ANEXO A. TABLAS AUXILIARES DICCIONARIO DE DATOS**.

- **Variable Dependiente o Explicada**

Para el modelo predictivo

Ya que en este trabajo de investigación se desea predecir el rendimiento académico valiéndose del promedio acumulado entonces la variable **promedio_acumulado**, que es de tipo numérico; se selecciona como la variable dependiente y se define así: “El promedio general acumulado y el semestral o anual de calificaciones de un estudiante, será el que resulte de calcular el promedio aritmético de todas las notas registradas, tomado en unidades, décimas y centésimas” (Nariño, 1998).

Para el modelo de segmentación

Las categorías que se utilizan para interpretarlo son: (Álvarez y García, 1996):

- Muy Bajo: cuando el promedio es inferior a 3.0
- Bajo: cuando el promedio está entre 3.0 y 3.49
- Medio: cuando el promedio está entre 3.5 y 3.99
- Alto: iguales o superiores a 4.

Estas categorías se proponen en el trabajo investigativo: Factores que Predicen El Rendimiento Universitario, de los autores María Teresa Álvarez y Hernán García, quienes hacen esta categorización teniendo en cuenta el puntaje del examen de estado y el rendimiento promedio del estudiante en el bachillerato.

5.1.11.3 Preparación de los datos

- **Obtención de la vista minable**

Para esta fase del trabajo de investigación no fue posible proceder con una tarea de imputación de los datos faltantes en las diferentes variables por las siguientes razones: La universidad no tiene otros repositorios de datos en los que haya más información socioeconómica de los estudiantes, de haber necesitado esto, se habría tenido que enviar una solicitud de averiguación a todos y cada uno de los estudiantes objeto de estudio de este trabajo. Sólo se permitió el acceso a la información socioeconómica contenida en la base de datos descrita en la sección:
5.1.11.1 Compresión de los datos

Inicialmente, el repositorio de datos contenía 15504 registros de los estudiantes de pregrado matriculados de todos los programas que ofrece la Universidad de Nariño. De estos se seleccionaron aquellos de los periodos académicos 2010-A a

2014-B resultando en 10199 ejemplos. El periodo 2010 a 2014 es el que interesa en este trabajo de investigación.

El conjunto de datos obtenido anteriormente se convierte en el repositorio base sobre el que se ejecuta una primera fase de exploración, organización y limpieza de datos, así:

Renombramiento de atributos

cod facultad a **cod_facultad**
nombre facultad a **nombre_facultad**

Creación de nuevos atributos con fines descriptivos

nombre_carrera_cor que reemplaza a la variable **nombre_carrera_lar**
nombre_facultad_cor que reemplaza a la variable **nombre de facultad**

Corrección de datos

La variable **barrio** contiene el nombre del barrio en el que vive el estudiante, se corrigió el dato de un barrio de **20-jul** a **20 de Julio** para algunos registros, ya que el dato **20-jul** se puede confundir con la fecha que representa.
Limpieza de datos

1. Eliminación de 27 registros de la variable **fecha_grado** del 2012 y 2013 porque son estudiantes que iniciaron sus estudios antes del 2010.

2. Se eliminaron 377 registros cuyo valor es “No informa” para la variable **estrato**. Es de interés para este trabajo conocer con precisión el estrato socioeconómico del estudiante porque puede influir en el rendimiento académico.

3. Se eliminó el atributo **tipo_exencion** porque no tiene incidencia en la investigación ya que interesa analizar todos los estudiantes sin importar si fueron exentos o no.

4. Se eliminaron 352 registros de la variable **tipo_residencia** porque su valor es “No informa”.

5. De la variable **p_p_total** se eliminaron 41 registros. 40 de ellos con dato 0 y el otro con dato nulo.

6. Se eliminó un registro para la variable **anios_ingreso** porque su valor 7 es atípico.

7. De la variable **promedio_acumulado** se filtraron 108 registros porque tenían dato 0 o vacío.

Eliminación de otros atributos

Se eliminaron los atributos: **lecturacritica**, **ciencias**, **razonamiento**, **competenciaciud** porque no tienen datos para ningún registro.

Luego de haber ejecutado los pasos de limpieza y organización de datos anteriores, se dispone de un repositorio de datos con 9293 registros.

La figura 9 muestra diferentes estadísticas para los atributos resultantes.

Seguidamente, para este nuevo repositorio, se ejecutaron las siguientes tareas de limpieza y organización de datos:

1. En figura 12, se puede observar que las variables: **fecha_grado**, **periodo_egreso**, **periodo_grado** tienen un alto porcentaje de valores nulos y además están correlacionadas entre ellas por lo tanto se decide suprimir estos atributos, además que estas variables no son predictoras.

2. Para 274 registros, se promedió las variables: **geografia** e **historia** y este valor se ubicó en la variable **sociales** porque la forma de evaluación de las materias de Historia y Geografía cambió en el ICFES del 2014, entonces sólo se calcula un valor promedio que se resume en la variable sociales.

3. Se eliminaron las variables **geografía** e **historia** porque su valor se generaliza en la variable **sociales**.

4. Se eliminaron 204 registros de las variables **vive_con_familia** y **no_hermano_univers** porque no reportaron ningún dato además no se pudo imputar o reemplazar el valor para estos atributos y de esos registros porque suponía una averiguación de su información y situación personal y familiar.

Figura 12. Diferentes estadísticas para los atributos resultantes del repositorio de datos con 9293 registros

Grupo	Variable	Mínimo	Maximo	Media	Desviacion Estandar	Varianza	Asimetria	Kurtosis	Suma General	Datos Perdidos	Porcenaje Perdidos
Personales	Anios	18	84	24.05	4.01	16.1	4.21	40.06	223526	0	0
	fecha_nacimiento									0	0
	Sexo									0	0
Demográficas	Barrio									356	3.83
	Ciudad									0	0
	ciudad_nacimiento_lm									0	0
	ciudad_per_lm									0	0
Socioeconómicas	ano_ingresos	0	34	30.51	3.54	12.53	-7.12	58.91	283544	0	0
	ano_pago_colegio	0	99	29.85	8.87	78.69	5.5	42.93	277363	0	0
	Estrato									0	0
	ingresos_familiare	0	261020000	9056690.45	9132903.97	8.34099E+13	6.31	96.01	84163824341	0	0
	jefe_familia									0	0
	mas_una_cargo									1	0.01
	no_hermano_univers									204	2.2
	pago_contado	0	4140331	228033.44	211644.73	44793493086	4.41	34.69	2119114804	0	0
	Puntaje	0	100	12.18	7.57	57.24	3.2	20.69	113171	0	0
	tipo_residencia									0	0
	valor_matric_coleg	0	2308500	47337.16	70236.56	4933174321	6.66	149.9	439904197	0	0
	vive_con_familia									204	2.2
Antecedentes Escolares	Biologia	0	99	55.38	8.7	75.68	0.59	1.33	514623.28	0	0
	Filosofia	0	96.7	51.82	9.37	87.73	0.26	1.36	481537.35	0	0
	Fisica	0	110	54.42	10.31	106.37	0.56	1.51	505693.88	0	0
	Idiomas	0	116.95	51.27	10.71	114.61	0.88	1.66	476467.11	0	0
	Lenguaje	0	94.52	56.98	7.68	59.02	0.32	1.05	529508.86	0	0
	matematicas	0	122	59.94	12.41	154.1	0.84	2.03	557030.78	0	0
	Quimica	0	118.8	55.67	9.83	96.61	0.91	2.11	517316.66	0	0
	Sociales	0	107	55.34	8.01	64.2	0.09	0.76	514269.545	0	0
	p_p_total	26.51	119.78	58.89	7.04	49.53	0.53	1.33	547246.84	0	0
	tipo_de_colegio									0	0
Universitarias	anios_ingreso	15	80	19.22	3.76	14.1	5	51.41	178632	0	0
	detalle_especial1									0	0
	esta_vigente									0	0
	estado_actual									0	0
	fecha_grado									8423	90.64
	nombre_carrera_cor									0	0
	nombre_facultad_cor									0	0
	periodo_academico									0	0
	periodo_egreso									7736	83.25
	periodo_grado									8423	90.64
	promedio_acumulado	**0.02**	**4.81**	**3.33**	**0.9**	**0.82**	**-1.43**	**1.83**	**30988.99**	**0**	**0**
	semestre_actual	1	110	15.95	30.46	927.67	2.71	5.53	148226	0	0
	vigente_actualmente									0	0

Fuente: este trabajo

Número total de registros hasta este punto: 9293

5. Se cambiaron 148 registros de la variable **especial** del valor guion (-) a vacío.

6. Se eliminaron las variables **esta_vigente** y **vigente_actualmente** porque el propósito de este trabajo es analizar el rendimiento académico a partir del promedio acumulado en ese momento del pasado.

7. Se eliminaron 356 registros que no tienen ningún valor diligenciado para la variable **barrio**.

Eliminación de Claves Candidatas

La variable **codigo** (identificador interno del estudiante) se reemplazó por un número consecutivo que sirva como identificador del registro más no del estudiante.

Se eliminaron los campos **cedula**, **cedula-1**, **nombres**, **apellidos**, **nombres-1**, **telefono**, **direccion-pasto**, **cod_icfes**, **ano_ingresos**, **puntaje**,

ano_pago_colegio ya que son atributos identificadores irrelevantes para esta investigación.

Generalización

Se creó una variable llamada **ciudad_nacimiento_lm** que generaliza los valores de la variable **ciudad_nacimiento** así: IPIALES, OTRAS, PASTO, TUMACO, TUQUERRES.

Se creó una variable llamada **ciudad_per_lm** que generaliza los valores de la variable **ciudad_per** así: IPIALES, OTRAS, PASTO, TUMACO, TUQUERRES.

Limpieza de valores atípicos

Este paso en la limpieza de los datos sólo se aplica a variables numéricas y sólo se elimina los registros que la prueba notifica, por ejemplo, en la tabla 16 se eliminan 10 registros porque estos resultaron atípicos, pero en la tabla 17 se eliminan 5 registros por la misma razón.

A continuación, se muestra las variables que fueron objeto de esta limpieza.

Variable **ingresos_familiare**

La tabla 16 muestra los casos de datos extremos (atípicos) para esta variable.

Tabla 16: Valores Extremos Variable **ingresos_familiare**

Valores Extremos					
			Caso Número	Código	Valor
ingresos_familiare	Más alto	1	8093	8831	261020000
		2	7771	8439	159832486
		3	2188	2415	141407000
		4	3055	3298	141407000
		5	8042	8777	117715000
	Más bajo	1	903	1001	0
		2	287	302	9
		3	3164	3409	10
		4	1030	1181	2010
		5	6731	7236	10000

Fuente: este trabajo

Claramente, se puede ver que hay varios registros con datos atípicos, esto puede influir de manera negativa en la construcción de un modelo de minería de datos, por lo tanto estos registros serán eliminados. En total se eliminan 10 ejemplos.

Variable **valor_matric_coleg**

El mismo procedimiento descrito para la variable anterior, se aplica a esta. A continuación, se presenta la tabla 17 con los valores atípicos.

Tabla 17: Valores Extremos Variable **valor_matric_coleg**

Valores Extremos					
			Caso Número	Código	Valor
valor_matric_coleg	Más alto	1	697	764	2308500
		2	6176	6641	1500000
		3	789	885	1260000
		4	6926	7442	1000068
		5	3634	3894	540000
	Más bajo				

Fuente: este trabajo

Como antes, los datos extremos se eliminan, 5 en total.

Variable **pago_contando**

A continuación se presenta la tabla 18 con los datos extremos para la variable **pago_contado**.

Tabla 18: Valores Extremos Variable **pago_contando**

Valores Extremos					
			Caso Número	Código	Valor
pago_contado	Más alto	1	8093	8831	4140331
		2	8034	8769	2568162
		3	7789	8457	2458124
		4	3970	4248	2411527
		5	7771	8439	2342199
	Más bajo				

Fuente: este trabajo

Como antes, los datos extremos se eliminan, 5 en total.

Luego de haber ejecutado esta fase de limpieza, corrección y eliminación de datos, se dispone de un repositorio de datos listo al que se puede aplicar diferentes algoritmos o técnicas de minería de datos y que cuenta con 8716 registros.

Las variables que se utilizarán en las diferentes técnicas de minería de datos son:

- Nombre Carrera Corto: nombre_carrera_cor
- Ciudad
- Jefe Familia: jefe_familia
- Mas Una Cargo: mas_una_cargo
- Estrato
- Tipo Residencia: tipo_residencia
- Vive Con Familia: vive_con_familia
- Ingresos Familiares: ingresos_familiares
- Valor Matrícula Colegio: valor_matric_coleg
- Pago Contado: pago_contado
- Ciudad Permanencia LM: ciudad_per_lm
- Edad de ingreso a la universidad: anios_ingreso
- Años: anios
- Puntaje Total: p_p_total
- Sexo
- Tipo Colegio: tipo_colegio
- Promedio Acumulado (Objetivo): promedio_acumulado
- Biología
- Matemáticas
- Filosofía
- Física
- Química
- Lenguaje
- Idiomas
- Sociales
- Semestre Actual: semestre_actual
- **Caracterización**

A continuación, se hace una caracterización de los estudiantes en relación a las variables personales, socioeconómicas, académicas e institucionales que resultaron luego del proceso de preparación de los datos y que son más representativas.

Variables Personales

Sexo

El 41.33% de los registros corresponden a mujeres y el 58.66% son hombres. En las siguientes facultades, hay mayor presencia masculina: facultad de derecho: 54%, facultad de ciencias económicas y administrativas: 53%, facultad de ciencias agrícolas: 56%, facultad de ingeniería agroindustrial 58%, y hay un porcentaje mucho mayor de estudiantes masculinos en las facultades de ingeniería: 82%, facultad de artes: 72% y facultad de ciencias exactas y naturales: 67%.

En las siguientes facultades, hay mayoría de estudiantes femeninas: facultad de educación: 59%, ciencias humanas 53%, ciencias de la salud 59%.

La única facultad que tiene una proporción casi igual de hombres y mujeres es: facultad de ciencias pecuarias con 50.5% para hombres y 49.5% para mujeres.

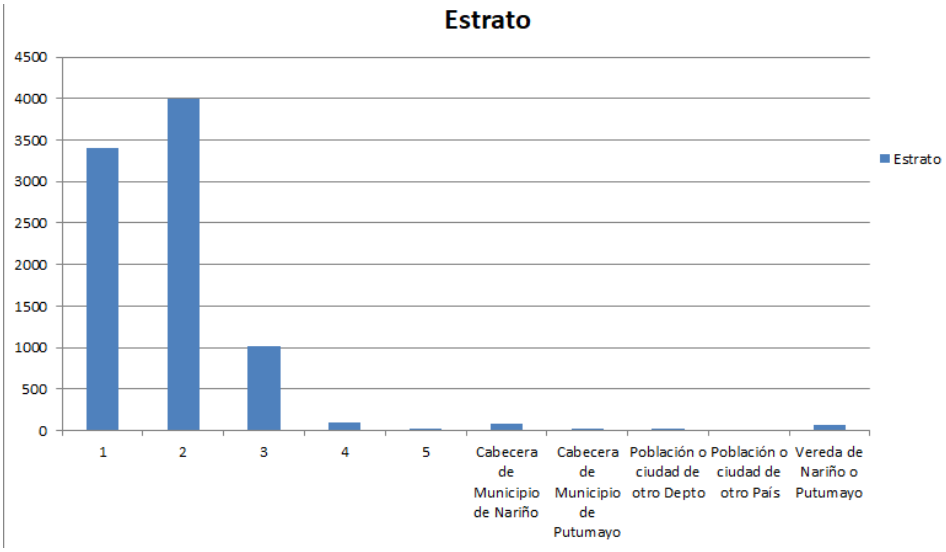
Variables Socioeconómicas

Estrato

El 84.90% de los estudiantes pertenecen a los estratos 1 y 2; el 11.61% están en el estrato 3 y el resto de estudiantes, el 3.48%; son de los otros estratos, ver figura 13.

Las facultades que concentran el mayor número de estudiantes en estratos 1 y 2 son: la facultad de educación con el 93.24%, la facultad de ciencias de la salud: 88.38%, y la facultad de ciencias humanas: 88.02%.

Figura 13. Gráfica de la distribución del estrato



Fuente: este trabajo

Tipo Residencia

Tabla 19: Variable **tipo_residencia** por facultad

Facultad	Arrendada/Anticresada	No es propia	Propia	Propia/paga cuotas
Derecho	23.01%		71.89%	5.09%
Facartes	21.11%	0.08%	74.38%	4.43%
Facea	26.25%		69.88%	3.87%
Facedu	19.74%		78.70%	1.56%
Facia	18.72%		76.68%	4.6%
Faciagro	24.57%		72.84%	2.59%
Facien	20.3%		77%	2.7%
Facihu	20.28%	0.07%	77.37%	2.28%
Facipec	18.98%		78.98%	2.03%
Facsalud	21.64%		75.17%	3.19%
Ingenieria	23.60%		73.03%	3.37%

Fuente: este trabajo

Se puede ver en la tabla 19 que la mayoría de los estudiantes en todas las facultades tienen casa propia, una relativa minoría arrienda, un pequeño porcentaje tiene casa propia pero aún la están pagando. Sólo en las facultades de Artes y Ciencias Humanas hay estudiantes que no tienen casa propia aunque su porcentaje es muy pequeño.

Variables Antecedentes Escolares

Puntaje Total

Tabla 20: Variable **p_p_total** promediada por carrera

CARRERA	PROMEDIO p_p_total
INGENIERIA EN PRODUCCION ACUICOLA	51.7
ZOOTECNIA	53.1
LICENCIATURA EN LENGUA CASTELLANA Y LITERATURA	53.1
LICENCIATURA EN FILOSOFIA Y LETRAS	53.2
GEOGRAFIA	53.9
TECNOLOGIA EN COMPUTACION	53.9
LICENCIATURA EN INFORMATICA	54.4
LICENCIATURA EDUCAC.BASICA ENFASIS CIENCIAS NATURALES-EDUC.AMBIENTAL	54.5
TECNOLOGIA EN PROMOCION DE LA SALUD	55
MERCADEO	55.3
LICENC. EN EDUC.BASICA CON ENFASIS EN HUMANIDA. LENGUA CASTEL E INGLES	55.7
SOCIOLOGIA	56.1
COMERCIO INTERNACIONAL	56.2

ADMINISTRACION DE EMPRESAS	56.3
LICENCIATURA EN EDUCACION BASICA CON ENFASIS EN CIENCIAS SOCIALES	56.8
INGENIERIA AGRONOMICA	57.1
INGENIERIA AGROINDUSTRIAL	57.8
FISICA	57.9
ECONOMIA	57.9
DISEÑO INDUSTRIAL	57.9
INGENIERIA AGROFORESTAL	58.1
LICENCIATURA EN MATEMATICAS	58.1
DISEÑO GRAFICO Y MULTIMEDIAL	58.5
INGENIERIA DE SISTEMAS	59.1
PSICOLOGIA	59.5
MEDICINA VETERINARIA	59.7
LICENCIATURA EN INGLES-FRANCES	60.4
QUIMICA	61.2
DERECHO	61.9
INGENIERIA ELECTRONICA	62
CONTADURIA PUBLICA	62.3
INGENIERIA AMBIENTAL	62.7
INGENIERIA CIVIL	63.2
BIOLOGIA	63.5
LICENCIATURA EN ARTES VISUALES	64.2
ARTES VISUALES	65.5
LICENCIATURA EN MUSICA	67.2
ARQUITECTURA	69.4
MEDICINA	71.1

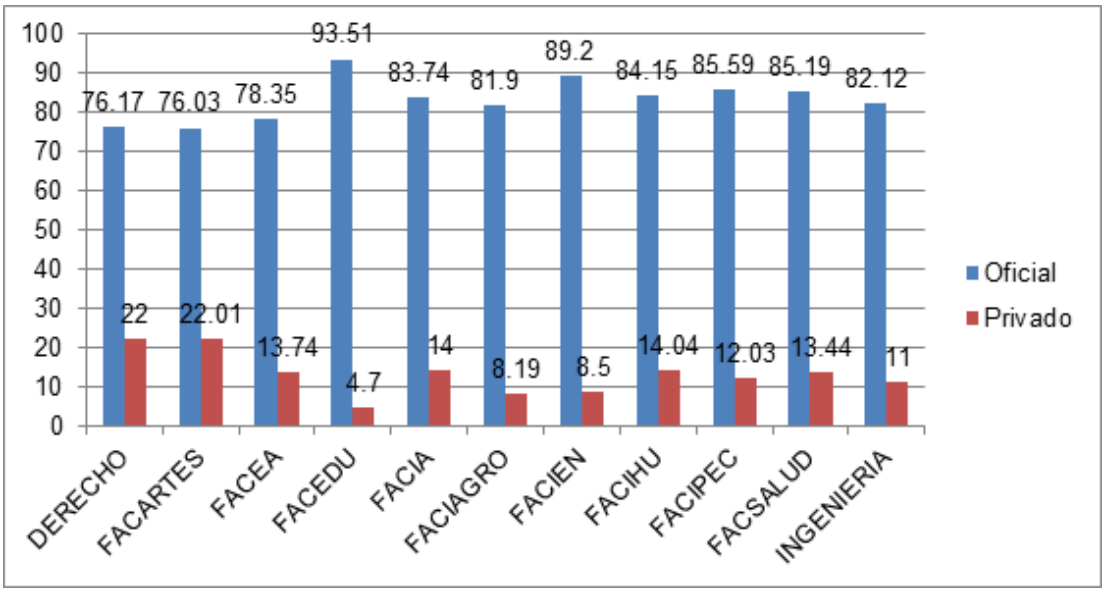
Fuente: este trabajo

El mayor promedio para las carreras técnicas es para Medicina con 71.1, mientras que para las humanísticas es Licenciatura en música con 67.2. Los promedios más bajos son para las carreras técnicas: Ingeniería en producción acuícola y Zootecnia, para las humanísticas es Licenciatura en Lengua Castellana y Literatura. No se puede detallar un dominio de un tipo de carrera sobre otro, ver tabla 20.

Tipo Colegio

En general, el 85.54% de los estudiantes provienen de colegios oficiales y el 14.46% son de colegios privados. Para todas las facultades, la mayoría de los estudiantes son de colegios oficiales seguidos de colegios privados, cabe resaltar que en la facultad de Educación el 93,50% son estudiantes de colegios oficiales, la gran mayoría, ver figura 14.

Figura 14. Gráfica de la variable tipo_colegio

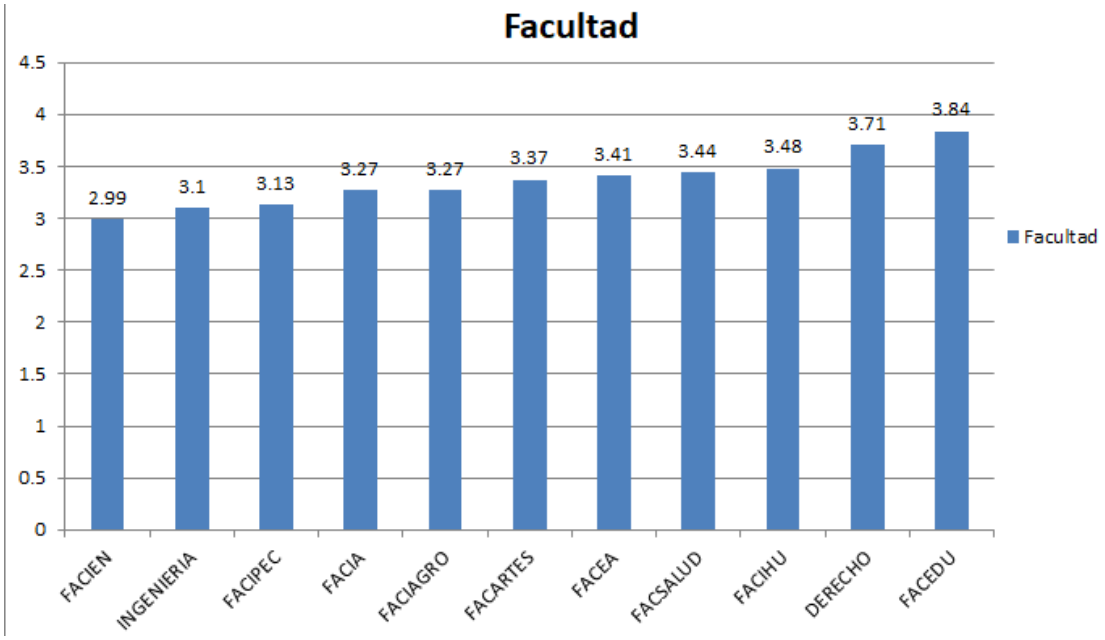


Fuente: este trabajo

Variables Universitarias

promedio_acumulado

Figura 15. Gráfica de la variable promedio_acumulado



Fuente: este trabajo

El promedio más bajo se registra para la facultad de Ciencias Exactas y Naturales seguida de la facultad de Ingeniería, este promedio está por debajo de la nota mínima aprobatoria (Nariño, 1998), el promedio más alto es para la facultad de Educación, seguida por Derecho, ver figura 15.

Variables Socioeconómicas

jefe_familia

Tabla 21: Tabla de la variable **jefe_familia**

Facultad	Inscrito	Madre	Otro	Padre
Derecho	5%	41%	12%	42%
Facartes	4%	40%	12%	44%
Facea	4%	38%	9%	48%
Facedu	5%	34%	13%	48%
Facia	2%	41%	11%	46%
Faciagaro	2%	34%	5%	59%
Facien	2%	40%	10%	48%
Facihu	5%	38%	12%	46%
Facipec	2%	37%	9%	53%
Facsalud	3%	38%	10%	49%
Ingenieria	2%	33%	8%	57%

Fuente: este trabajo

Se puede ver que para todas las facultades, los jefes de familia *padre* y *madre* son mayoría, las categorías *otro* e *inscrito* tienen un menor porcentaje, ver tabla 21.

5.1.11.4 Modelado

- **Selección de la técnica para el modelo de predicción**

Ya que el objetivo es determinar cuáles de las variables independientes predicen mejor el rendimiento académico (**promedio_acumulado**), se ejecutó una prueba para determinar qué modelo es el mejor en términos de cuál arroja la mejor correlación con el menor error relativo y el menor número de campos usados, ver tabla 22.

Tabla 22: Comparación de prueba de técnicas de predicción

Técnica	Correlación	# Campos Usados	Error Relativo
SVM	0.919	25	0.159
CHAID	0.776	12	0.398
Lineal Generalizado	0.771	26	0.406
Red Neuronal	0.768	26	0.411

C&RT	0.748	21	0.441

Fuente: este trabajo

La prueba consistió en ejecutar cado uno de los algoritmos listados en la tabla anterior en un flujo alimentado con las variables independientes finales mencionadas en la sección **5.1.11.2 Obtención de la vista minable** y la variable dependiente continua (**promedio_acumulado**), anotar el resultado de su correlación, número de campos usados y el error relativo, organizar los algoritmos teniendo en cuenta como criterio, primero, la correlación, segundo, el error relativo, y tercero, el número de campos usados. El flujo se creó y ejecutó en el software SPSS Modeler versión 18.

De los resultados anteriores, se puede ver claramente que las Máquinas de Vectores de Soporte tienen la mejor correlación (91.9%), y el menor error relativo (15.9%) aunque el número de campos utilizados sea uno de los mayores (25). Para entender mejor esta situación, (Kantardzic, 2011) aclara que una correlación positiva alta muestra una relación fuerte entre la variable objetivo y las variables explicativas más importantes, (IBM, 2016) dice que un error relativo bajo indica una mejor precisión del modelo y que entre menos campos se utilicen para construirlo, se puede tener mejor rendimiento y generalización cuando se esté probando con instancias nuevas, entonces a pesar de que las máquinas de vectores de soporte usan un buen número de campos para construir el modelo, estas tienen la mejor correlación y de igual forma el error relativo.

Vale la pena recordar que no se puede tener una medida de rendimiento como el Error de Clasificación pues la tarea de este modelo predictivo es sobre una variable dependiente continua (para regresión) y no sobre una variable dependiente categórica (para clasificación), además que para medir la bondad de ajuste entre el promedio acumulado observado y el promedio acumulado predicho se utiliza la correlación, entonces se selecciona el algoritmo que mejor correlación tenga con el menor error relativo por ende las máquinas de vectores de soporte son ese algoritmo.

Ya que a este trabajo de investigación interesa el rendimiento académico y usando como medida para este el promedio acumulado, se decide dividir el repositorio de datos en 2 grupos correspondientes a las carreras técnicas y a las humanísticas, esta división se realiza teniendo en cuenta la tabla de ponderaciones de las tarjetas ICFES del año 2006 al 2014-I, ver **Anexo B – Tabla de Ponderaciones con Tarjetas ICFES**, que establece la oficina de registro académico de la Universidad de Nariño, OCARA, con el porcentaje que cada una de las áreas debe tener para el ingreso a la carreras ofrecidas; en el caso de las carreras que se han clasificado como técnicas los departamentos a los que se encuentran adscritas estas carreras asignan un mayor porcentaje a las áreas de matemáticas, lenguaje, biología y química; para el caso de las carreras que se han clasificado como humanísticas, los departamentos académicos asignan un mayor porcentaje a las

áreas de lenguaje, ciencias sociales y filosofía. Esta división se propone dado que el rendimiento en una carrera técnica puede ser diferente, no mejor ni peor, que en una carrera humanística.

El repositorio de las carreras técnicas está compuesto por 5380 registros y lo integran las carreras: Administración de Empresas, Arquitectura, Biología, Comercio Internacional, Contaduría Pública, Diseño Gráfico y Multimedia, Diseño Industrial, Economía, Física, Geografía, Ingeniería Agroforestal, Ingeniería Agroindustrial, Ingeniería Agronómica, Ingeniería Ambiental, Ingeniería Civil, Ingeniería Electrónica, Ingeniería en Producción Acuícola, Ingeniería de Sistemas, Medicina Veterinaria, Medicina, Mercadeo, Química, Tecnología Computacional y Zootecnia.

El repositorio de las carreras humanísticas contiene 3336 registros y está compuesto por las carreras: Artes Visuales, Derecho, Licenciatura en Artes Visuales, Licenciatura en Educación Básica con Énfasis en Ciencias Sociales, Licenciatura en Educación Básica con Énfasis en Humanidad, Lengua Castellana e Inglés, Licenciatura en Educación Básica con Énfasis en Ciencias Naturales y Educación Ambiental, Licenciatura en Filosofía y Letras, Licenciatura en Informática, Licenciatura en Inglés y Francés, Licenciatura en Lengua Castellana y Filosofía, Licenciatura en Matemáticas, Licenciatura en Música, Psicología, Sociología y Tecnología en Promoción de la Salud.

- **Configuración del modelo de Máquina de Vectores de Soporte de Regresión**

En general, para los 2 repositorios, se configuró un modelo de máquinas de vectores de soporte de regresión para cada conjunto de datos así:

Las variables explicativas mencionadas al final de la sección **5.1.11.2 Obtención de las vista minable** se utilizaron como variables de entrada, la variable objetivo es **promedio_acumulado**, el kernel seleccionado fue **RBF Gaussiano**, debido a que la literatura lo recomienda como una primera elección ya que, por lo general; tiene buenos resultados en la mayoría de problemas y sólo tiene un parámetro para afinar (γ) por lo tanto, el tiempo requerido para mejorar será menor que seleccionando otro kernel como el polinómico (Kantardzic, 2011).

Como es usual en problemas de minería de datos, se configuró una validación cruzada de los repositorios usando 2 particiones, una de entrenamiento con el 70% de los datos y otra de prueba con el 30%. Así, se construye el modelo usando la primera partición y se prueba la generalización y el rendimiento con la segunda. La idea de usar validación cruzada permite evitar situaciones de sobreajuste (overfitting), este problema se origina cuando la estructura del

algoritmo utilizado se adhiere casi perfectamente a la estructura de los datos de la partición de entrenamiento de forma que cuando se evalúa con una instancia desconocida el modelo es incapaz de clasificarla ya que no puede generalizar correctamente.

- **Afinamiento de los parámetros del modelo para el repositorio de las carreras técnicas**

La tabla 23 muestra los diferentes valores que se asignaron al modelo para afinarlo y lograr un mejor rendimiento.

Tabla 23: Afinación de parámetros para la máquina de vectores de soporte para el repositorio de las carreras técnicas

Configuración	C	γ	Entrenamiento			Prueba		
			Error Medio	Error Medio Absoluto	Correlación	Error Medio	Error Medio Absoluto	Correlación
1	1	1.0	-0.066	0.192	0.941	-0.056	0.523	0.561
2	1	10	-0.091	0.204	0.942	-0.072	0.602	0.292
3	1	100	-0.096	0.21	0.941	-0.073	0.626	0.076
4	10	0.5	0.005	0.104	0.987	0.021	0.458	0.677

Fuente: este trabajo

La razón por la cual los parámetros C y γ tienen esos valores en la tabla anterior, es porque luego de haber revisado varios documentos relacionados con máquinas de vectores de soporte y haber visto qué valores usaron los autores, los más comunes son los que están en la mencionada tabla, por ejemplo ver a (Pachano, 2008). La misma razón se aplica al repositorio de las carreras humanísticas.

Se puede observar que a medida que se incrementa el valor para el parámetro gama (Kernel RBF) y se conserva el valor del parámetro C (costo de error) y considerando también que en el entrenamiento la correlación es muy buena, se puede ver que el modelo se sobreajusta ya que la correlación es baja en la partición de prueba y el error medio absoluto es muy alto para las 3 primeras configuraciones.

Si se permite que haya un costo de error un poco más grande y dejar el valor de gamma entre 10 y 0.1 (0.5) se puede lograr una correlación muy alta en el entrenamiento y mejorar notoriamente la correlación en la partición de prueba, no obstante, el modelo se sobreajusta un poco porque la correlación en la prueba no puede ser más alta.

- **Afinamiento de los parámetros del modelo para el repositorio de las carreras humanísticas**

Tabla 24: Afinación de parámetros para la máquina de vectores de soporte para el repositorio de las carreras humanísticas

Configuración	C	γ	Entrenamiento			Prueba		
			Error Medio	Error Medio Absoluto	Correlación	Error Medio	Error Medio Absoluto	Correlación
1	1	1.0	0.072	0.209	0.93	-0.034	0.533	0.508
2	1	10	-0.092	0.212	0.93	-0.055	0.596	0.261
3	1	100	-0.095	0.214	0.931	-0.051	0.62	0.115
4	10	0.5	0.014	0.106	0.985	0.059	0.465	0.674

Fuente: este trabajo

La situación que se presenta con el afinamiento de los parámetros del modelo para las carreras técnicas también sucede para las humanísticas y la mejor configuración que se encontró es la cuatro que provee la mejor correlación y el menor error medio absoluto todo esto con un poco de detrimento en el rendimiento y la generalización del modelo (sobreajuste), ver tabla 24.

- **Segmentación**

Para este paso, se creó una variable categórica a partir del promedio académico acumulado llamada: **categoría_promedio_acumulado**, el dominio de valores es el mismo que se utiliza en el capítulo 5, sección **5.1.11.1 Comprensión de los datos**, estos son: Muy Bajo, Bajo, Medio y Alto.

La técnica seleccionada para el objetivo de segmentación fue el algoritmo K-Means (K-medias) por su popularidad en el uso con grandes bases de datos y por su relativa facilidad.

El software utilizado para ejecutar el algoritmo K-Means fue SPSS Modeler en su versión 18.

La tabla 25 muestra los diferentes valores provistos a los parámetros del algoritmo K-Means en búsqueda de los mejores resultados posibles con los datos disponibles.

Tabla 25: Afinación de los parámetros para el algoritmo K-Means

Generación del modelo				Resultados		
Configuración	# Clusteres	Iteraciones	Tolerancia	Iteraciones Resultantes	Silueta	Error
1	5	20	0	20	0.2	0

2	5	20	0.1	6	0.2	0.067
3	5	20	0.3	3	0.2	0.278
4	4	20	0.3	4	0.2	0.239
5	4	20	0.1	6	0.2	0.09
6	4	20	0	20	0.2	0
7	3	20	0	20	0.2	0.001
8	3	50	0	50	0.2	0
9	3	20	0.1	8	0.2	0.026
10	3	20	0.3	3	0.1	0.274
11	3	20	0.01	9	0.3	0.01

Fuente: este trabajo

Inicialmente, la configuración contempló la creación de 5 clústeres (configuraciones 1, 2, 3), lográndose un buen error pero con un factor de silueta bastante bajo (no aceptable); se continuo con la reducción de clústeres generados a 4, con la esperanza de que estos abarquen las 4 categorías de la variable **categoria_promedio_acumulado** (configuraciones 4, 5, 6); y aunque hay resultados con errores bajos, el factor de silueta sigue siendo no aceptable; finalmente, se reduce en uno el número de clústeres (configuraciones 7, 8, 9, 10, 11) y esta vez, asignando a la Tolerancia un valor relativamente pequeño se logra un resultado con una silueta aceptable (0.3) y un error muy bajo.

Se sabe que un valor de la Silueta (calidad del modelo) de un resultado de clustering debe ser mayor o igual a 0.5 para considerarse bueno y que para este caso el máximo logrado es 0.3, el cual es aceptable; no se pudo lograr una calidad mayor con los datos disponibles (SPSS Modeler, 2016) .

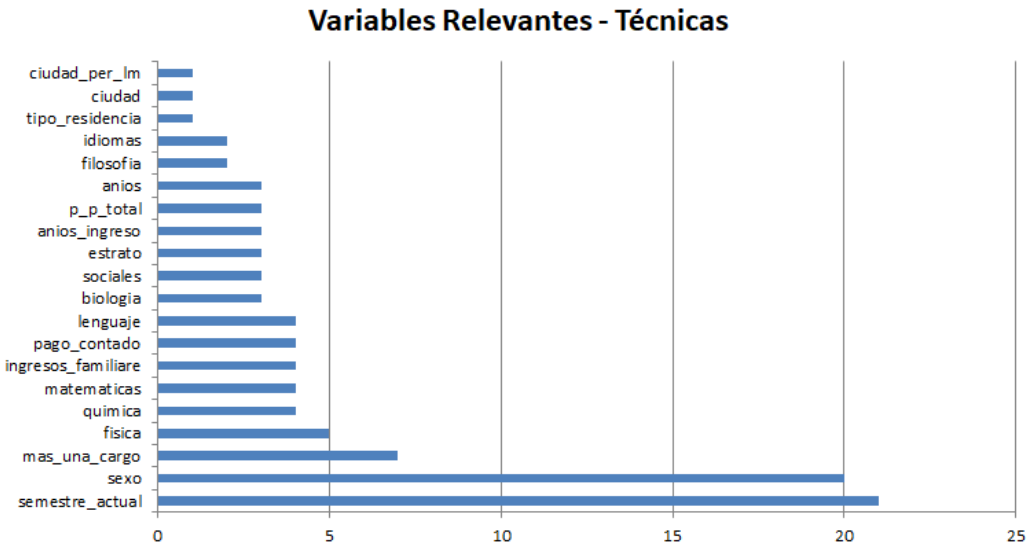
5.1.11.5 Evaluación

- **Evaluación del modelo para las carreras técnicas utilizando máquinas de vectores de soporte para regresión**

Variables influyentes

Las variables que influyen en la construcción del modelo según su importancia para el repositorio de las carreras técnicas se pueden ver en la figura 16.

Figura 16. Gráfica de las variables más influyentes en la construcción del modelo para las carreras técnicas



Fuente: este trabajo

La importancia de cada variable que se muestra en el gráfico anterior, no está relacionada con la precisión del modelo. Sólo está relacionada con la importancia de cada predictor a la hora de realizar una predicción no con si esta es precisa o no (IBM, 2016).

Las variables aparecen de esa forma en el gráfico anterior porque se tiene en cuenta la importancia relativa de cada una de ellas cuando se realiza una predicción. Mírese la siguiente tabla que tiene todas las variables más relevantes con su importancia relativa, aunque en el capítulo **6. RESULTADOS Y DISCUSIÓN** se hablará al respecto.

La importancia es relativa porque la sumatoria de esta para todas las variables es 1, mírese la tabla 26 importancia relativa de las variables.

Tabla 26: Tabla con la importancia relativa de las variables más relevantes para el repositorio de las carreras técnicas

Variable	Importancia Relativa %
semestre_actual	21
sexo	20
mas_una_cargo	7
fisica	5
quimica	4
matematicas	4
ingresos_familiares	4
pago_contado	4
lenguaje	4

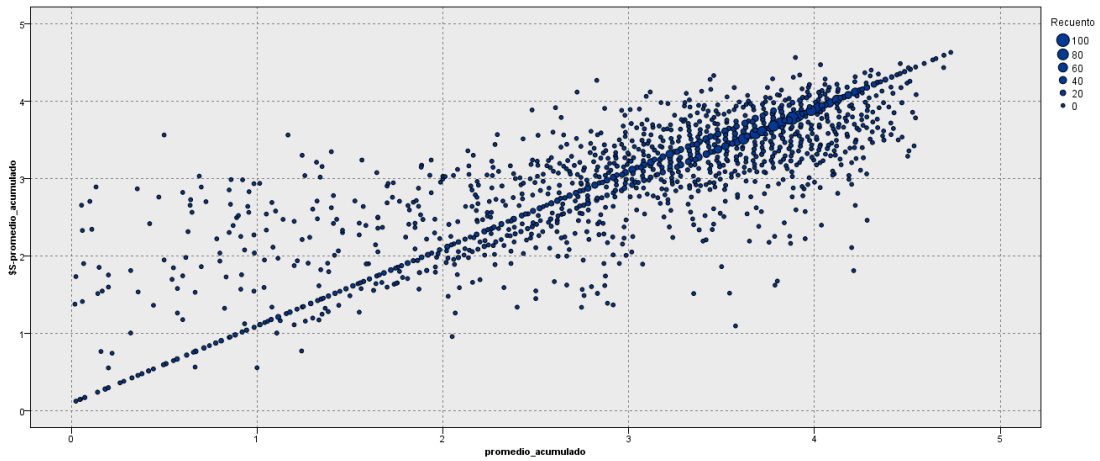
biologia	3
sociales	3
estrato	3
anios_ingreso	3
p_p_total	3
anios	3
filosofia	2
idiomas	2
tipo_residencia	1
ciudad	1
ciudad_per_lm	1

Fuente: este trabajo

De la figura 16 se puede ver que las variables: **semestre_actual** y **sexo** son por mucho las más importantes, y en su orden siguen: **física**, **química**, **matemáticas**, etc.

En la figura 17, se puede confirmar lo mencionado en el apartado anterior que concierne al sobreajuste del modelo. Hay una varianza mayor en la predicción de los valores para la variable objetivo menores o iguales que 3, en cambio se puede ver un poco más de uniformidad para los promedios mayores que 3.

Figura 17. Gráfica de la variable **promedio_acumulado observado** versus **promedio_acumulado predicho** carreras técnicas



Fuente: este trabajo

Esta figura se construye teniendo en cuenta los valores observados y los predichos para la variable dependiente, en este caso **promedio_acumulado**.

Entonces en el eje X están los valores observados y en el eje Y los predichos. El software SPSS Modeler versión 18 permite generar esta gráfica a través de un nodo.

- **Evaluación del modelo para las carreras humanísticas utilizando máquinas de vectores de soporte para regresión**

Variables influyentes - Variables influyentes

Como antes, para el repositorio de las carreras técnicas, las variables más relevantes para las carreras humanísticas aparecen gráficamente en la figura 18 de acuerdo a su importancia relativa, sin embargo en la tabla 27 están estos porcentajes.

Se hablará de esta importancia en el capítulo **6. RESULTADOS Y DISCUSIÓN**.

Tabla 27: Tabla con la importancia relativa de las variables más relevantes para el repositorio de las carreras humanísticas

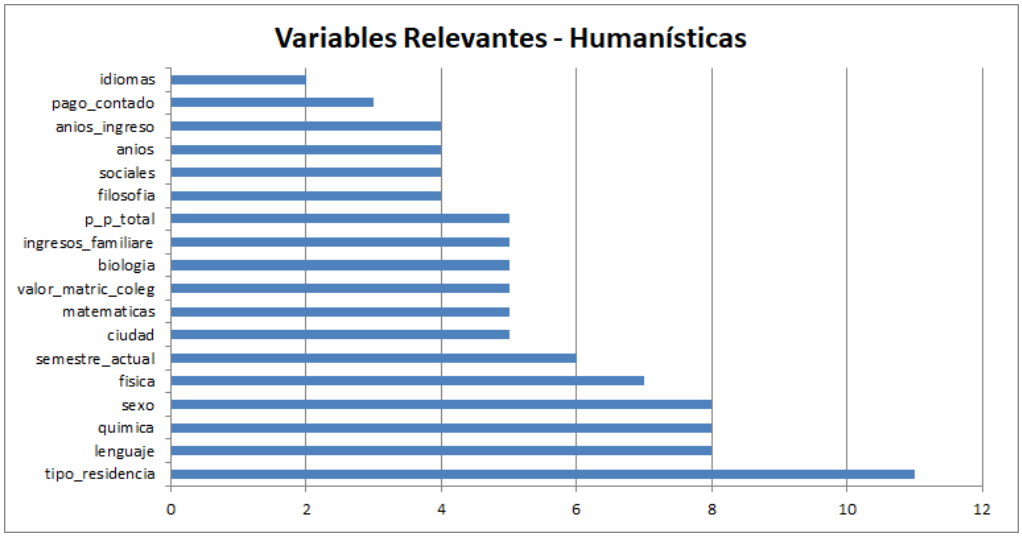
Variable	Importancia Relativa %
tipo_residencia	11
lenguaje	8
quimica	8
sexo	8
fisica	7
semestre_actual	6
ciudad	5
matematicas	5
valor_matric_coleg	5
biologia	5
ingresos_familiare	5
p_p_total	5
filosofia	4
sociales	4
anios	4
anios_ingreso	4
pago_contado	3
idiomas	2

Fuente: este trabajo

Para el caso de las carreras humanísticas, es claro que la variable más importante es **tipo_residencia** con un 11%, seguida por **lenguaje**, **química** y **sexo** con 8%, **física** con 7%, etc.

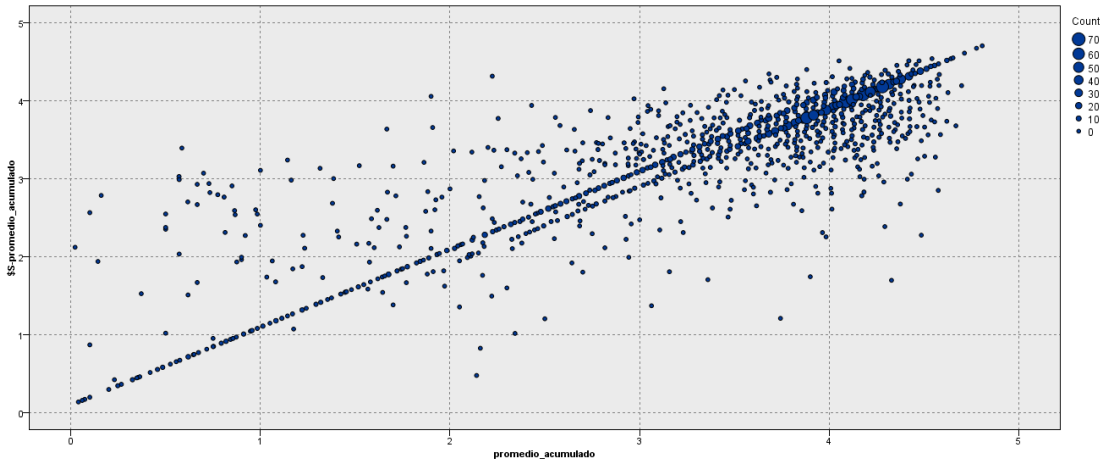
El mismo problema de sobreajuste para el modelo de las carreras técnicas también se presenta para el de las carreras humanísticas. De la figura 19 se puede observar una varianza significativa en los promedios predichos cuando son menores que 3.

Figura 18. Gráfica de las variables más influyentes en la construcción del modelo para las carreras humanísticas



Fuente: este trabajo

Figura 19. Gráfica de la variable **promedio_acumulado observado** versus **promedio_acumulado predicho** – carreras humanísticas



Fuente: este trabajo

La construcción de la figura 19 y el software utilizado es igual que para la figura 17 Gráfica de la variable **promedio_acumulado observado** versus **promedio_acumulado predicho** carreras técnicas.

- **Evaluación general de los modelos predictivos**

En general, aunque no se puede lograr que los modelos tengan un mejor rendimiento en términos de que su correlación sea más alta y el error medio absoluto sea más bajo, se pueden aceptar como viables para una primera fase de prueba posterior a este proyecto, en la que se pueda medir el rendimiento de los promedios acumulados predichos con registros desconocidos pero adheridos a la estructura que espera el modelo. Esto permitiría tomar una decisión más acertada que contemple o no implementar este modelo en un ambiente académico real.

Como soporte adicional para aceptar los modelos se presenta en las tablas 28 y 29 el cálculo de la correlación de las variable **promedio_acumulado observado** con **promedio_acumulado predicho** para ambos repositorios.

Tabla 28: Tabla de correlación entre el promedio acumulado predicho con el observado para las carreras técnicas

Correlación	**promedio_acumulado predicho**	**Relación**
promedio_acumulado observado	0.909	Fuerte

Fuente: este trabajo

Tabla 29: Tabla de correlación entre el promedio acumulado predicho con el observado para las carreras humanísticas

Correlación	**promedio_acumulado predicho**	**Relación**
promedio_acumulado observado	0.906	Fuerte

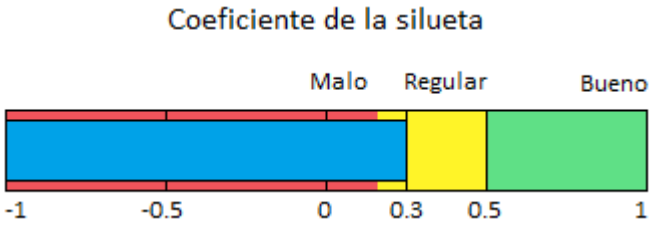
Fuente: este trabajo

Se puede notar que en ambos modelos la relación de la variable predicha con la observada es lo suficientemente fuerte como para poder confiar en el poder de regresión.

- **Evaluación del modelo de segmentación**

De las 26 variables, 25 seleccionadas en la obtención de la vista minable, y la variable 26 creada para el propósito de la segmentación **categoria_promedio_acumulado**, el algoritmo K-Means seleccionó a 18 de ellas como las más influyentes y proveyó una calidad de 0.3 medido con el coeficiente de silueta. Como se mencionó en la sección de modelado, se decidió crear 3 clústeres luego de probar con diferentes configuraciones.

Figura 20. Gráfica de la calidad del modelo de segmentación K-Means



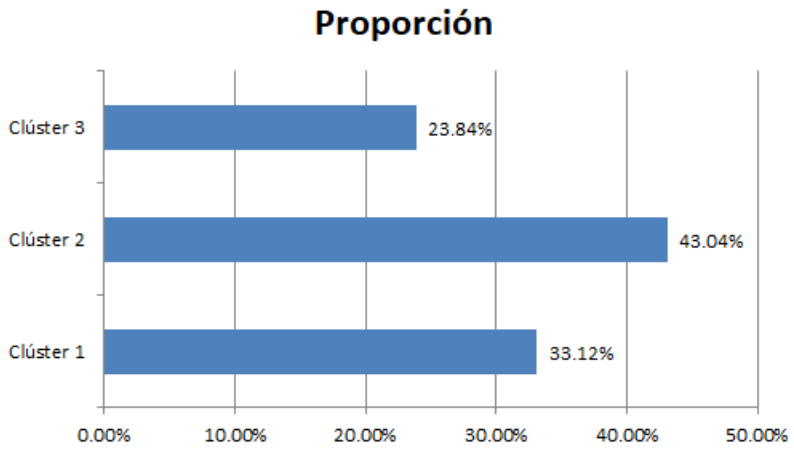
Fuente: este trabajo

El coeficiente de silueta con valor 0.3, obtenido para esta tarea de clústering; señala que la calidad del modelo es regular, esto se ve reflejado en cuán diferentes son los clústeres entre ellos. Para que haya una buena calidad y que esta se vea reflejada en un valor de silueta mayor o igual que 0.5 hasta 1.0, debe haber una diferencia significativa de los clústeres en cuanto a las características (valores de las variables seleccionadas para construirlos), ver figura 20.

La poca diferenciación de los grupos creados se verá con mayor detalle en el capítulo 6, sección **6.2 SEGMENTACIÓN**.

A continuación se presenta en la figura 21 la proporción y recuento de los clústeres.

Figura 21. Gráfica de la proporción y recuento de los clústeres



Recuento

Clúster 3	2078
Clúster 2	3751
Clúster 1	2887

Fuente: este trabajo

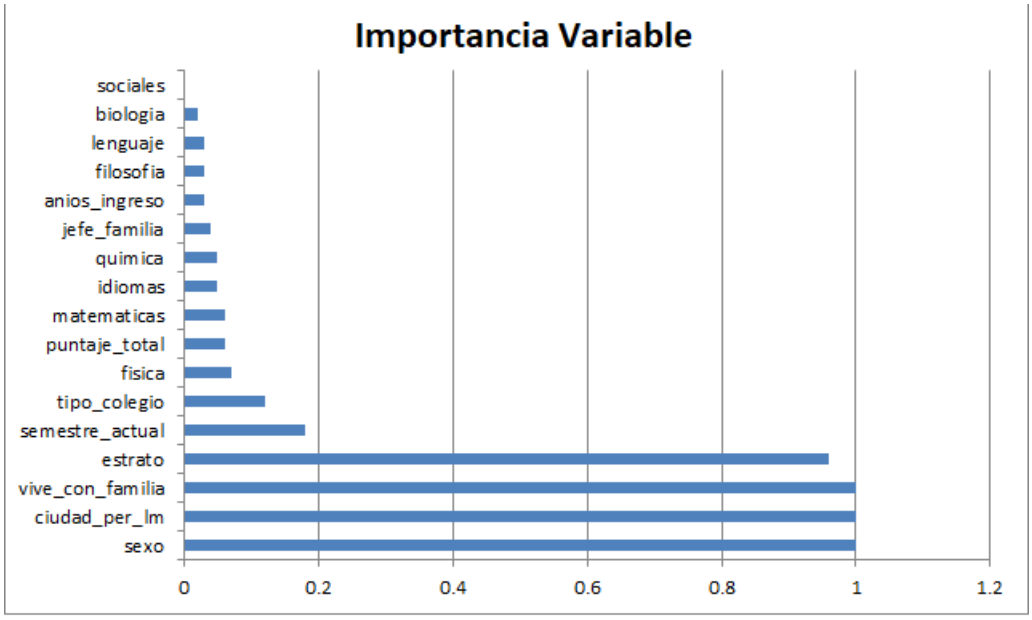
Variables influyentes

Para el algoritmo K-Means, las variables más apropiadas al momento de generar los clústeres son:

En orden de importancia: sexo, ciudad_per_lm, vive_con_familia, estrato, semestre_actual, tipo_de_colegio, física, p_p_total, matemáticas, idiomas, química, jefe_familia, anios_ingreso, filosofía, lenguaje, biología, sociales, ver figura 22.

Que una variable tenga un valor de importancia mayor que otra no es una prueba en si misma de que es mejor, sino que es más adecuada para el propósito del clustering.

Figura 22. Gráfica de las variables más influyentes para la segmentación con K-Means



Fuente: este trabajo

6. RESULTADOS Y DISCUSIÓN

Lo que hace que una variable sea importante es la fuerza de la influencia y el número de casos de la influencia.

6.1 MODELOS PREDICTIVOS

La importancia de las variables predictoras en la construcción del modelo predictivo que tiene como variable dependiente o respuesta el **promedio_acumulado** son:

- **Para el repositorio de las carreras técnicas**

Variables Categóricas

Variable **semestre_actual**

La variable **semestre_actual** tiene una influencia (importancia relativa) del 21% y en general se ve que con el pasar de los semestres, el **promedio_acumulado** mejora, por ejemplo; en los primeros (1, 2) semestres el promedio observado es 1.82 y 2.51 respectivamente y el predicho es 1.87 y 2.65; para los últimos semestres (9, 10) el observado es 3.58 y 3.67 respectivamente y el predicho es 3.58 y 3.72 respectivamente.

Variable **sexo**

La variable **sexo** tiene una influencia del 20% y el mejor **promedio_acumulado** es para las mujeres, explícitamente el promedio observado para el sexo femenino es 3.42 y para el masculino es 3.13, el predicho para el femenino es 3.44 y para el masculino es 3.17.

Variable **mas_una_cargo**

La variable **mas_una_cargo** tiene una influencia del 7% y el mejor rendimiento es para la categoría 1 con un valor observado del 4.11 y un valor predicho del 3.71, aquí se puede evidenciar que cuando hay pocas observaciones para una categoría, la máquina de vectores tiende a generar errores.

Variable **estrato**

La variable **estrato** influye con un 3% en el modelo, el mejor rendimiento académico es para 'Población o ciudad de otro Dpto' con un promedio acumulado observado 3.67 y uno predicho 3.75; el más bajo rendimiento es para el estrato 5 con un promedio observado 2.74 y uno predicho 2.94.

Variable **tipo_residencia**

La variable **tipo_residencia** contribuye con una influencia del 1%, el mejor rendimiento académico observado 4.02 y predicho 3.75 es para la categoría 'No es propia'. La máquina de vectores predijo el **promedio_acumulado** muy por debajo del observado porque esta categoría tiene pocas frecuencias en el dataset.

Variable **ciudad**

La variable **ciudad** tiene una influencia del 1% en el modelo, se puede ver que el mejor **promedio_acumulado observado** 4.06 y predicho 3.63 es para 'Samaniego', el error en la predicción se debe a las pocas observaciones de esta categoría.

Variable **ciudad_per_lm**

La variable **ciudad_per_lm** influye con el 1% y, el mejor rendimiento lo tiene 'Otras' con uno observado 3.29 y predicho 3.29 y el más bajo es para Tumaco, con un **promedio_acumulado observado** 2.78 y uno predicho 2.9.

Variables Numéricas

Las variables numéricas que influyen en el modelo y en su orden son: con el 5% **física**; con el 4% **química**, **matemáticas**, **ingresos_familiares** y **lenguaje**; con el 3% **biología**, **sociales**, **anios_ingreso**, **p_p_total** y **anios** finalmente; con el 2% **filosofía** e **idiomas**.

- **Para el repositorio de las carreras humanísticas**

Variables Categóricas

Variable **tipo_residencia**

La variable **tipo_residencia** contribuye con una influencia del 11% en el modelo, y el mejor **promedio_acumulado** es para 'Arrendada/Anticresada' con un observado 3.6 y uno predicho 3.46. Dado que esta variable se repite tanto en el modelo de las carreras técnicas y humanísticas, se puede decir que realmente influye en el rendimiento académico.

Variable **sexo**

La variable **sexo** influye con el 8%, como en las carreras técnicas, las mujeres tienen el mejor rendimiento académico. Observado 3.75, predicho 3.66

Variable **semestre_actual**

La variable **semestre_actual** tiene una influencia del 6% en el modelo, también como en el caso de las carreras técnicas, el **promedio_acumulado** mejora con el pasar de los semestres. Primeros semestres (1 y 2) el promedio observado es 1.75 y 2.65 respectivamente, el predicho 2.15 y 2.82 respectivamente, para los

últimos semestres (9, 10) el observado es 3.82 y 3.81 respectivamente y el predicho 3.74 y 3.61 respectivamente.

Variable **ciudad**

La variable **ciudad** influye un 5%, el mejor promedio es para 'Pasto' con un observado 3.55 y uno predicho 3.49.

Variables Numéricas

Las variables numéricas que influyen en el modelo y en su orden son: con el 8% **lenguaje** y **química**; con el 7% **física**; con el 5% **matemáticas**, **valor_matric_coleg**, **biología**, **ingresos_familiare** y **p_p_total**; con el 4% **filosofía**, **sociales**, **anios** y **anios_ingreso**; con el 3% **pago_contado** e **idiomas** con el 2%.

En términos de las pruebas ICFES, las variables que correspondes a los puntajes de cada materia evaluada en dicha prueba, están presentes en ambos modelos: carreras técnicas y humanísticas por ende, se concluye que estas variables influyen ciertamente en el rendimiento académico.

6.2 SEGMENTACIÓN

- **Interpretación de los clústeres**

A continuación, se describe las características de los 3 clústeres generados para esta fase de segmentación, el **ANEXO C – GRÁFICA DE CLÚSTERES RENDIMIENTO ACADÉMICO - MODELER** contiene una gráfica detallada que permite observar cada una de las variables y sus porcentajes que influyen en cada uno de los clústeres.

Por ejemplo, en el **ANEXO C**, para el clúster 1 Rendimiento alto, estos son los porcentajes para las primeras variables. ciudad_per_lm: Pasto (84%), sexo: F (100%), vive_con_familia: S (94.6%), estrato: 2(51.6%), entonces de acuerdo estos valores se construyó la caracterización del clúster 1 Rendimiento alto que está abajo. La misma idea se aplicó para los 2 clústeres restantes.

Clúster 1. Rendimiento alto

Los estudiantes que pertenecen a este clúster se caracterizan por ser todas mujeres, gran parte de ellas provienen de la ciudad de Pasto, un alto porcentaje

de ellas vive con su familia; un poco más de la mitad pertenecen al estrato 2; un 81.3% de estas estudiantes proviene de un colegio oficial, e ingresan a la universidad alrededor de los 18 años, un 51% de ellas tiene como jefe de familia al padre y cursan octavo semestre. El puntaje promedio con el que ingresan a la universidad es de 58.12. En cuanto a las áreas que evalúa las pruebas ICFES, las estudiantes que pertenecen a este clúster se caracterizan por tener en física, filosofía e idiomas, puntajes promedios entre 52 y 53, y en áreas como química, biología y sociales promedios entre 54 y 56 y en las áreas de matemáticas y lenguaje promedios entre 57 y 59.

Clúster 2. Rendimiento muy bajo

Los estudiantes que pertenecen a este clúster se caracterizan por ser hombres, gran parte de ellos provienen de la ciudad de Pasto y viven en su gran mayoría con su familia; un poco más de la mitad pertenecen al estrato 2 y la mitad de ellos tienen como jefe de familia al padre; un gran porcentaje de estos estudiantes proviene de un colegio Oficial, ingresan a la universidad alrededor de los 19 años y cursan sexto semestre. El puntaje promedio con el que ingresan a la universidad es de 59.54. En cuanto a las áreas que evalúa las pruebas ICFES, estos estudiantes se caracterizan por tener en áreas como sociales, biología, lenguaje, química, y física puntaje promedios entre 54 y 57, y en áreas como idiomas y filosofía promedios entre 51 y 52 y un caso especial para el área de matemáticas con un promedio alrededor de 60.

Clúster 3. Rendimiento medio

Los estudiantes que pertenecen a este clúster se caracterizan por ser en un 65.5% hombres, cerca del 69% de estos estudiantes provienen de otras ciudades diferentes a Pasto, gran parte de ellos no vive con su familia; el 70.5% de ellos pertenecen al estrato 1 y un poco menos de la mitad de ellos tienen como jefe de familia a la madre; un gran porcentaje de estos estudiantes proviene de un colegio Oficial, ingresan a la universidad alrededor de los 19 años y cursan sexto semestre. El puntaje promedio con el que ingresan a la universidad es de 59.63. En cuanto a las áreas que evalúa las pruebas ICFES, los estudiantes que pertenecen a este clúster se caracterizan por tener en idiomas y filosofía, puntajes promedios entre 49 y 52; en áreas como física, lenguaje, biología, sociales y química, promedios entre 55 y 58; y en el área de matemáticas promedios alrededor de 62.

6.3 CONCLUSIONES, RECOMENDACIONES Y SUGERENCIAS PARA TRABAJOS FUTUROS

- **Conclusiones**

Específicamente para esta base de datos y trabajo de investigación, se puede decir que:

1. De los modelos de minería de datos analizados para predicción, el mejor de todos fue las máquinas de vectores de soporte, esta misma situación se presenta en los antecedentes revisados que utilizaron esta técnica y que de verdad la recomiendan para problemas de rendimiento académico.

2. Como están provistos los datos, el modelo de segmentación no permite identificar claramente un clúster para el rendimiento académico bajo.

3. Teniendo en cuenta los antecedentes y los resultados arrojados por esta investigación se puede decir que la minería de datos es una herramienta muy utilizada para abordar problemas tan complejos como el rendimiento académico porque permite ver la situación desde una perspectiva técnica y educacional entonces al final de un proceso de minería los datos provistos al algoritmo, la construcción del modelo y el modelo como tal se resumen en reportes, cuadros comparativos, gráficos, etc., que permiten tener un mapa mental de toda la problemática. Sin detrimento de lo dicho anteriormente, construir un modelo adecuado para los datos que sirva para propósitos de predicción o segmentación resulta en una tarea difícil ya que los datos tienen problemas desde el origen y aunque se haga una labor ardua en la limpieza y organización de los mismos a veces resulta insuficiente.

4. Como es común en trabajos de minería de datos, se tuvo problemas con el sobreajuste en la partición de entrenamiento y con la generación de un modelo de predicción tolerable utilizando la partición de prueba que tenga una correlación alta y un error absoluto medio menor, situación que desembocó en el retorno a las fases de limpieza y organización de datos en repetidas ocasiones, esto generó la eliminación de registros que no cumplieron con las técnicas aplicadas, a pesar de esto, el sobreajuste se redujo muy poco.

5. En general, para generar un modelo de predicción con resultados óptimos, es decir; que se pueda hacer predicciones con pocos errores resulta en una tarea complicada no sólo por el problema natural que hay en las estructuras de los datos sino también que el rendimiento académico se puede abordar de diferentes perspectivas y llevar esto a un modelo no siempre se ajusta

como se desea. El clustering tampoco está exento de esta situación, por ejemplo; este trabajo expone un cluster con un coeficiente de silueta de 0.3 que evidencia que encontrar patrones en el rendimiento académico es difícil.

6. Luego de haber analizado los modelos de predicción, se puede ver que el rendimiento es más bajo en los 2 primeros semestres. Esta situación se evidencia en el análisis de las variables para el problema de predicción en ambos repositorios de las carreras técnicas y las humanísticas.

7. Se puede ver una situación interesante en las variables más relevantes para los modelos de predicción de las carreras técnicas y humanísticas, ambos comparten todas las variables de los puntajes del examen ICFES como son: **biología**, **matemáticas**, **física**, **química**, **lenguaje**, **idiomas**, **sociales**, **filosofía** esto muestra que en realidad los puntajes de las pruebas de estado influyen en el rendimiento académico.

8. Resulta curioso que la variable **semestre_actual** tenga una correlación alta en el modelo de predicción de las carreras técnicas y no sea así en el modelo de predicción para las humanísticas ya que luego de haber visto los resultados del primer modelo se esperaba lo mismo para el segundo.

9. De las variables más importantes para la segmentación, se puede ver que la primera es el **sexo**, luego **ciudad_permanencia_lm** y le siguen las variables socioeconómicas **vive_con_familia** y **estrato** esto puede dar a entender que de verdad el sexo, la situación socioeconómica y demográfica del estudiante afecta su rendimiento académico.

10. De acuerdo al modelo de segmentación, la variable **p_p_total** (promedio ponderado con el que ingresa el estudiante al programa que eligió) es mayor para los clústeres con rendimiento muy bajo y medio, el menor valor de esta variable es para el clúster del rendimiento alto, esto podría dar a entender que tener un alto promedio ponderado (**p_p_total**) no es un indicador de buen rendimiento académico, recordando la definición para el **promedio_acumulado** en la sección **5.1.11.1 Comprensión de los datos**, se define este variable como: “El promedio general acumulado y el semestral o anual de calificaciones de un estudiante, será el que resulte de calcular el promedio aritmético de todas las notas registradas, tomado en unidades, décimas y centésimas” (Nariño, 1998); y las categorías que se utilizan para interpretarlo son: Muy Bajo: cuando el promedio es inferior a 3.0, Bajo: cuando el promedio está entre 3.0 y 3.49, Medio: cuando el promedio está entre 3.5 y 3.99, Alto: iguales o superiores a 4.

11. El rendimiento medio es compartido por hombres y mujeres.

12. El rendimiento muy bajo es una característica mayoritariamente masculina.

13. Las mujeres tienen mejor rendimiento académico que los hombres, se puede evidenciar en los modelos de predicción y segmentación.

- **Recomendaciones**

1. Una recomendación luego de haber ejecutado este proyecto es que la universidad mejore la recolección de datos de sus estudiantes en etapas como: inscripción, durante los estudios y al egreso. Esto no sólo serviría para hacer trabajos de investigación con minería de datos sino con otras técnicas o metodologías que se desarrollen para múltiples propósitos. Por ejemplo, se puede diseñar un formulario de inscripción o matrícula enfocado específicamente para tareas de minería de datos.

2. Partiendo de la conclusión 6, se puede sugerir que se centren esfuerzos para mejorar el rendimiento académico en los 2 primeros semestres de todas las carreras de pregrado que ofrece la universidad, se podría crear y ejecutar algún tipo de política preventiva.

3. Mientras se ejecutaba el proceso de limpieza de datos, se pudo ver que hay variables importantes que permiten la omisión del dato, entonces se aconseja que los formularios de ingreso de información sean un poco más estrictos con el diligenciamiento de variables realmente importantes.

4. Aunque esta investigación arroja algunos resultados interesantes respecto del rendimiento académico, estos sólo se aplican al caso de la Universidad de Nariño, porque este problema está sujeto a las características que cada institución educativa tenga entonces se sugiere encausar el análisis de estos sólo a esta institución.

5. Como pequeño aporte de este estudio de investigación, se puede decir que se abordó el problema del rendimiento académico de forma general para todas las carreras de pregrado y de todos los semestres, mientras que en trabajos anteriores se hace sólo para algunos semestres, generalmente los primeros, o para algunas materias en particular, aquello no está mal pues a los investigadores les interesaba un resultado más preciso mientras que a este trabajo le parece adecuado un resultado más global.

- **Sugerencias para trabajos futuros**

1. Implementar este modelo usando instancias desconocidas cuya estructura se ajuste a la esperada para mejorar la predicción y el rendimiento en general. Esto podría mejorar aún más si se decide expandir este trabajo investigativo con los periodos académicos más actuales.

2. Se exhorta a la universidad la implementación de un una solución de minería de datos que aborde el problema del rendimiento académico como una herramienta de uso frecuente para consultar la situación académica particular de cada estudiante, de forma que se pueda generar estrategias preventivas contra rendimientos bajos.

3. De los resultados de este trabajo investigativo, se puede notar que para ambos modelos de predicción de las carreras técnicas y humanísticas las variables más relevantes son casi las mismas en ambos casos (**semestre_actual**, **sexo**, **física**, **química**, etc), aunque sus porcentajes de correlación sean diferentes, entonces sería interesante que un trabajo futuro aborde el grupo de datos como uno solo, es decir sin dividirlo, esto podría mejorar la precisión, el error medio absoluto y el desempeño en general.

4. De los antecedentes consultados, ninguno promueve como trabajo futuro la medición del rendimiento académico para estudiantes de postgrado, sería interesante ver cómo un estudiante de postgrado que se desempeñe profesionalmente rinde académicamente cuando tiene una carga laboral encima.

7. BIBLIOGRAFÍA

Alcover, R et al. «Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos».

Álvarez, Maria Teresa;, y Hernan García. 1996. *Factores que Predicen el Rendimiento Universitario*. San Juan de Pasto (Colombia): Universidad de Nariño.

Bramer, Max A. 2007. *Principles of Data Mining*. http://books.google.co.kr/books?id=xVW7NslHNHsC.

Carvajal Olaya, Patricia, Julio César Mosquera, y Irina Artamonova. 2009. «MODELOS DE PREDICCIÓN DEL RENDIMIENTO ACADEMICO EN MATEMÁTICAS I EN LA UNIVERSIDAD TECNOLÓGICA DE PEREIRA». http://revistas.utp.edu.co/index.php/revistaciencia/article/download/2323/1237.

Chapman, Pete et al. 2000. «Crisp-Dm 1.0». *CRISP-DM Consortium*: 76.

Deng, Naiyang, Yingjie Tian, y Chunhua Zhang. 2013. *Support Vector Machines Optimization Based Theory, Algorithms, and Extensions*.

Erazo, Oscar. 2012. «El rendimiento académico, un fenómeno de múltiples relaciones y complejidades». *Revista Vanguardia Psicológica Clínica Teórica y práctica* 2(2): 144-73.

Gallardo Arancibia, José Alberto. 2009. «Metodología para la definición de requisitos en proyectos de data mining (ER-DM)». : 317. http://oa.upm.es/1946/1/JOSE_ALBERTO_GALLARDO_ARANCIBIA.pdf.

Garbanzo, Guiselle María. 2007. «Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública.» *Educación* 31(1): 43-63. http://www.redalyc.org/pdf/440/44031103.pdf.

Harwati, Ardita Permata Alfiani, y Febriana Ayu Wulandari. 2015. «Mapping Student's Performance Based on Data Mining Approach (A Case Study)». *Agriculture and Agricultural Science Procedia* 3: 173-77. http://linkinghub.elsevier.com/retrieve/pii/S2210784315000352.

Hastie, Trevor, Robert Tibshirani, y Jerome Friedman. 2009. «The Elements of Statistical Learning». *Elements* 1: 337-87. http://www.springerlink.com/index/10.1007/b94608.

IBM, Corp. 2016. *Nodos de Modelado de IBM SPSS Modeler 18.0*.

Kantardzic, Mehmed. 2011. 12 *Data Mining Concepts, Models, Methods, and Algorithms*. IEEE Press.

Ktona, Ana, Denada Xhaja, y Ilia Ninka. 2014. «Extracting Relationships between Students' Academic Performance and Their Area of Interest Using Data Mining Techniques». *2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks*: 6-11. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7059136.

Lanzarini, Laura, Emilia Charnelli, y D Javier. 2015. «Academic Performance of University Students and its Relation with Employment».

Merchán, S M, y J A Duarte. 2016. «Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance». *IEEE Latin America Transactions* 14(6): 2783-88.

Nariño, Universidad de. 1998. *ESTATUTO ESTUDIANTIL DE PREGRADO DE LA UNIVERSIDAD DE NARIÑO*. http://www2.udenar.edu.co/recursos/wp-content/uploads/2017/08/document-est.pdf.

Navarro, Rubén. 2003. «El rendimiento académico: concepto, investigación y desarrollo». *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación* 1(2): 1-16. http://www.redalyc.org/pdf/551/55110208.pdf%0Ahttp://www.ice.deusto.es/rinace/reice/vol1n2/Edel.pdf EL.

Orallo, Jose Hernandez, Maria José Ramirez Quintana, y Cèsar Ferri Ramírez. 2004. Introduccion a la mineria de datos *Introducción a La Minería De Datos*. http://users.dsic.upv.es/~flip/LibroMD/.

Pachano, Liz Jeannette Aranguren. 2008. «Identificacion De Patrone De Consumo De Los Venezolanos Mediante Maquinas De Vectores De Soporte». Universidad de los Andes, Mérida.

Pereira, Ricardo Timarán. 2009. «Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos». *iiisorg*: 5. http://www.iiis.org/CDs2009/CD2009CSC/CISCI2009/PapersPdf/C692YV.pdf.

Pérez López, César, y Daniel Santín Gonzalez. 2007. *Minería de datos. Técnicas y herramientas*. Paraninfo.

Perez Marqués, Maria. 2015. *Mineria de datos a través de ejemplos*.

Pérez Marques, María. 2013. *Técnicas de Mineria de Datos. Modelos Predictivos*. Createspace.

Poh, Norman, y Ian Smythe. 2015. «To what extend can we predict students' performance? A case study in colleges in South Africa». *IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CIDM 2014: 2014 IEEE Symposium on Computational Intelligence and Data Mining, Proceedings*: 416-21.

Porcel, Eduardo Adolfo; Dapozo, Gladys Noemí;, y Maria Victoria Lopéz. 2010. «Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa». https://redie.uabc.mx/redie/article/view/264/730.

Rousseeuw, Peter J. 1987. «Silhouettes: A graphical aid to the interpretation and validation of cluster analysis». *Journal of Computational and Applied Mathematics* 20(C): 53-65.

Saltelli, a. 2002. «Making best use of model valuations to compute sensitivity indices». *Computer Physics Communications* 145: 280-97.

Saltelli, Andrea, Stefano Tarantola, Francesca Campolongo, y Marco Ratto. 2004. *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons.

Slim, Ahmad, Gregory L. Heileman, Jarred Kozlick, y Chaouki T. Abdallah. 2015. «Predicting student success based on prior performance». *IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CIDM 2014: 2014 IEEE Symposium on Computational Intelligence and Data Mining, Proceedings*: 410-15. http://ieeexplore.ieee.org.ezproxy.utp.edu.co/document/7008697/.

Sorour, Shaymaa E., Tsunenori Mine, Kazumasa Godaz, y Sachio Hirokawax. 2014. «Comments data mining for evaluating student's performance». *Proceedings - 2014 IIAI 3rd International Conference on Advanced Applied Informatics, IIAI-AAI 2014*: 25-30. http://ieeexplore.ieee.org.ezproxy.utp.edu.co/document/6913261/.

SPSS Modeler. 2016. «Algorithms Guide».

Tobergte, David R., y Shirley Curtis. 2013. 53 Journal of Chemical Information and Modeling *Machine learning with R*.

Tuffery, S. 2011. Wiley *Data Mining and Statistics for Decision-Making*.

Vialardi, César et al. 2011. «A data mining approach to guide students through the enrollment process based on academic performance». *User Modeling and User-Adapted Interaction* 21(1-2): 217-48. http://link.springer.com.ezproxy.utp.edu.co/article/10.1007%2Fs11257-011-9098-4.

Wendler, Tilo, y Sören Gröttrup. 2016. *Data Mining with SPSS Modeler*. http://link.springer.com/10.1007/978-3-319-28709-6.

ANEXOS

ANEXO A – TABLAS AUXILIARES DICCIONARIO DE DATOS

Las siguientes son tablas auxiliares que ayudan a entender los posibles valores de las variables mencionadas en la sección **5.1.10.1 Compresión de los datos**.

Valores posibles para la variable Sexo

Código	Descripción
M	Masculino
F	Femenino

Fuente: este trabajo

Valores posibles para la variable Ciudad Nacimiento

Descripción
Ipiales
Pasto
Tumaco
Túquerres
Otras

Fuente: este trabajo

Valores posibles para la variable Ciudad Proveniencia

Descripción
Ipiales
Pasto
Tumaco
Túquerres
Otras

Fuente: este trabajo

Valores posibles para la variable Ciudad Sede Universidad

Descripción
Ipiales
La Unión
Pasto
Samaniego
Túquerres
Tumaco

Fuente: este trabajo

Valores posibles para la variable Estrato

Descripción
1
2
3
4
5
Cabecera de Municipio de Nariño
Cabecera de Municipio de Putumayo
Población a Ciudad de otro Departamento
Población o ciudad de otro País
Vereda de Nariño o Putumayo

Fuente: este trabajo

Valores posibles para la variable Jefe Familia

Descripción
Inscrito
Madre
Padre
Otro

Fuente: este trabajo

Valores posibles para la variable Más de Una Persona a Cargo

Código	Descripción
1	Una persona a cargo
N	Ninguna persona a cargo
S	Más de una persona a cargo

Fuente: este trabajo

Valores posibles para la variable Tipo de Residencia

Descripción
Arrendada/Anticresada
No es Propia
Propia
Propia/paga cuotas

Fuente: este trabajo

Valores posibles para la variable Vive Con Familia

Código	Descripción
N	No
S	Si

Fuete: este trabajo

Valores posibles para la variable Tiene Hermanos Universitarios

Código	Descripción
N	No
S	Si

Fuente: este trabajo

Valores posibles para la variable Tipo Colegio

Descripción
Desaparecido
No informa
Oficial
Privado
Valido ICFES

Fuente: este trabajo

Valores posibles para la variable Ingreso a la Universidad con Cupo Especial

Descripción
CUPO LICEO UDENAR
CUPO REGULAR
DEPORTISTA DESTACADO
DESPLAZADOS NARIÑO-PUTUMAYO
EXTRANJEROS
HIJO DE VICTIMA DE SECUESTRO
HIJO DE VICTIMA DE DESAPARICION
MEJOR BACHILLLER DEPARTAMENTO DE NARIÑO
MUNICIPIOS DEPRIMIDOS SOCIO-ECONOMICAMENTE
NEGRITUDES DE LA ZONA PACIFICA NARIÑENSE
PERTENECIENTE A CABILDOS INDIGENAS de NARIÑO
PERTENECIENTE CABILDO INDIGENA PUTUMAYO
PROFESIONALES
REINSERTADOS

Fuente: este trabajo

Valores posibles para la variable Estado Actual

Descripción
Egresado sin Título
Graduado
Regular

Fuente: este trabajo

Valores posibles para la variable Vigente Actualmente

Código	Descripción
N	No
S	Si

Fuente: este trabajo

Valores posibles para la variable Semestre Actual

Código	Descripción
1	Semestre 1
2	Semestre 2
3	Semestre 3
4	Semestre 4
5	Semestre 5
6	Semestre 6
7	Semestre 7
8	Semestre 8
9	Semestre 9
10	Semestre 10
15	Egresado
110	Graduado

Fuente: este trabajo

Valores posibles para la variable Nombre de la Carrera

Código	Descripción
LIC. INFOR	Licenciatura en informática
LIC. FIL. LET	Licenciatura en filosofía y letras
FISICA	Física
LIC. EDU. BAS. HUM. CAS. ING	Licenciatura en educación básica con énfasis en humanidad, lengua castellana e ingles
ING. AGRONOM	Ingeniería agronómica
ADM. EMPRESAS	Administración de empresas
ECONOMIA	Economía
DERECHO	Derecho
ART. VISUALES	Artes visuales
LIC. MUS	Licenciatura en música
MEDICINA	Medicina
ING. AGROIND	Ingeniería agroindustrial
SOCIOLOGIA	Sociología
DIS. INDUSTRIAL	Diseño industrial
LIC. LEN. CAS. LIT	Licenciatura en lengua castellana y literatura
ING. PROD. ACUI	Ingeniería en producción acuícola
ING. SISTEMAS	Ingeniería de sistemas
ING. CIVIL	Ingeniería civil
ING. ELECT	Ingeniería electrónica
TEC. COMP	Tecnología en computación
COM. INTERN	Comercio internacional
ING. AGROF	Ingeniería agroforestal
DIS. GRAF. MULT	Diseño gráfico y multimedia
ARQUITECTURA	Arquitectura
LIC. MAT	Licenciatura en matemáticas
LIC. ING. FRAN	Licenciatura en inglés-francés
LIC. ART. VIS	Licenciatura en artes visuales
ZOOTECNIA	Zootecnia

TEC. PROM. SAL	Tecnología en promoción de la salud
ING. AMB	Ingeniería ambiental
BIOLOGIA	Biología
PSICOLOGIA	Psicología
MED. VET	Medicina veterinaria
LIC. EDU. BAS. NAT. AMB	Licenciatura educación básica énfasis ciencias naturales y educación ambiental
LIC. EDU. BAS. CIEN. SOC	Licenciatura en educación básica con énfasis en ciencias sociales
GEOGRAFIA	Geografía
QUIMICA	Química
CONT. PUBLICA	Contaduría publica

Fuente: este trabajo

Valores posibles para la variable Nombre de la facultad

Código	Descripción
FACIEN	Ciencias Exactas y Naturales
FACIHU	Ciencias Humanas
FACIA	Ciencias Agrícolas
FACEA	Ciencias Económicas y Administrativas
DERECHO	Derecho
FACARTES	Artes
FACSALUD	Ciencias de la Salud
FACIAGRO	Ingeniería Agroindustrial
INGENIERÍA	Ingeniería
FACEDU	Educación
FACIPEC	Ciencias pecuarias

Fuente: este trabajo

Valores posibles para la variable Estado Vigente

Código
N
S
T

Fuente: este trabajo

ANEXO B – TABLA DE PONDERACIONES CON TARJETAS ICFES

Programa	LENG. %	MATE %	CIEN. SOC. %	FILOS %	BIO %	QUIM %	FÍSICA %	IDIOMAS %
1. Administración de Empresas	30	20	20	15	0	0	0	15
2. Arquitectura	20	15	20	20	5		20	
3. Artes Visuales	20	10	20	20	10	5	5	10
4. Biología	15	20	10	5	30	10	10	
5. Comercio Internacional	20	20	30	10	0	0	0	20
6. Contaduría Pública	30	30	20	20				
7. Derecho	30	5	30	20	5	5	5	
8. Diseño Gráfico	25	15	30	15	5	5	5	
9. Diseño Industrial	25	15	10	15	5	15	15	
10. Economía	20	20	25	20	5	5	5	0
11. Física	15	20	10	5	10	15	25	
12. Geografía	20	20	35		10		10	5
13. Ingeniería Agroforestal	5	20	10		25	25	10	5
14. Ingeniería Agroindustrial	15	20			20	20	15	10
15. Ingeniería Agronómica	5	20	10		25	25	10	5
16. Ingeniería Ambiental	5	20	10		25	25	10	5
17. Ingeniería Civil	15	25	10	5	5	15	25	
18. Ingeniería de Sistemas	15	25	10	5	5	15	25	
19. Ingeniería Electrónica	15	25	10	5	5	15	25	
20. Ingeniería en Producción Acuícola	15	20	10	5	20	20	10	
21. Lic. Educación Básica con 1Énfasis en Ciencias Naturales y Educación Ambiental	20	5	15	10	20	15	15	
22. Licenciatura en Artes Visuales	20	10	20	20	10	5	5	10
23. Licenciatura en Educación Básica con Énfasis en Ciencias Sociales	20	10	30	20	5	5	5	5
24. Licenciatura en Educación Básica: Lengua Castellana e Inglés	35	5	10	10	5	0	0	35
25. Licenciatura en Filosofía y Letras	25	10	25	25	5	5	5	
26. Licenciatura en Informática	25	30	10	15	5	5	10	
27. Licenciatura en Inglés – Francés	35	5	10	10	5	0	0	35
28. Licenciatura en Lengua Castellana y Literatura	20	5	30	30	5	5	5	
29. Licenciatura en Matemáticas	30	30	5	15	5	5	10	
30. Licenciatura en Música	20	5	30	20	10	5	10	
31. Medicina	15	10	5	5	25	25	15	
32. Mercadeo	15	20	30	15	0	0	0	20
33. Medicina Veterinaria	20	10	10	5	25	20	5	5
34. Promoción de la Salud	20	15	15	5	20	15	10	
35. Psicología	30	20	10	10	10	5	5	10
36. Química	15	15	10	5	10	30	15	
37. Sociología	20	20	35	20	5			
38. Tecnología en Computación	15	15	20	10	10	15	15	
39. Zootecnia	10	15	10	5	20	20	10	10

Fuente:http://apolo.udenar.edu.co/admisiones/documentos/PUNTAJES_Y_PONDERADOS_A_DE_2015.pdf

ANEXO C – GRÁFICA DE CLÚSTERES RENDIMIENTO ACADÉMICO - MODELER

Clústeres

Importancia de entrada (predictor)

1,0 0,8 0,6 0,4 0,2 0,0

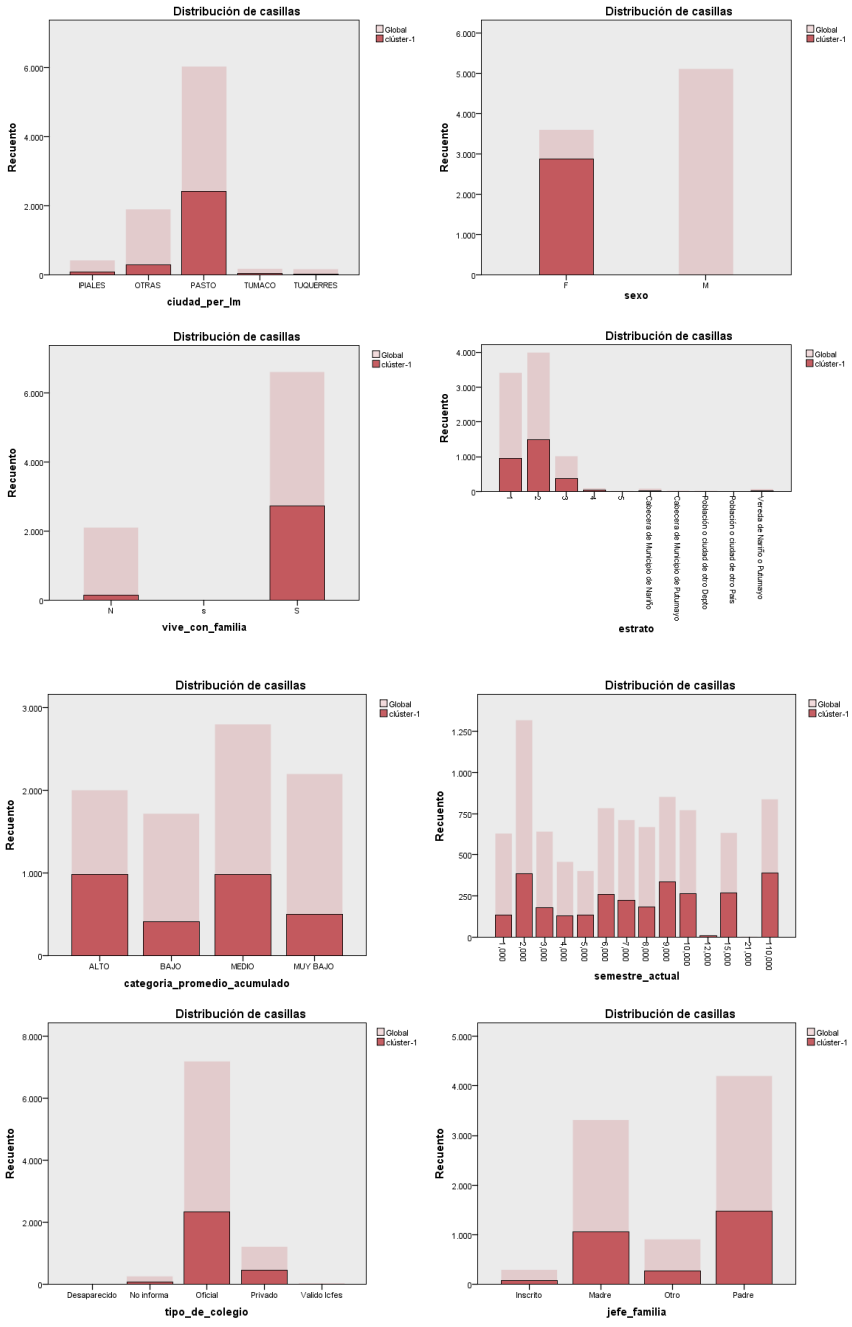
Clúster	clúster-2	clúster-1	clúster-3
Etiqueta			
Descripción			
Tamaño	43,0% (3751)	33,1% (2887)	23,8% (2078)
Entradas	ciudad_per_lm PASTO (88,9%)	ciudad_per_lm PASTO (84,0%)	ciudad_per_lm OTRAS (68,8%)
	sexo M (100,0%)	sexo F (100,0%)	sexo M (65,5%)
	vive_con_familia S (94,2%)	vive_con_familia S (94,6%)	vive_con_familia N (83,4%)
	estrato 2 (55,2%)	estrato 2 (51,6%)	estrato 1 (70,5%)
	categoria_promedio_acumulado	categoria_promedio_acumulado	categoria_promedio_acumulado

semestre_actual 6,000	semestre_actual 8,000	semestre_actual 6,000
tipo_de_colegio Oficial (78,3%)	tipo_de_colegio Oficial (81,3%)	tipo_de_colegio Oficial (91,8%)
fisica 54,91	fisica 53,06	fisica 55,89
p_p_total 59,54	p_p_total 58,12	p_p_total 59,63
matematicas 60,59	matematicas 58,48	matematicas 61,45
idiomas 51,86	idiomas 52,29	idiomas 49,73
quimica 55,55	quimica 55,02	quimica 57,28

	jefe_familia Padre (50,1%)	jefe_familia Padre (51,0%)	jefe_familia Madre (43,4%)
	anios_ingreso 19,47	anios_ingreso 18,85	anios_ingreso 19,32
	filosofia 51,32	filosofia 52,85	filosofia 51,87
	lenguaje 56,85	lenguaje 57,84	lenguaje 56,66
	biologia 55,37	biologia 54,94	biologia 56,36
	sociales 55,69	sociales 55,35	sociales 55,33

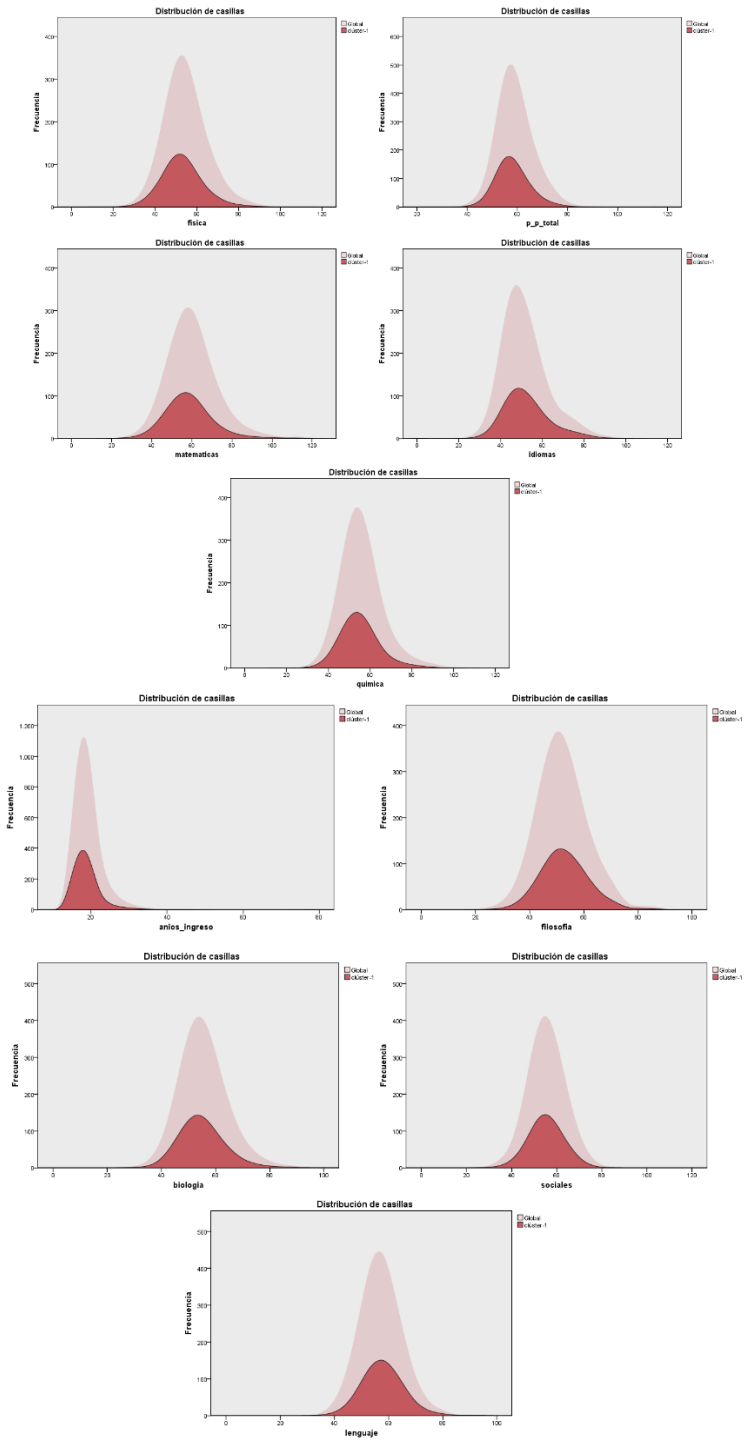
ANEXO D – GRÁFICAS VARIABLES CLÚSTER 1

Gráfica variables categóricas – Clúster 1



Fuente: este trabajo

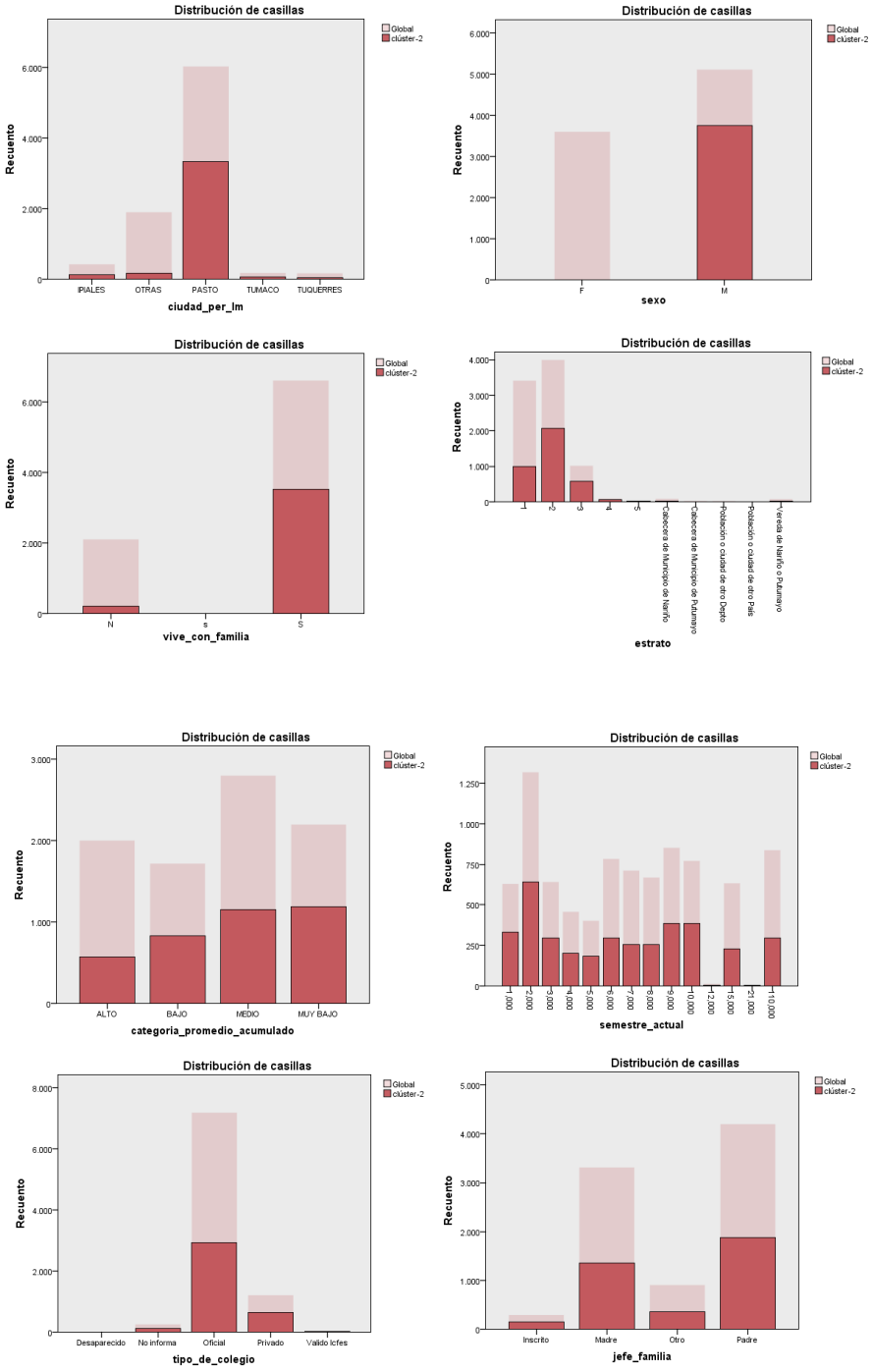
Gráfica variables numéricas – Clúster 1



Fuente: este trabajo

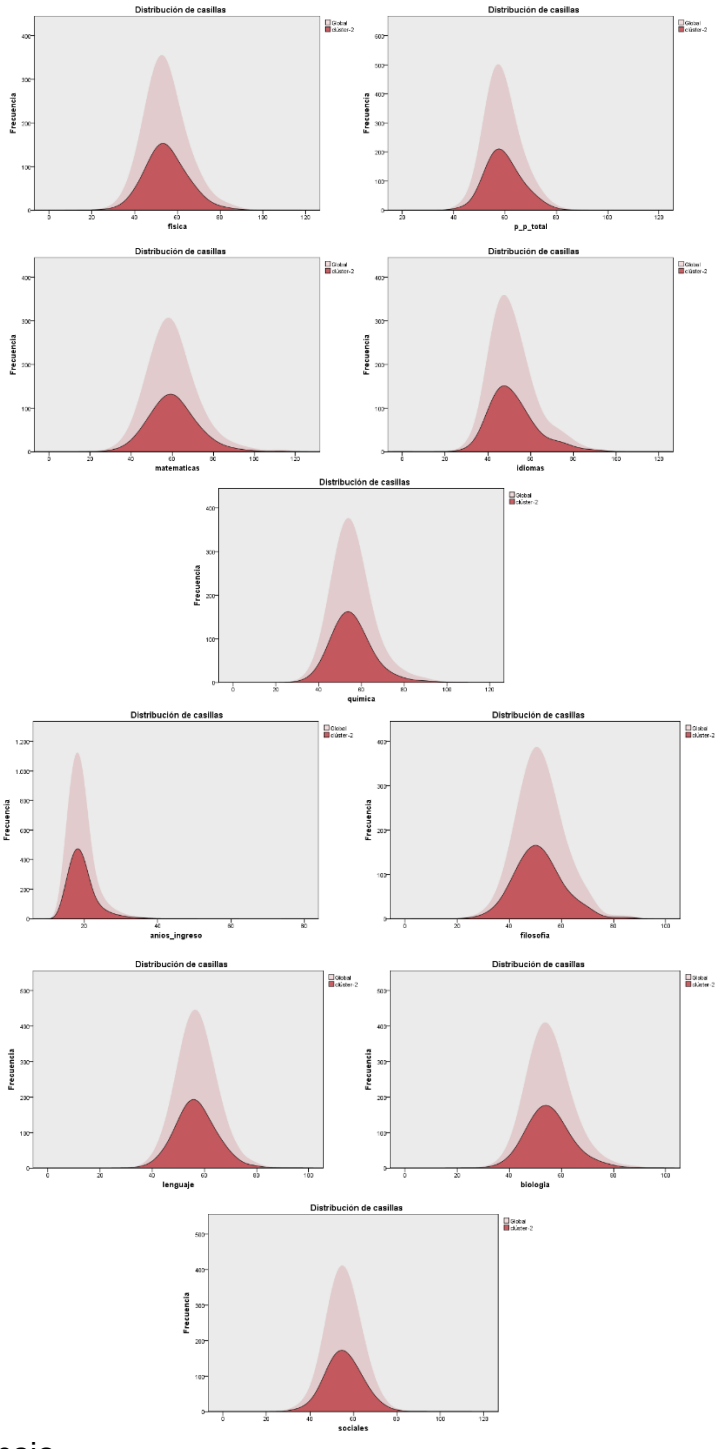
ANEXO E – GRÁFICAS VARIABLES CLÚSTER 2

Gráfica variables categóricas – Clúster 2



Fuente: este trabajo

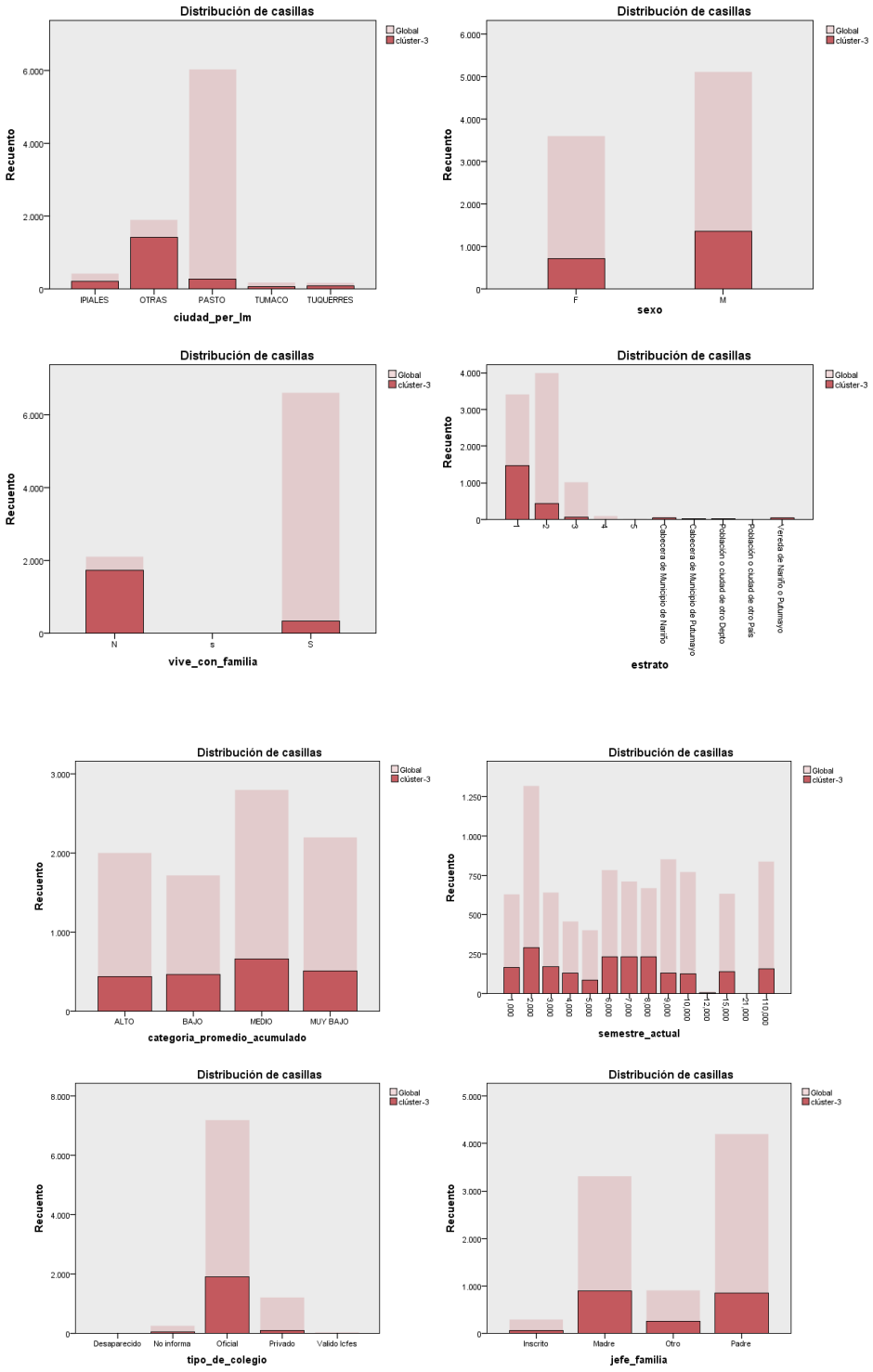
Gráfica variables numéricas – Clúster 2



Fuente: este trabajo

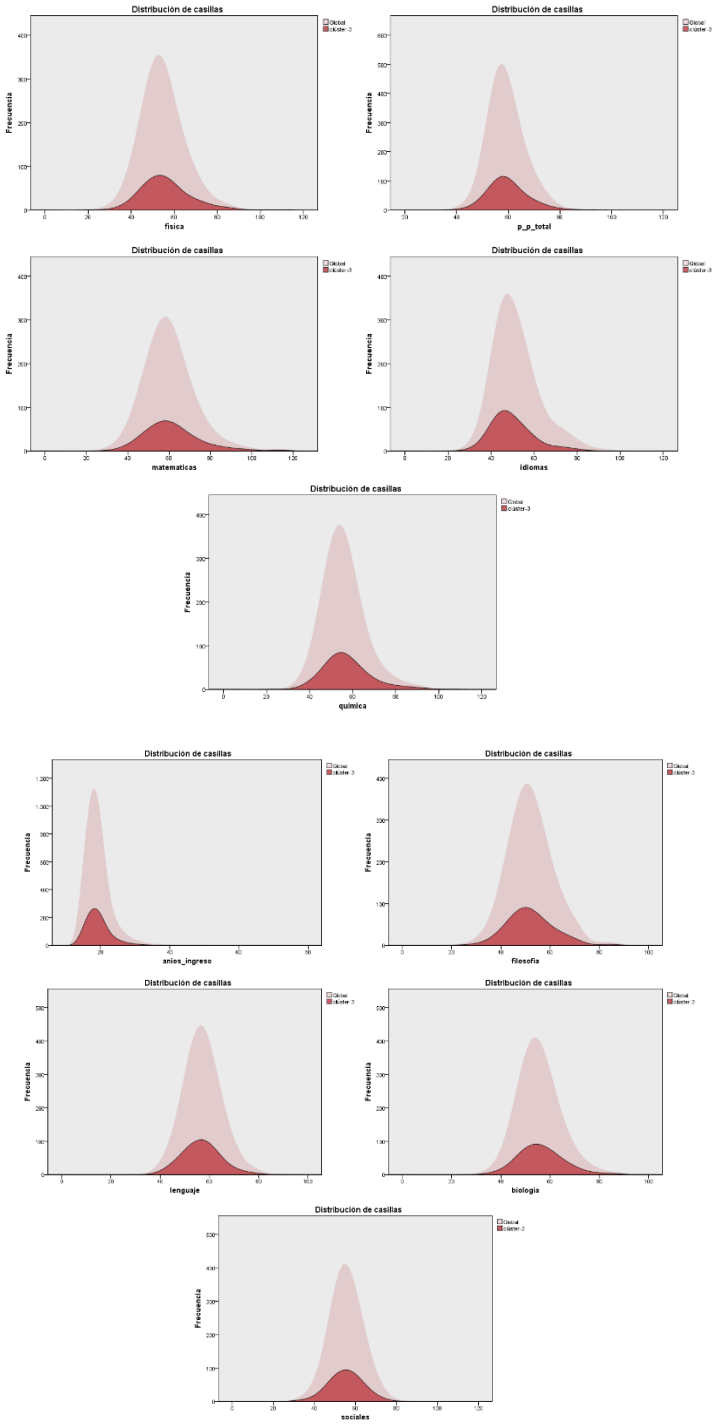
ANEXO F – GRÁFICAS VARIABLES CLÚSTER 3

Gráfica variables categóricas – Clúster 3



Fuente: este trabajo

Gráfica variables numéricas – Clúster 3



Fuente: este trabajo

3166

San Juan de Pasto, 21 de Noviembre de 2017

Vo Bo
Ing. Juan Carlos, favor atender la solicitud preservando el anonimato de los estudiantes..

Doctor
Javier Caicedo Zambrano
Vicerrector Académico (e)
Universidad de Nariño

Cordial saludo,

Somos, Sandra Viviana Escobar Madroñero y Jaime Harvey Enríquez Tulcán; estudiantes de último semestre de la Maestría en Investigación de Operaciones y Estadística de la universidad Tecnológica de Pereira (UTP) en convenio con la Universidad de Nariño; y en ejecución del trabajo de grado en minería de datos enfocado al rendimiento académico.

Por el anterior motivo, solicitamos que se nos otorgue un permiso escrito sobre la base de datos de los estudiantes de pregrado de la Universidad de Nariño entre los años 2010 y 2014, a cuyos registros se les elimina columnas como la cédula de ciudadanía, código interno estudiantil, dirección, teléfono y cualquier otra que implique algún tipo de identificación, también se remueve filas que no correspondan al periodo académico mencionado; con el propósito de poder llevar a cabo nuestro trabajo de grado.

Muchas gracias por su tiempo y colaboración.

Rdo: 23-11-17

Atentamente

UNIVERSIDAD DE NARIÑO
VICERRECTORIA ACADEMICA
21 NOV. 2017
Hora: 09:42am
Responsable:

Sandra Viviana Escobar M
Sandra Viviana Escobar Madroñero
C.C.: 37 083 551 de Pasto
3105376717

Jaime Enríquez
Jaime Harvey Enríquez Tulcán
C.C.: 1085 252903 de Pasto

Universidad de Nariño
Centro de Informática
27 NOV. 2017
CI No. 768 Nº Folios 3
Recibido 1341

Correo: sandraescobar2008@hotmail.com
belerofonte@hotmail.co.uk

ACTA DE COMPROMISO DE CONFIDENCIALIDAD PARA EL SUMINISTRO DE INFORMACIÓN SOBRE ESTUDIANTES DE LA UNIVERSIDAD DE NARIÑO

Yo Sandra Viviana Escobar M, identificado con el cédula número 37'083.551, en mi calidad de estudiante, del Programa de Maestría en Investigación Operativa y Estadística de la Universidad de Nariño en convenio con la UTP (Pereira), me comprometo a no usufructuar, ni a utilizar la información a la que tuve acceso directo o indirecto con un fin distinto al del proyecto, investigación o informe titulado Identificación de patrones de rendimiento académico de los estudiantes de pre-grado de la Universidad de Nariño entre los años 2010 y 2014 utilizando Minería de datos

Mantendré el respeto por la confidencialidad y reserva de la misma, de conformidad con las normas sobre el suministro de información personal y las contenidas en la Ley 1266 de 2008 y demás normas concordantes.

Dentro del Acta de Compromiso se establece que tanto el solicitante y la Universidad de Nariño aceptan cumplir lo siguiente:

1. El acceso a información y/o datos estadísticos restringidos estará limitado al solicitante registrado en la solicitud.
2. Copias de la información y/o datos estadísticos no serán reproducidos o puestos a disposición de otra entidad o persona diferente a las que menciona en esta acta.
3. La información suministrada deberá utilizarse únicamente para reportar la información agregada, y no para investigar o reportar a personas u organizaciones específicas. Los datos no podrán utilizarse en ninguna forma para efectos administrativos, judiciales, de propiedad exclusiva, o para la ejecución de alguna ley.
4. Cualquiera de los libros, artículos, documentos de conferencias, tesis, disertaciones, informes u otras publicaciones que utilizan datos obtenidos deben citar la fuente de los mismos.
5. Si hay algún cambio en la especificación del proyecto, investigación o informe para el cual se ha solicitado esta información, es responsabilidad del Investigador principal que la Oficina de Control y Registro Académico de la Universidad de Nariño acepte estos cambios.

Se firma en San Juan de Pasto a los 27 días del mes de Noviembre del año 2017

Nombre: Sandra Viviana Escobar
C .C. No. 37'083.551.
Firma: Sandra Viviana Escobar M.

Universidad de Nariño
Centro de Informática
27 NOV. 2017
CI No. 328 Nº Folios 3
Recibido 1789

ACTA DE COMPROMISO DE CONFIDENCIALIDAD PARA EL SUMINISTRO DE INFORMACIÓN SOBRE ESTUDIANTES DE LA UNIVERSIDAD DE NARIÑO

Yo Jaime Harvey Enríquez Tulcán, identificado con el cédula número 1085 252 903 en mi calidad de estudiante del Programa de maestría en investigación de operaciones y estadística de la Universidad de Nariño en convenio con la UTP (pereira), me comprometo a no usufructuar, ni a utilizar la información a la que tuve acceso directo o indirecto con un fin distinto al del proyecto, investigación o informe titulado Identificación de patrones de rendimiento académico de los estudiantes de la universidad de Nariño entre los años 2010 y 2014 usando minería de datos

Mantendré el respeto por la confidencialidad y reserva de la misma, de conformidad con las normas sobre el suministro de información personal y las contenidas en la Ley 1266 de 2008 y demás normas concordantes.

Dentro del Acta de Compromiso se establece que tanto el solicitante y la Universidad de Nariño aceptan cumplir lo siguiente:

1. El acceso a información y/o datos estadísticos restringidos estará limitado al solicitante registrado en la solicitud.
2. Copias de la información y/o datos estadísticos no serán reproducidos o puestos a disposición de otra entidad o persona diferente a las que menciona en esta acta.
3. La información suministrada deberá utilizarse únicamente para reportar la información agregada, y no para investigar o reportar a personas u organizaciones específicas. Los datos no podrán utilizarse en ninguna forma para efectos administrativos, judiciales, de propiedad exclusiva, o para la ejecución de alguna ley.
4. Cualquiera de los libros, artículos, documentos de conferencias, tesis, disertaciones, informes u otras publicaciones que utilizan datos obtenidos deben citar la fuente de los mismos.
5. Si hay algún cambio en la especificación del proyecto, investigación o informe para el cual se ha solicitado esta información, es responsabilidad del Investigador principal que la Oficina de Control y Registro Académico de la Universidad de Nariño acepte estos cambios.

Se firma en San Juan de Pasto a los 27 días del mes de Noviembre del año 2017

Nombre: Jaime Enríquez
C .C. No. 1085 252 903
Firma: Jaime Enríquez

Universidad de Nariño
Centro de Informática
27 NOV. 2017
CI No. 318 N° Folios 3
Recibido 1739