Identificación de variables relevantes en la deserción de estudiantes del departamento de Ingeniería Electrónica y Telecomunicaciones mediante algoritmos no supervisados

Ana María Giraldo Marín

Proyecto de investigación para optar por el título de Ingeniera Electrónica

Asesor Interno
Claudia Victoria Isaza Narváez, PhD

Profesora Facultad de Ingeniería

Universidad de Antioquia
Facultad de Ingeniería

Departamento de Electrónica y Telecomunicaciones

Medellín, Colombia

2022

Cita		Giraldo Marín [1]
Referencia Estilo IEEE (2020)	[1]	A. Giraldo Marín, “Identificación de variables relevantes en la deserción de estudiantes del departamento de Ingeniería Electrónica y Telecomunicaciones mediante algoritmos no supervisados”, Proyecto de Investigación, Ingeniería Electrónica, Universidad de Antioquia, Medellín, 2022.

Grupo de Investigación SISTEMIC.

Resumen

En este proyecto de investigación, se propone una metodología que permite identificar las variables más relevantes en el fenómeno de deserción de los estudiantes del Departamento de Ingeniería Electrónica y Telecomunicaciones de la Universidad de Antioquia, realizando un análisis de estas variables para tres grupos: Estudiantes de todo el departamento, estudiantes en modalidad virtual y estudiantes de modalidad presencial, asignando a cada uno de los grupos encontrados en los tres análisis un nivel de riesgo de deserción y un perfil de los estudiantes pertenecientes a los grupos con distribución de desertores más relevantes (mayor y menor porcentaje), estos perfiles tienen asociadas variables de interés que son las responsables de que el estudiante quede clasificado en un grupo en particular. Se entrega también, una herramienta para la predicción de riesgo de deserción para cada uno de los modelos analizados.

Palabras Claves: Deserción, Clustering, KMeans, Aprendizaje no supervisado, Variables Relevantes.

Tabla de contenido

Lista de Imágenes

Lista de Tablas

Introducción

Las instituciones de educación superior se enfrentan ante la gran problemática de la deserción estudiantil [1]. Este fenómeno es de gran relevancia porque no solo afecta a las instituciones educativas, también con esto, se ve truncado el proyecto de vida de las personas y se ve afectado el desarrollo social y económico de la comunidad [2][3]. En Colombia, la tasa de deserción acumulada en el año 2014 a nivel nacional, según el Ministerio de Educación Nacional fue de un 50% [9], lo que quiere decir que la mitad de los estudiantes que empiezan un proceso de educación superior abandonan sus estudios. La Universidad de Antioquia no es ajena a esta problemática, en especial, este fenómeno afecta a la Facultad de Ingeniería, donde, al cabo de 10 semestres se tiene un 51% de deserción en programas con modalidad presencial y un 76% en programas con modalidad virtual [4].

En estudios realizados sobre deserción se concluye que este fenómeno es multicausal, siendo los factores individuales, académicos, sociales e institucionales los principales determinantes [5]. Cada uno de estos cuatro factores, se definen haciendo uso de diversas variables, lo cual hace que el proceso de análisis del perfil del desertor sea difícil de realizar manualmente por la cantidad de variables a considerar, es aquí, donde empieza a ser de gran ayuda la implementación de técnicas de minería de datos para la obtención de información a partir de las relaciones ocultas entre variables [8].

Para entender más a fondo este fenómeno, desde el Departamento de Ingeniería Electrónica y de Telecomunicaciones se adelanta un proyecto orientado a identificar el perfil de los estudiantes con tendencia a la deserción y lograr así encaminar de mejor manera los esfuerzos que hace la facultad para disminuir la deserción. Al tratar un fenómeno multicausal y sobre el cual no se ha llegado a respuestas claras y absolutas, se concluye que es importante trabajar con algoritmos donde no se limite el análisis con respuestas esperadas (técnicas supervisadas), por esto, se abordara el proyecto con técnicas no supervisadas. Las técnicas no supervisadas agrupan los datos de acuerdo a su similitud, y a partir del análisis del patrón de cada grupo encontrado se puede identificar las variables que llevaron a tener este tipo de división. En este trabajo se usaron técnicas no supervisadas para diferenciar entre el perfil de estudiantes desertores y no desertores.

Objetivos

Objetivo General

Identificar las variables de relevancia en el fenómeno de deserción estudiantil de los estudiantes del Departamento de Ingeniería Electrónica y Telecomunicaciones de la Universidad de Antioquia, mediante técnicas de aprendizaje no supervisado, con el propósito de determinar el perfil de los estudiantes con tendencia a la deserción.

Objetivos Específicos

- Establecer posibles variables candidatas a partir del análisis de literatura de trabajos previos realizados en el tema de deserción.
- Implementar algoritmo de agrupamiento en lenguaje Python para agrupar a los estudiantes con perfiles similares analizando los centroides (vector con valores en cada variable) de cada grupo.
- Asociar niveles de riesgo de deserción según los grupos encontrados separando el problema entre deserción temprana y tardía.
- Identificar las variables que llevan a hacer el agrupamiento a partir del análisis de los cambios de los centros de los grupos encontrados.

Marco Teórico

Fenómeno de Deserción en Educación Superior

La deserción en educación superior es un fenómeno multicausal extremadamente complejo, tanto así que no es posible realizar una definición única capaz de captar en totalidad toda su esencia, por esto, la definición de deserción se realiza de acuerdo a los intereses y metas de la investigación, buscando la que mejor se ajusta [6]. Para este estudio se adoptará la definición dada por el Ministerio de Educación Nacional Colombiano en base a la definición de Tinto (1982) y Giovagnoli (2002):Se llama deserción a la situación a la que se enfrenta un estudiante cuando no logra concluir su proyecto educativo, considerándose como desertor aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos lo cual equivale a un año de inactividad [7]

Dada la complejidad del fenómeno en análisis, diferentes autores coinciden en que la deserción no solo depende de factores académicos, sino también de factores individuales, socioeconómicos e institucionales [7]. En el estado del arte, se evidencian como variables relevantes en la decisión de desertar de un programa académico: edad, género, capacidad económica, lugar de residencia, desempeño académico en el programa, si el estudiante trabaja, habilidades lógicas, habilidades en lectoescritura, entre otras. Estas variables se incluyeron en el perfil de cada estudiante para ser analizadas como posibles indicadores de riesgo de deserción.

Aprendizaje no supervisado

Este estudio hace uso de técnicas que permiten la construcción de un modelo de ingeniería descriptiva, donde el objetivo es identificar grupos naturales en los datos y verificar si estos grupos se asocian al nivel de riesgo de deserción del estudiante. Se trabaja con una técnica no supervisada porque no se tiene certeza de los grupos que van a permitir asociar los estudiantes a los diferentes niveles de riesgo de deserción.

Clustering (Agrupamiento)

El agrupamiento es una técnica de aprendizaje no supervisado, la cual presenta éxito en la detección y clasificación de conjuntos de datos. Para esto se extrae la información que caracteriza los rangos de cada individuo para ser asignado a un grupo. Cada grupo está compuesto por un conjunto de individuos que son similares

entre sí, y al mismo tiempo diferentes a los individuos de otros grupos [10]. El objetivo es encontrar grupos con características similares. La mayoría de algoritmos de agrupamiento encuentran los grupos a partir de maximizar las distancias entre grupo y minimizar las distancias entre los datos del mismo. Cada grupo encontrado tiene un centroide, que corresponde al perfil más representativo de los datos del mismo grupo. [8]. En el caso de los algoritmos basados en distancia, el centroide corresponde al punto central (centro) de cada grupo de datos. A continuación, se muestra gráficamente el proceso realizado por un algoritmo de *clustering*. Los algoritmos encuentran los grupos (en la figura se representa cada grupo por un color) sin tener en cuenta una etiqueta (en la figura se representa por tener puntos sin color).

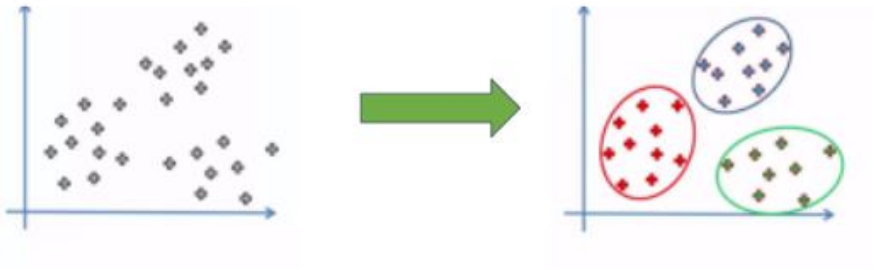


Imagen 1. Proceso de clustering

Proceso de agrupamiento, a la izquierda se muestra la población que se analizará, representada por las cruces grises y a la derecha está el resultado de implementar el algoritmo de agrupación, allí se puede evidenciar que la población quedó agrupada en tres grupos, cada uno representado con un color diferente.

Uno de los algoritmos más usados para realizar agrupamiento es el *KMeans*, este se basa en el análisis de distancia entre datos. El *KMeans* realiza el siguiente procedimiento [14]:

1. Se inicializan los centroides (centro geométrico del cluster) uno para cada *cluster*.
2. Se procede a calcular para cada dato la distancia (euclidiana cuadrada) con respecto a todos los centroides, se determina el centroide más cercano a cada uno de estos, y cada dato se anexa al clusters del centroide que fue seleccionado.
3. Se actualiza el valor de los centroides asignándoles la posición del promedio de los objetos pertenecientes al *cluster*.

4. Se realiza la verificación de convergencia, algunas de las condiciones utilizadas son:
   - El número de iteraciones.
   - Los centroides obtenidos en dos iteraciones sucesivas no cambian su valor.
   - La diferencia entre los centroides de dos iteraciones sucesivas no supera cierto umbral.
   - No hay transferencia de objetos entre grupos en dos iteraciones sucesivas.

   Si no se logra la convergencia, se repetirán los pasos 2,3 y 4 hasta que se cumplan las condiciones.

Una vez converge el algoritmo, los centroides pueden ser utilizados para realizar los análisis pertinentes.

Medidas de validación de calidad de agrupamiento

Al ser los algoritmos de agrupamiento una técnica de aprendizaje no supervisada, el conjunto de datos se distribuirá en el número de clases elegido por el usuario, independiente de si el número de grupos es apropiado para el universo analizado, es por esto que al implementar un algoritmo de agrupamiento se deben utilizar métricas para poder determinar la calidad y robustez de los grupos hallados, y elegir así el agrupamiento más adecuado. Para esto, existen medidas internas y externas, las medidas internas miden la homogeneidad y separación de los conjuntos y las externas miden los resultados de acuerdo al conocimiento de las clases o agrupaciones correctas (etiquetas externas) para los datos bajo análisis, en este proyecto no es posible realizar validación externa ya que no se cuenta con información sobre las clases reales.[12]

Las medidas de validación internas que se utilizaron en el proyecto se presentan en la siguiente tabla:

Tabla 1. Índices de Validación Interna de Calidad de Agrupación de Clúster [11]

Índice de Validación Interna	Ecuación	Interpretación
Silhouette	$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$ $SC = \frac{1}{K}\sum_{i=1}^{K} s(x)$ K= Número de grupos. $a(x)$= Distancia promedio de x a los puntos en el mismo grupo. $b(x)$= Distancia promedio de x a los puntos del grupo más cercano.	Entrega un valor entre 1 y -1, siendo 1 el indicador de una buena agrupación.
Davies-Bouldin	$DB = \frac{1}{k}\sum_{i=1, i\neq j}^{k} \max\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$ K= Número de grupos. σ_i=Distancia promedio entre los puntos del grupo i y su centroide. σ_j=Distancia promedio entre los puntos del grupo j y su centroide. $d(c_i, c_j)$=Distancia entre los centroides de los grupos i y j.	Se considera mejor la calidad de agrupación cuando el valor obtenido es menor.
Dunn	$D = \frac{\min_{1\leq i<j\leq n} d(i,j)}{\max_{1\leq x\leq n} d'(x)}$ $d(i,j)$=Distancia entre los grupos i y j. $d'(X)$= distancia de el punto analizado a los otros puntos del grupo.	Entre más alto es el valor obtenido, mejor será la calidad de agrupación.

Metodología para identificar las variables más relevantes en la deserción:

Se presenta a continuación, la metodología que se planteó para el desarrollo del proyecto:

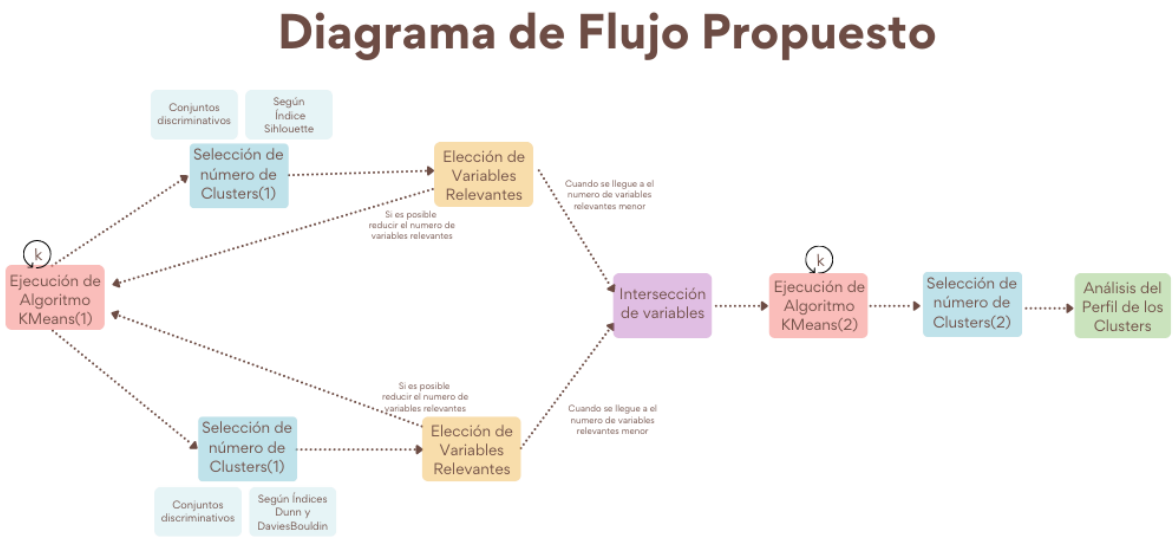


Imagen 2. Metodología propuesta

Pasos propuestos para la identificación de las variables relevantes en el fenómeno de deserción, para luego finalizar encontrando el perfil de los clusters mediante el análisis de los centroides resultantes de la ejecución del algoritmo KMeans.

Antes de entrar a detallar lo realizado, es importante aclarar que actualmente no existe un método estándar para realizar la selección del número apropiado de clases, igualmente, tampoco se tiene un proceso estándar para la elección de variables relevantes, por lo que el método utilizado fue diseñado durante la ejecución del proyecto.

Ejecución de Algoritmo KMeans (1)

Partiendo de que para el proyecto es necesario usar algoritmos de agrupamiento, se elige trabajar con el algoritmo KMeans, con las siguientes especificaciones:

Tolerancia = 1e-6

Lo siguiente que se debe especificar es el número de conjuntos en los que se agrupará el universo de datos, como se desconoce el número óptimo de grupos se da un rango de 2 a

25, los cuales serán analizados posteriormente para elegir así el número de conjuntos en los que se agrupe adecuadamente a los estudiantes.

Para el desarrollo de esta etapa se utilizó la librería de *Python Scikit-Learn (sklearn)*.

Selección de Numero de Clusters (1)

Para la selección del número de conjuntos en los que se debe separar el universo de estudiantes, se utilizaran los siguientes conceptos:

Medidas de calidad de agrupación: Como se indica en la sección del marco teórico, los índices Silhouette, Dunn y Davies-Bouldin dan información sobre la calidad de agrupación al dividir el universo de estudiantes en K clases, estos índices son calculados para cada número de grupos en los que se divide el universo de estudiantes. Para poder hacer una comparación entre estos índices, se normalizan todos los valores encontrados, al final se obtienen unas graficas como las mostradas en la imagen 3:

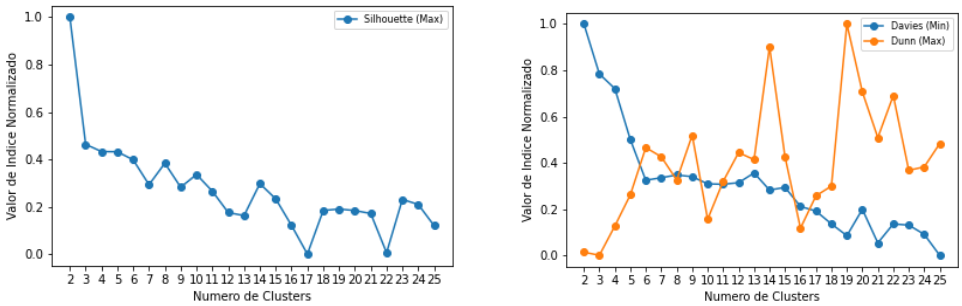


Imagen 3. Ejemplo de Graficas de Índices de Calidad de Agrupación

Se presentan ejemplos de graficas de índices de calidad de agrupación normalizados. En la parte izquierda se tiene una representación del índice Silhouette (agrupación optima cuando es máximo). En la parte derecha de la imagen, se grafican los índices Davies-Bouldin en color azul (agrupación optima cuando es mínimo) y Dunn en color naranja (agrupación optima cuando es máximo).

Distribución de desertores entre clases: Esta medida se halla encontrando el porcentaje de desertores que contiene cada conjunto, para luego analizar si existe una diferencia considerable entre estos porcentajes de cada *cluster*. Si es así, se considera que al dividirse los estudiantes en K grupos, se obtiene una distribución de desertores entre conjuntos discriminativa, de lo contrario, la distribución será no discriminativa. A continuación, un ejemplo de una distribución de desertores entre agrupaciones no discriminativa (Imagen 4) y una discriminativa (Imagen 5).

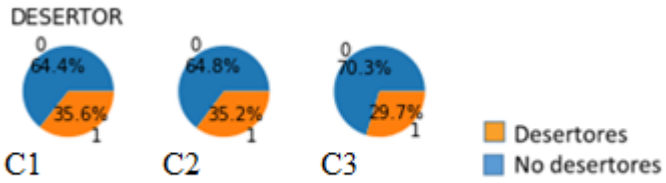


Imagen 4. Distribución no discriminativa de desertores entre clases

Al analizar el porcentaje de desertores que contiene cada uno de los grupos, se evidencia que no existe una diferencia considerable, por esto, se puede concluir que al agrupar a los estudiantes en tres grupos no se discriminará a los estudiantes desertores de los no desertores.

Por el contrario, analizando el porcentaje de estudiantes desertores en los 6 conjuntos, se puede concluir que es una distribución que discrimina entre estudiantes desertores y no desertores, pues existen clases como el C5 donde más de la mitad de los estudiantes son desertores y clases como el C6 donde los estudiantes desertores son muy pocos

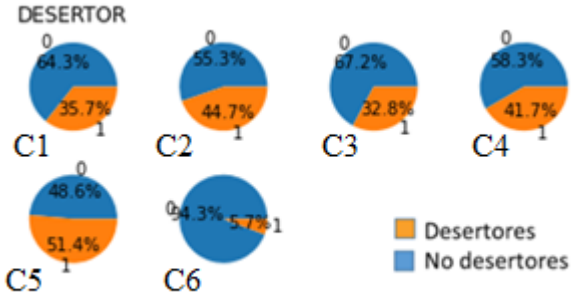


Imagen 5. Distribución discriminativa de desertores entre clases

En resumen, una vez se tienen los estudiantes clasificados en grupos, se procede a elegir el número óptimo de clases, para esto nos basamos en las medidas de calidad de agrupación y la distribución de desertores entre clases de la siguiente forma:

En este punto el análisis se divide en dos partes:

1. Se analiza la gráfica del índice Silhouette, y se toma el número de K donde se obtiene el valor máximo del índice, posteriormente se analiza la distribución de desertores entre clases para el numero K. Si no es una distribución discriminativa, se procede a analizar el siguiente valor más alto del índice Silhouette, si la distribución es

discriminativa, se dice que K es un numero óptimo de grupos para dividir el universo de estudiantes analizados y se detiene el análisis.

2. Se analiza la gráfica del Índice Dunn y Davies-Bouldin, y se toma el número de K donde se cumpla que el valor del índice Dunn es alto y el del índice Davies-Bouldin es bajo, luego se analiza la distribución de desertores para los K grupos y si es una distribución no discriminativa se busca otro K donde se cumpla con las condiciones mencionadas anteriormente, si la distribución es discriminativa se dice que K es un numero óptimo de clases para dividir a los estudiantes.

Los dos análisis (Silhouette, Dunn y Davies-Bouldin) se hacen en paralelo pues cada uno propone número de *clusters* diferente. Al final se compara las variables obtenidas con los diferentes índices.

Elección de Variables Relevantes

La distribución de los estudiantes en las clases se da de acuerdo a los valores que tiene cada estudiante para las 203 variables analizadas inicialmente, sin embargo, no todas las variables influyen de manera importante para hacer el agrupamiento de datos. Se buscó identificar las que mayor importancia tienen para agrupar un estudiante en un *cluster* y no en otro. Es aquí donde se fundamenta el concepto de variables relevantes, haciendo referencia a las variables que "hacen la diferencia" en el momento de la asignación de clase a un sujeto, gráficamente se vería así:

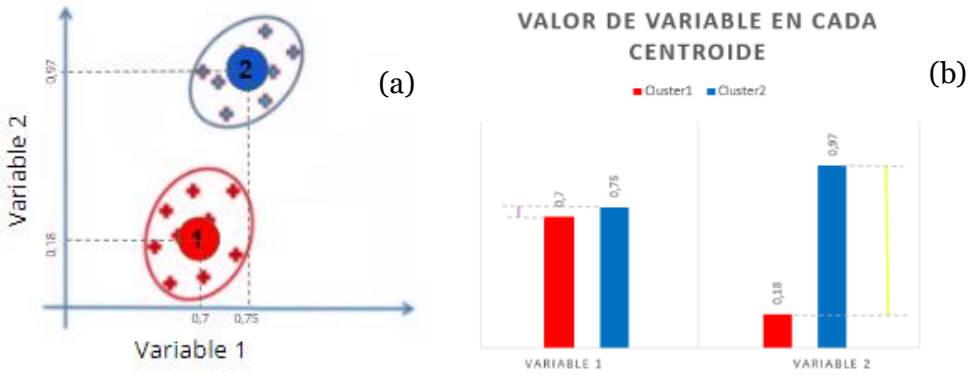


Imagen 6. Variables relevantes **(a)**Clases y la posición de sus centroides **(b)**Diferencia entre los valores de variables en cada centroide.

Para entender más fácil el concepto, se tienen a sujetos con dos variables: 'Variable 1' y 'Variable 2', separados en dos grupos, el 1 (color rojo) y el 2 (color azul), al analizar los valores de los centroides para cada una de las variables, al lado derecho de la imagen, se encuentra que la diferencia que existe entre los valores de los centroides para la 'Variables 1' es mínima, cosa contraria que sucede al analizar la diferencia de los valores de los centroides para la 'Variable 2'

De la imagen 6 (a), se puede concluir que la Variable 2 es una variable relevante ya que el valor que tenga el estudiante en esta afectara su asignación de clase, por el contrario, la Variable 1 sería una variable no relevante, pues el valor que tenga el estudiante en esta variable no afectara su asignación de clase.

Realizando un mayor análisis de la situación anterior, se puede concluir que las variables relevantes son aquellas en las que existe una mayor diferencia en el valor que tiene la variable en mención en cada uno de los centroides de los conjuntos analizados, como se evidencia en la imagen 6 (b).
Esto quiere decir que, si se desea encontrar las variables más relevantes, se deberá encontrar la diferencia de los valores que tiene la variable X para todos los centroides y repetir el proceso con todas las variables para luego seleccionar las que tienen una mayor diferencia entre centroides. Para esto, después de hallar esta diferencia para todas las variables, se prosigue a ordenar estos valores de mayor a menor, para luego hallar el punto de inflexión, el cual se toma como punto de corte, es decir, de este punto hacia la izquierda están las variables relevantes y de este punto hacia la derecha se encuentran las variables no relevantes, como se muestra en la Imagen 7.

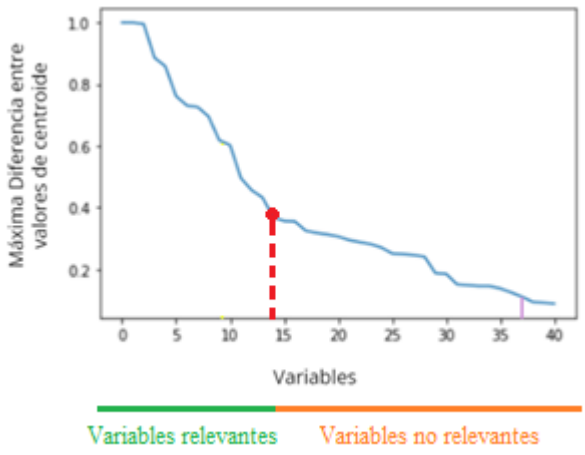


Imagen 7. Diferencia entre centroides y punto de inflexión

En la gráfica se encuentra el valor máximo de la diferencia entre los valores de los centroides para cada una de las variables analizadas, ordenado de mayor a menor. También, en color rojo, se marca el punto de inflexión, el cual separa a las variables relevantes de las variables no relevantes.

Una vez se encuentran las variables relevantes, se debe analizar si estas son el número mínimo con el que se garantiza que se puede discriminar estudiantes desertores y no

desertores. Para esto se hace la siguiente iteración: Este proceso se repetirá hasta que se llegue a una de las dos situaciones mencionadas a continuación, en este caso se tomará la base de datos utilizada en el proceso anterior:

- El criterio de punto de inflexión no aplique (esto será cuando la inflexión esté ubicada en los extremos, incluya a todas las variables o ninguna)
- Las distribuciones a las que se llegan con los diferentes números de grupos no son discriminativas: No exista una diferencia notable entre la distribución de desertores y no desertores en los grupos.

Si el número de variables se puede reducir, se deberá repetir los pasos mencionados hasta el momento; modificando la base de datos con la que se alimenta al algoritmo KMeans, si el número de variables es el mínimo, se prosigue con el siguiente paso del método.

En resumen, se selecciona el número de grupos de acuerdo a un índice interno (Silhouette, Dunn y Davies-Bouldin), se analizan las variables que son relevantes para la agrupación a la que se llega, esto mediante un análisis de los valores de los centroides de cada cluster, realizando una resta entre el valor de la variable en cada centroide con el valor de la variable en los demás centroides. El resultado de estas operaciones se ordena de mayor a menor y se haya el punto de inflexión, que será el punto de corte. Las variables que tengan una diferencia mayor a la que tiene la variable donde se encuentra el punto de corte, serán las variables relevantes.
Con este nuevo grupo de variables, se realiza el análisis de discriminación de los grupos y si se está en un extremo de la gráfica de distancias para saber si se debe repetir el análisis o se sigue con el siguiente paso.

Es importante aclarar que este proceso se debe realizar en paralelo para los dos resultados encontrados en el anterior paso, el hallado con el Índice Silhouette y los Índices Dunn y Davies-Bouldin.

Intersección de variables

Una vez se tienen las variables más relevantes de cada una de las agrupaciones analizadas, la encontrada por el Índice Silhouette y la encontrada usando a los Índices Dunn y Davies-Bouldin, se realiza la intersección de estas variables para encontrar así el conjunto final de las variables relevantes.

Ejecución de Algoritmo KMeans (2)

En este paso, se ejecuta nuevamente el algoritmo KMeans variando su valor de K desde 2 hasta 25, pero esta vez siendo alimentado solo por las variables identificadas como relevantes y que resultan después de la intersección.

Selección de Número de *Clusters* (2)

Una vez se tienen los estudiantes separados por clases, se elegirá el número de conjuntos que mejor discrimina entre desertores, para esto nos basamos en las medidas de calidad de agrupación (Silhouette, Dunn y Davies-Bouldin) y en la distribución de desertores entre clases y a diferencia de lo que se realiza en el paso 'Selección de Numero de Clusters (1)', aquí se realiza un análisis en conjunto de los índices, esto quiere decir, que se tomara como K apropiado, para el número de clases que cumpla con maximizar los valores de los índices Silhouette y Dunn y minimice el valor del índice Davies-Bouldin (los tres índices al tiempo), además de tener en cuenta que se llegue a una distribución de desertores entre clases discriminativa.

Análisis del Perfil de los Clusters

Una vez encontrado el número de conjuntos adecuado para la agrupación de los estudiantes, se le asigna un nivel de riesgo a cada clase de acuerdo al porcentaje de estudiantes desertores que contiene. Se procede a realizar un análisis de las clases extremas, es decirlos los dos grupos que tienen el mayor porcentaje de estudiantes desertores y los dos que tienen el menor porcentaje de deserción de estudiantes. Para obtener el perfil de los conjuntos en los que se clasifican los estudiantes, se realiza un análisis de los centroides de cada grupo, pues son estos los que caracterizan el grupo, es decir, para encontrar el perfil de cada clase, se debe analizar su vector central [13], como se muestra en la Imagen 8.

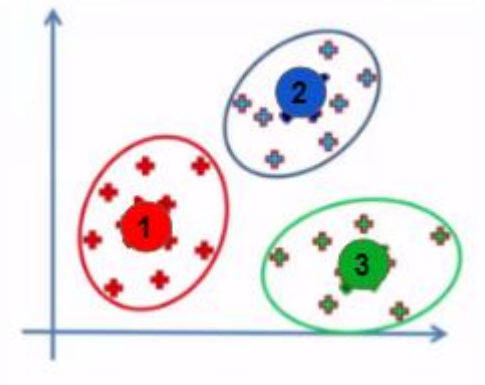


Imagen 8. Clases y sus centroides

Cada centro corresponde al elemento más representativo. Por tanto, si se analizan los valores del centro de cada grupo, se identifica el patrón que asoció a los estudiantes de ese grupo.

Este proceso se realiza con la ayuda de la aplicación PowerBI, ya que con esta es posible relacionar información de forma gráfica y tener así una mayor claridad sobre los datos.

Casos de estudio

Base de Datos

Para el desarrollo del proyecto se contó con 4 bases de datos, a continuación, se presentarán las características más importantes de cada una de estas:

Base de Datos de Bienestar: Esta base de datos contiene la información anonimizada de todos los estudiantes del departamento que han llenado la encuesta de caracterización (necesaria para acceder a los servicios de Bienestar), contiene variables de tipo Socio-Familiar, Psicopedagógico, Psicológico, Socio-Económico, Nutrición, Deporte, recreación y cultura, sexualidad y afectividad

Teniendo en cuenta el estado del arte que se realiza para la ejecución de este proyecto se seleccionan 154 variables de esta base de datos

Base de Datos 'Biblia': Esta base de datos generada cada semestre por la Facultad de Ingeniería contiene información tanto personal como académica de los estudiantes activos de la facultad.

De esta base de datos se seleccionan 49 variables.

Historia Académica Estudiantes: En esta se encuentra la información de las materias cursadas por cada estudiante de la Facultad de Ingeniería y el semestre en el que la cursó, con esto es posible determinar si algún estudiante es desertor, analizando su periodo de inactividad.

Esta base de datos es utilizada para conocer a los estudiantes desertores del departamento y así poder evaluar la eficacia del algoritmo generado

Listado Estudiantes Activos: Esta base de datos, facilitad por el Departamento de Ingeniería Electrónica y Telecomunicaciones, contiene la información actualizada de la situación de los estudiantes del departamento.

Igual que la anterior, esta base de datos es utilizada para conocer a los estudiantes desertores del departamento y evaluar el desempeño del algoritmo.

Después de la selección de las variables de acuerdo con lo hallado en el estado del arte (sin aún aplicar la propuesta de selección de variables realizada) y seleccionar los registros pertenecientes a nuestra población de interés que son los estudiantes del Departamento de

Ingeniería Electrónica y Telecomunicaciones que ingresaron antes de pandemia, se obtiene una base de datos con las siguientes dimensiones:

$$804\ sujetos\ con\ 203\ variables.$$

Casos de estudio

Los 804 sujetos que conforman el universo de estudiantes a analizar están compuestos por estudiantes de Ingeniería Electrónica (programa presencial), Ingeniería en Telecomunicaciones (programa presencial) e Ingeniería en Telecomunicaciones (programa virtual), por lo que se considera pertinente analizar los siguientes casos de estudio:

- **Caso 1:** Estudiantes de todo el departamento (804 sujetos)
- **Caso 2:** Estudiantes de modalidad presencial (831 sujetos)
- **Caso 3:** Estudiantes de modalidad virtual (53 sujetos)

Resultados y análisis

Se aplica el método expuesto anteriormente para cada caso de estudio, mostrando como resultado las variables relevantes y los grupos de variables a los que pertenecen estas, el nivel de riesgo de deserción de cada clase y el perfil de las clases extremas.

Caso 1: Estudiantes de todo el Departamento

Variables relevantes

Después de realizar el proceso para la elección de variables relevantes, se llega a 28 variables, listadas a continuación:

TRABAJAACTUALMENTE
INTENSIDADHORARIA
BORRAC_CON
SOCIALES
CHATEAR
CREDAPROBPROGPASAN_x
INTERNET_1
PIPA_CON
MARIHUA_CON
ESCRITORIO
APOYO
PROMPROG_10
MARIHUA_ED
BORRAC_ED
BORRAC_DOS
PIPA_ED
PERSONADEP
ILUMINADO
PIPA_DOS
SILLA
PROMPROG_9
VIDEO
MARIHUA_DOS
AUDIFONOS
LICOR_CON
SILENCIO

DEPORTEGRUPO
TECoLOGICO
ENERGIA_CON
QUERIDO
TELEFONO
TRABAJO

NOCHE
LEER
SEXO_1
SMARTPHONE
ENERGIA_DOS

Las cuales pertenecen a 8 grupos de variables:

Trabajo (4) Preguntas relacionadas con la tenencia de trabajo y las condiciones de este.
Drogas (12) Preguntas relacionadas con el consumo de drogas.
Tiempo Libre (8) Preguntas relacionadas con las actividades que realiza en su tiempo libre.
Académicos (3) Preguntas relacionadas con su situación académica
Ambiente de Estudio (7) Preguntas relacionadas con las condiciones en las que normalmente realiza sus actividades académicas.
Grado de estudio (1) Pregunta sobre otros estudios realizados anteriormente
Pertenece a un grupo (1) Preguntas relacionadas con la práctica de actividades extracurriculares que se realizan en grupo
Satisfacción Carrera (1) Preguntas relacionadas con la imagen que tiene de la carrera que cursa

Perfil del Estudiante

A continuación, se presenta la proporción de estudiantes desertores y no desertores para cada uno de los grupos finales.

El riesgo de deserción está dado por el porcentaje de estudiantes desertores en cada conjunto

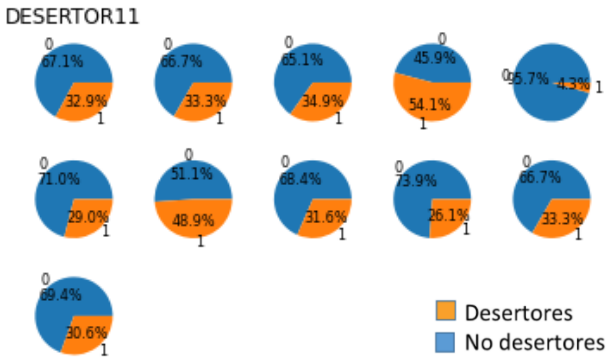
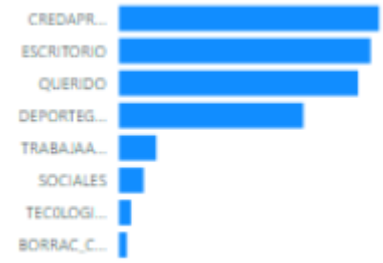
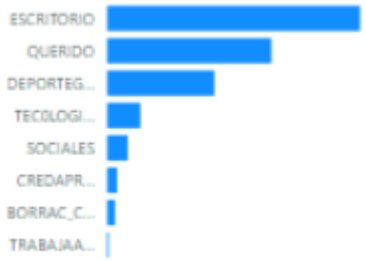
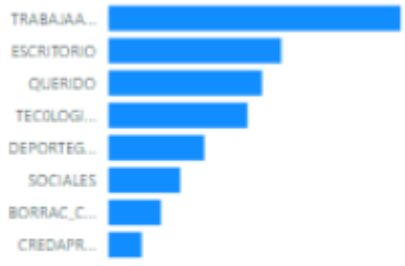


Imagen 9. Distribución de estudiantes desertores y no desertores en cada cluster – Todo el departamento

En la tabla 2, se da una descripción del perfil de los 2 grupos con menor riego de deserción y de los 2 grupos con mayor riesgo

Tabla 2. Perfil de clases extremas – Todo el departamento

Clústeres	Distribución	Descripción
C4(4,3%) Baja deserción		Estudiantes de últimos niveles que han querido estudiar la carrera, la mayoría practican deportes grupales, hay pocos tecnólogos
C8(26,1%)		Estudiantes que no consumen mucho alcohol, la mayoría cree que las redes sociales afectan su vida, pocos utilizan en escritorio y pocos trabajan
C6(48,9%) Alta deserción		Estudiantes que no trabajan, son de nivel académico bajo, creen que las redes no afectan su vida y pocos practican deporte en grupo
C3(54,1%) Alta deserción		Todos los estudiantes trabajan, casi la mitad son tecnólogos, pocos son de niveles altos, pocos practican deporte en grupo

Del perfil de los estudiantes perteneciente a las clases analizadas, llaman la atención variable cómo el nivel de la carrera en el que está el estudiante, si trabaja y si participa en deportes grupales, siendo esta última una variable que no se encuentra en el estado del

arte realizado para el proyecto, pero que evidentemente, según el proyecto es un factor importante para la deserción o no del estudiante.

Caso 2. Estudiantes de modalidad presencial

Variables relevantes

Después de realizar el proceso para la elección de variables relevantes, se llega a 26 variables, listadas a continuación:

BORRAC_CON
TRABAJAACTUALMENTE
LICOR_CON
MARIHUA_CON
SOCIALES
ENERGIA_CON
ESCRITORIO
SILLA
INTENSIDADHORARIA
INTERNET_1
CHATEAR
PERSONADEP
TRABAJO
PIPA_CON
DEPORTEGRUPO
MARIHUA_ED
BORRAC_ED
DEPORTE
ENERGIA_ED
LEER
BORRAC_DOS
APOYO
LICOR_ED
QUERIDO
AUDIFONOS
MARIHUA_DOS

Las cuales pertenecen a 8 grupos de variables:

Drogas (11) Preguntas relacionadas con el consumo de drogas.
Trabajo (4) Preguntas relacionadas con la tenencia de trabajo y las condiciones de este.
Tiempo Libre (6) Preguntas relacionadas con las actividades que realiza en su tiempo libre.
Ambiente de Estudio (2) Preguntas relacionadas con las condiciones en las que normalmente realiza sus actividades académicas.
Pertenece a un grupo (1) Preguntas relacionadas con la práctica de actividades extracurriculares que se realizan en grupo.
Deporte (1) Preguntas relacionadas con la práctica de deporte.
Satisfacción Carrera (1) Preguntas relacionadas con la imagen que tiene de la carrera que cursa

Perfil del Estudiante

A continuación, se presenta el riesgo de deserción para cada uno de los grupos, el cual está dado por el porcentaje de estudiantes desertores.

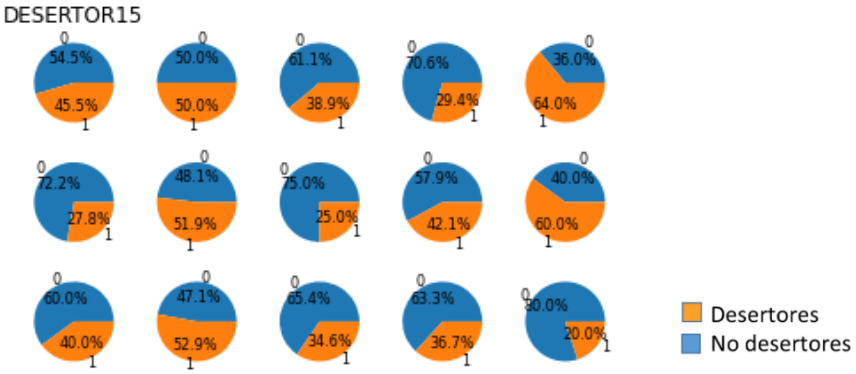


Imagen 10. Distribución de estudiantes desertores y no desertores en cada cluster – Modalidad Presencial

En la tabla 3, se da una descripción del perfil de los 2 grupos con menor riego de deserción y de los 2 grupos con mayor riesgo

Tabla 3. Perfil de clases extremas – Modalidad Presencial

Clústeres	Distribución	Descripción
C14(20,0%) Baja deserción		La mayoría practican deportes grupales, pocos trabajan, es la carrera que habían querido estudiar y tienen buenos hábitos de estudio
C7(25,0%) Baja deserción		Estudiantes que consumen mucho alcohol, creen que las redes sociales no afectan su vida, la mayoría utiliza el escritorio y pocos trabajan

C9(60,0%) Alta deserción		Practican deporte en grupo, estudiantes que consumen alcohol
C4(64,0%) Alta deserción		Pocos practican deporte en grupo, utilizan escritorio en ambiente de estudio estudiantes que no se embriagan

Analizando los perfiles encontrados, se puede ver que los perfiles de los estudiantes pertenecientes a las clases extremos son parecidos a los hallados en el caso de estudio 1, llamando la atención también es este caso la variable relacionada con la práctica de deporte grupal. Por otra parte, también se tiene coincidencia en los grupos de variables que se encontraron relevantes para identificar el nivel de deserción: Drogas, Trabajo, Tiempo Libre, Ambiente de estudio, Pertenece a un grupo, Deporte y Satisfacción Carrera.

Caso 3: Estudiantes de modalidad virtual

Variables relevantes

Después de realizar el proceso para la elección de variables relevantes, se llega a 28 variables, listadas a continuación:

INTENSIDADHORARIA
TRABAJAACTUALMENTE
PERSONADEP
APOYO
INGRESOTOTAL
EGRESOTOTAL
LEER
DNP_MUN_PADRES
SOCIALES
SILLA

DEPENDE
INGRESOSFAMILIARES
DNP_MUN_NACE
PORTATIL
PERPRUEBA
CINE
EXPOSICION
TIEMPO
TECoLOGICO
INTERCAMBIO
ESCRITORIO
ESTASISBENIZADO
VINCULACION
DEPORTIVO
ENERGIA_CON
MARIHUA_CON
SEXO
ENERGIA_ED

Las cuales pertenecen a 11 grupos:

Trabajo (4) Preguntas relacionadas con la tenencia de trabajo y las condiciones de este.
Económico (5) Preguntas relacionadas con la situación económica de la familia
Tiempo Libre (2) Preguntas relacionadas con las actividades que realiza en su tiempo libre.
Información familia y estudiante (3) Preguntas relacionadas con el sexo y el lugar de nacimiento y permanencia de ellos y sus familias
Ambiente de Estudio (4) Preguntas relacionadas con las condiciones en las que normalmente realiza sus actividades académicas.
Académicos (1) Preguntas relacionadas con su situación académica
Presentaciones Culturales (3) Preguntas relacionadas con la práctica de actividades extracurriculares que se realizan en grupo
Grado de estudio (1) Pregunta sobre otros estudios realizados anteriormente
Sisbén (1) Pregunta sobre tenencia o no de Sisbén
Elección Universidad (1) Preguntas relacionadas con el por qué eligió estudiar en la Universidad
Drogas (3) Preguntas relacionadas con el consumo de drogas.

Perfil del Estudiante

A continuación, se presenta el riesgo de deserción para cada uno de los clusters, el cual está dado por el porcentaje de estudiantes desertores.

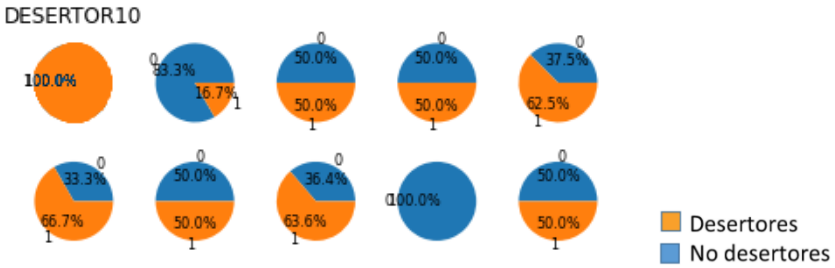
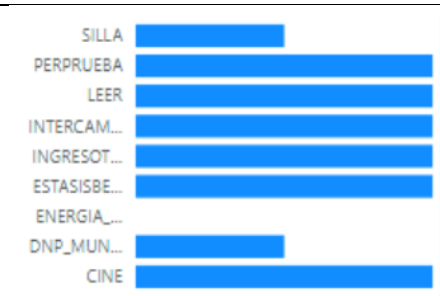
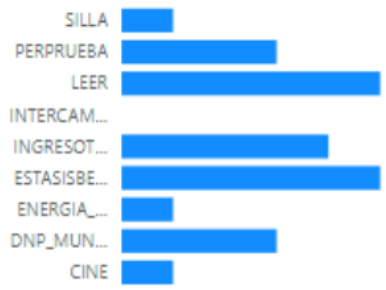
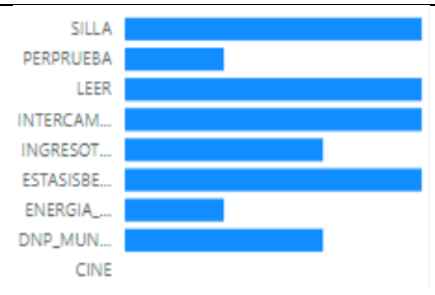
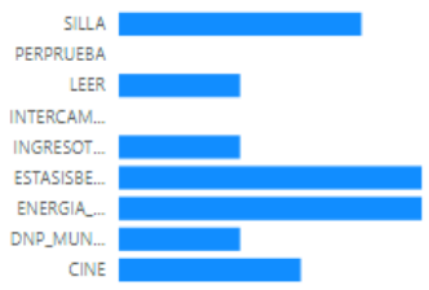


Imagen 11. Distribución de estudiantes desertores y no desertores en cada cluster – Modalidad Virtual

En la tabla 4, se da una descripción del perfil de los 2 grupos con menor riego de deserción y de los 2 grupos con mayor riesgo

Tabla 4. Perfil de clases extremas – Telecomunicaciones Virtual

Clústeres	Distribución	Descripción
8 Cluster (0%) No deserción		Están en actividades culturales (cine), les interesa un intercambio, han estado en periodo de prueba, todos son tecnólogos, nadie trabaja, ingresos familiares de $500.000 a $1 millón
1 Cluster (16,7%)		La mitad ha estado en periodo de prueba, ninguno trabaja, ingresos familiares entre $200.000 a $500.000
5 Cluster (66,7%)		No están en actividades culturales, tiene interés en un intercambio, tiene silla para estudiar, cree que leer afecta su desempeño, nadie trabaja
0 Cluster (100%)		Consumen energizante, no le interesa intercambio, no ha estado en periodos de prueba, muchos son tecnólogos, todos trabajan, ingresos familiares entre $1millón y $1,5millones

A diferencia de los resultados obtenidos del análisis de todo el departamento y en la modalidad presencial, en la modalidad virtual se encuentra una correlación más fuerte entre deserción y variables económicos que entre la deserción y las variables académicas.

Este análisis se realiza con un numero de datos muy bajo (53 sujetos) ya que de la base de datos total solo 53 estudiantes pertenecen al programa de telecomunicación virtual, por lo que los resultados pueden ser poco confiables, pues no cuenta con una relevancia estadística, sin embargo, el proceso se realizó para tener un comparativo de cuando se realice el análisis con un número significativo de muestras.

Conclusiones

- Se propuso una metodología basada en técnicas de agrupamiento que permite identificar las variables más relevantes en el fenómeno de deserción en el departamento de Ingeniería Electrónica y Telecomunicaciones.
- A diferencia de lo que existía antes, donde se clasificaba entre desertores y no desertores, los grupos resultantes permiten identificar el riesgo de deserción de los estudiantes.
- Se crearon tres modelos que permiten identificar el riesgo de deserción de los estudiantes
    - Llama la atención la relación que existe entre la pertenencia a grupos que hacen actividades extracurriculares y la no deserción. Estudiantes no tienen patrones asociados a la deserción, son estudiantes que realizan actividades extracurriculares.
    - Cómo era de esperarse, las variables relacionadas con el trabajo tienen una relación directa con el fenómeno de deserción. Se identificó que un grupo importante de estudiantes que trabajan tienen patrones de estudiantes desertores.
    - En la modalidad virtual se evidencia que la deserción está más relacionada con temas económicos que con temas académicos

Trabajos Futuros

- Crear un aplicativo más elaborado (a nivel usuario) en base a la herramienta de predicción de deserción, donde de forma automática reciba la información del estudiante y ejecute el script de predicción.
- Repetir el análisis realizado con un mayor número de sujetos para telecomunicaciones virtual.
- Plantear a bienestar la relación encontrada entre la participación de estudiantes en deportes grupales y la no deserción, para que analicen la posibilidad de realizar campañas para que aumentes estas practicas
- Verificar semestralmente la situación económica de los estudiantes de Telecomunicaciones virtual para poder hacer un acompañamiento temprano.
- Analizar el porcentaje de estudiantes que de teleco virtual pasan a programas presenciales, esto puede mostrar que la tendencia no es en realidad a desertar sino a cambiar a la modalidad presencial.

Referencias Bibliográficas

[1]: C. Díaz Peralta, "MODELO CONCEPTUAL PARA LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA CHILENA". *Estudios Pedagógicos,* vol. 34, n°2, pp. 65-86, 2008, [en línea]. Disponible en: https://www.redalyc.org/articulo.oa?id=173514136004

[2]: E. Castaño, S. Gallón, J. Vásquez, K. Gómez, "Análisis de los factores asociados a la deserción y graduación estudiantil universitaria". *Lecturas de Economía*, n°65, pp. 9-35, 2006, [en línea]. Disponible en: https://www.redalyc.org/articulo.oa?id=155213357001

[3]: E. Castañeda, Del otro al que llamamos desertor de ese otro que soy yo, [Tesis de maestría], Universidad de Antioquia, Medellín, 2010.

[4]: G. de I. Ingeniería y Sociedad, "¿La deserción y la graduación no diferencian a los programas de pregrado de la Facultad de Ingeniería de la Universidad de Antioquia?", *Ingeniería y Sociedad*, vol. 2, n.º 8, pp. 40–48, nov. 2014.

[5]: E. Castaño, S. Gallón, K. Gómez, y J. Vásquez, "Deserción estudiantil universitaria: una aplicación de modelos de duración", *Lecturas De Economía*, vol. 60, n.º 60, pp. 39–65, oct. 2009.

[6]: V. Tinto, "Definir la deserción: Una cuestión de perspectiva" *Revista de la Educación Superior*, vol. 18, n°71, sept. 1989.

[7]: Ministerio de Educación Nacional. "Deserción estudiantil en la educación superior en Colombia: Elementos para su diagnóstico y tratamiento", 1 ed., Bogotá, Ministerio de Educación Nacional, 2008.

[8]: A. Oñate, "Análisis de la Deserción y Permanencia Académica en la Educación Superior Aplicando Minería De Datos". [Tesis de maestría], Universidad Nacional de Colombia, 2016.

[9]: A. Úsuga, "La deserción estudiantil universitaria: análisis relacional del fenómeno en la Universidad de Antioquia para la cohorte 2009-I". [Trabajo de grado], Universidad de Antioquia, 2017.

[10]: Y. Salamonson, L. M. Ramjan, S. van den Nieuwenhuizen, L. Metcalfe, S. Chang, B. Everett, "Sense of coherence, self-regulated learning and academic performance in first year nursing students: A cluster analysis approach". *Nurse education in practice*, vol. 17, pp. 208–213. 2016, [en línea]. Disponible en: https://doi.org/10.1016/j.nepr.2016.01.001

[11]: B. Desgraupes, "Clustering Indices". University of Paris Ouest-Lab Modal'X, 2013.

[12]: D. Nazareno, “Índices de validación para algoritmos de agrupamiento”. [Tesis de doctorado]. Universidad Nacional del Litoral, 2019

[13]: H. H. Bock, “Clustering methods: a history of k-means algorithms”. *Selected contributions in data analysis and classification*, pp. 161-172, 2007.

[14]: H. G. Lara Gutiérrez, M. G. Lara Ruiz, V. Hernández Hernández, B. Hernández Hernández, y G. Hernández Hernández, «Análisis de un caso práctico aplicando el algoritmo K means mediante weka (Waikato environment for knowledge analysis)», ESH, vol. 4, n.º 7, ene. 2016.

Anexos

1. Instructivo de Uso Código:

Este código permite repetir la metodología propuesta para selección de variables significativas con nuevas bases de datos.

Elección Variables más Significativas

1. Cree una carpeta para realizar el proceso de forma ordenada, en esta copie los siguientes archivos:

- Normalizacion.py
  En este script se realiza la normalización de los datos ingresados
- KMeans_Graficos.py
  Script donde se realiza la ejecución del algoritmo KMeans, se almacenan los datos y se crean las gráficas de distribución de desertores por cluster
- Analisis_Codo.py
  Script donde se seleccionan las variables más relevantes, mediante la diferencia entre centroides y se encuentra el punto de inflexión
- BaseDatos.xlsx
  Archivo excel donde se encuentran los 804 estudiantes analizados con sus 203 variables analizadas
- Desertores.xlsx
  Archivo excel donde se tiene almacenada la situación del estudiante (Desertor ->1, no desertor ->0)

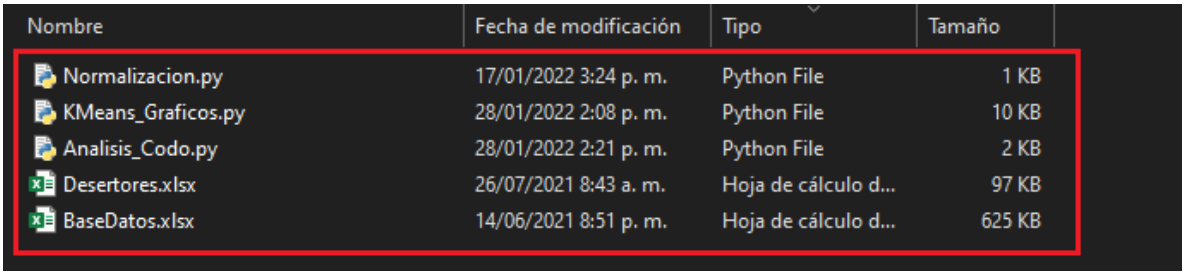


Figura 1

2. Ejecute el script **"Normalizacion.py",** este debe crear un archivo .xlsx con la información normalizada de la base de datos.

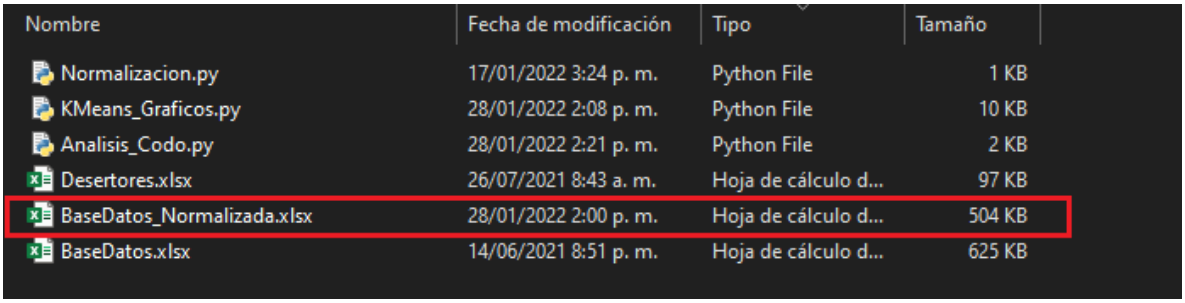


Figura 2

3. Ejecute el script **"KMeans_Graficos.py"**, este genera:

- Una carpeta donde se almacenan archivos con los centroides y *labels* encontrados durante la ejecución del algoritmo KMeans. (Fig. 3)
- Una carpeta donde se almacenan las gráficas de la distribución de desertores en cada uno de los *clusters*. (Fig. 3)

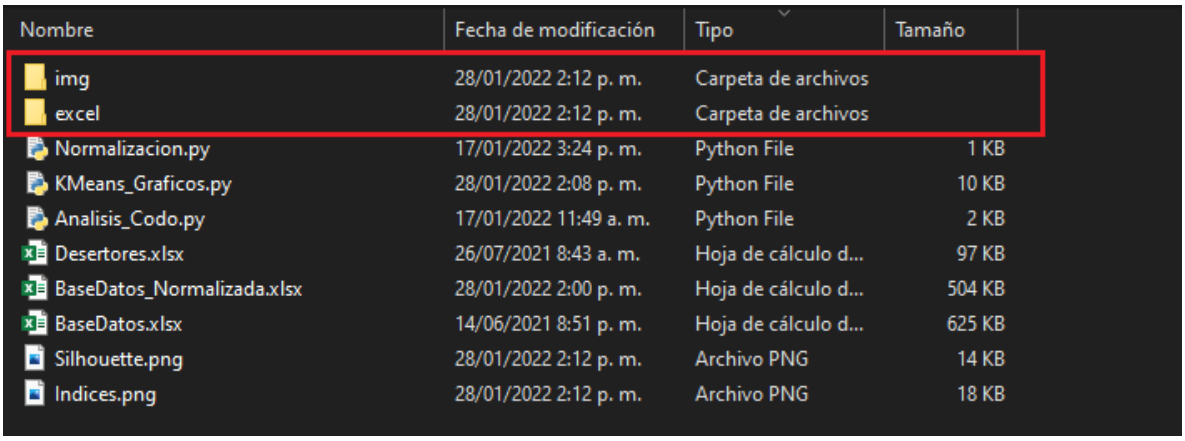


Figura 3

- Una lista con las agrupaciones más discriminativas (que permite separar mejor entre desertores y no desertores) en orden descendente. (Fig. 4)
  Nota: La lista está compuesta por el número de clusters, es decir, si aparece el numero 25, es por que cuando se agrupo en 25 clusters se llegó a la distribución más discriminativa

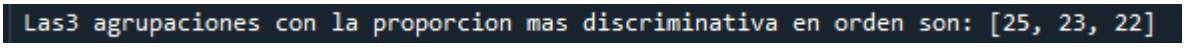


Figura 4

- Graficas de los índices Silhouette y DaviesBouldin-Dunn, estos dan la calidad de la agrupación

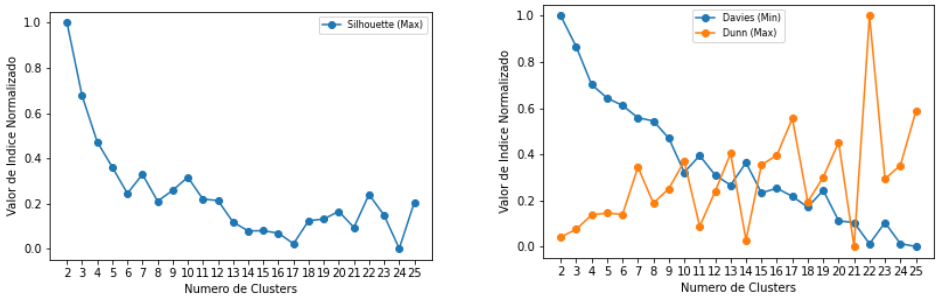


Figura 5.

4. Analice cual es la mejor agrupación, para esto básese en:
   - El orden dado por la lista de las agrupaciones más discriminativas, pues en ese orden se tiene una mayor discriminación entre desertores y no desertores.
   - Las gráficas de los índices Silhouette y DaviesBouldin-Dunn, que arroja el Script **"KMeans_Graficos.py"** cuando termina de ejecutarse.
     Elija el número de Clusters que mejor cumpla con ambos criterios, Silhouette (alto), Davies (Bajo), Dunn (Alto).

     Por ejemplo, para la Figura 5, el número de clústeres ideal sería 2, sin embargo, al analizar la distribución (Figura 6) nos damos cuenta de que no se llega a una distribución discriminativa, por esto se descarta. Analizando los índices, podemos ver que otra agrupación buena (Silhouette->alto, Dunn->Alto, Davies->Bajo) se encuentra con 10 clusters, entonces se analiza la distribución con 10 (Figura 7), y se concluye que esta si es una buena distribución porque hay clusters con alto número de desertores y otros con baja cantidad (El C4 tiene un 21% y el C8 tiene un 55,6%).

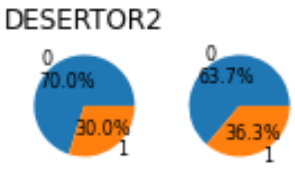


Figura 6.

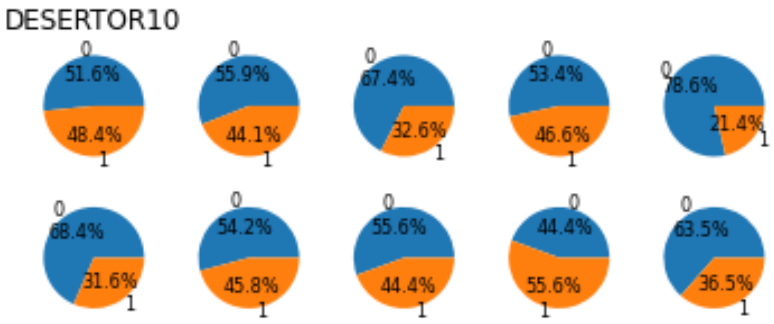


Figura 7

5. En el script **"Analisis_Codo.py"**, modifique la línea 39:

   centroids=pd.read_excel("excel/centroids2.xlsx",index_col=0)

   cambiando el número que acompaña a 'centroide' por el número de clusters que usted encuentra apropiado en el ítem 4, por ejemplo, si eligió 5 clusters debe escribir:

   centroids=pd.read_excel("excel/centroids5.xlsx",index_col=0)

```
35
36  #---------- Definicion de variables --------------------
37
38  #DataFrames
39  centroids=pd.read_excel("excel/centroids2.xlsx",index_col=0) # DF con los centroides de la agrupacion elegida
40  bd=pd.read_excel("BaseDatos_Normalizada.xlsx",index_col=0) # DF con la BD normalizada
41  bd_new= pd.DataFrame() # DF donde se almacenaran las columnas significativas
42  bd_new['a']=range(0,len(bd)) # Se especifica el numero de registros con que se trabajara
43
```

Figura 8

6. Ejecute el script **"Analisis_Codo.py"**, este genera un archivo .xlsx con las variables más significativas, en orden de relevancia (La primera columna corresponde a la variable de mayor relevancia).

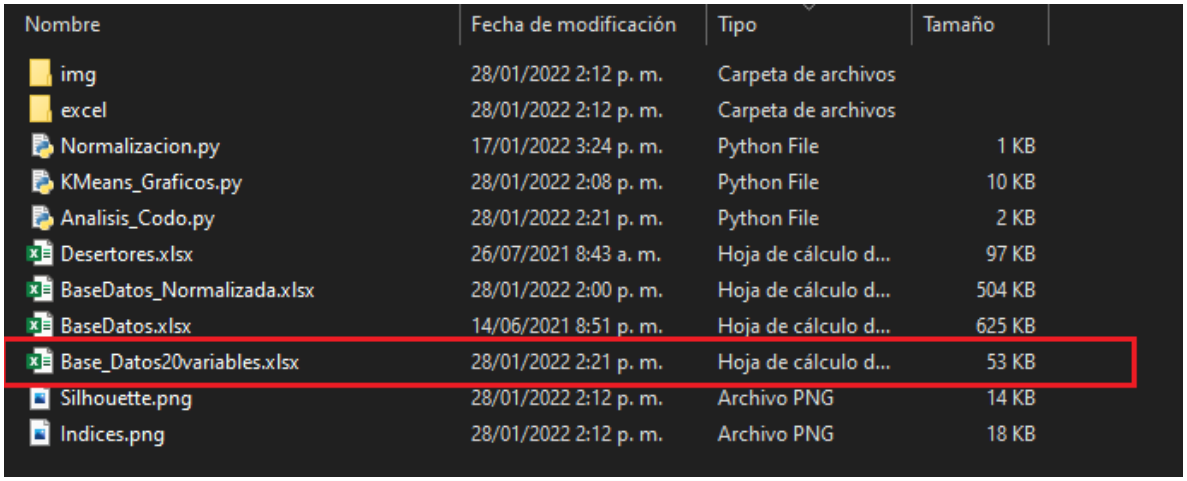


Figura 9

7. Repita el proceso (sin realizar el ítem 2), reemplazando el archivo 'BaseDatos' por el archivo resultante del ítem 6.

   Este proceso se repetirá hasta que se llegue a una de las dos situaciones mencionadas a continuación, en este caso se tomará la base de datos utilizada en el proceso anterior:

   - El criterio del codo no aplique (esto será cuando el codo este ubicado en los extremos, incluya a todas las variables o ninguna)

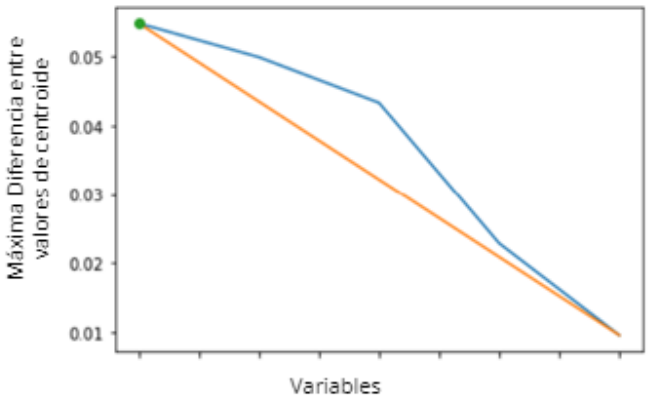


Figura 10

Nota: La Figura 10, representa la máxima diferencia que existe entre los valores de los centroides para cada una de las variables, estas variables están ordenadas de dorna descendente de acuerdo a la diferencia.

   - Las distribuciones a las que se llegan con los diferentes números de clusters no son discriminativas: No exista una diferencia notable entre la distribución de desertores y no desertores en los clusters.

8. Cuando encuentre las variables más relevantes, ejecute nuevamente el algoritmo KMeans, usando el script **"KMeans_Graficos.py"** alimentándolo con la base de datos que contiene a las variables relevantes, analice la distribución de desertores y los índices de calidad de clusters; el número de clusters que tenga mejor estas características será el elegido para el modelo.

2. Herramienta para la Predicción de Riesgo de Deserción

Esta herramienta permite, basada en los modelos identificados con las bases de datos del Departamento de Ingeniería Electrónica y Telecomunicaciones, permite identificar el riesgo de deserción que puede tener un estudiante.

Contexto Teórico:

Predicción: Una vez caracterizado cada *cluster*, es posible conocer el grado de pertenencia que un estudiante nuevo tiene a cada *cluster*, esto se conoce cómo predicción.

Este proceso se realiza calculando la distancia del nuevo sujeto a los centroides de todos los *clusters*, el sujeto será asignado al *cluster* con el que tenga la distancia menor

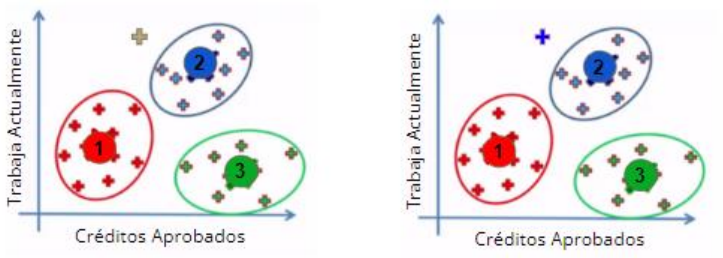


Figura 1.

Uso de la herramienta

La herramienta se se encuentra en una carpeta llamada 'Prediccion', la cual contiene tres subcarpetas, una por cada modelo:

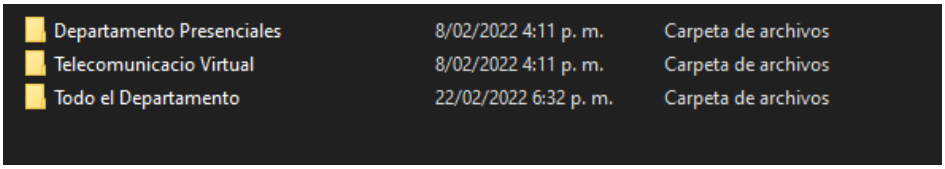


Figura 2.

Se debe elegir bajo que modelo se analizara el estudiante, una vez elegido, se ingresa a la carpeta (para esta ejecución se analizara al estudiante con el modelo de Todo el Departamento), encontrando los siguientes archivos:

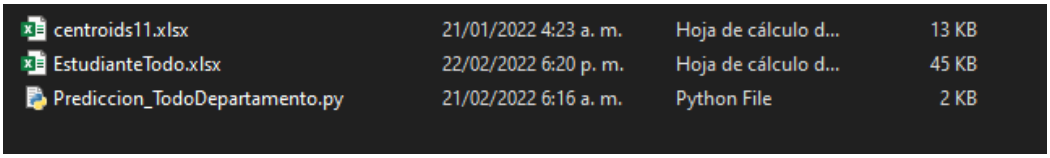


Figura 3.

En esta carpeta, igual que en las otras carpetas de los modelos, se encuentran tres archivos:

- 'centroids11xlsx': Que es el archivo que contiene en si el modelo, el valor de cada uno de los clusters
- 'EstudiantesTodo.xlsx': En este archivo se ingresará el perfil del estudiante, es decir, las respuestas que de las preguntas relacionadas a cada variable.

Variable	Pregunta	Cuantificación	Valores para el Estudiante
TRABAJAACTUALMENTE	¿Trabaja actualmente?	SI/NO → 1/0	0
INTENSIDADHORARIA	Seleccionar la cantidad más frecuente de tiempo que dedica al trabajo	Por horas (menos de medio tiempo) → 1 Medio tiempo → 2 Tiempo completo → 3 No aplica → 0	0
BORRAC_CON	¿Consumió esta sustancia en el último año?	SI/NO → 1/0	0
SOCIALES	¿Considere que el uso de redes sociales han afectado cualquier aspecto de su vida, como salud, tiempo, dinero, relaciones familiares, sociales, académicas o laborales en los últimos seis meses?	SI/NO → 1/0	1
CHATEAR	¿Considere que el chatear ha afectado cualquier aspecto de su vida, como salud, tiempo, dinero, relaciones familiares, sociales,	SI/NO → 1/0	

Figura 4.

- 'Prediccion_TodoDepartamento': este script es el que se debe ejecutar una vez se tenga la información de estudiante para obtener así, la predicción.

Una vez se realiza la ejecución del script, en la consola saldrá un mensaje indicando los *clusters* más cercanos al estudiante, y también especifica cual es el *cluster* con perfil más cercano al estudiante para conocer mejor la situación del estudiante y si este queda en deserción las variables a las que se debería prestar más atención.

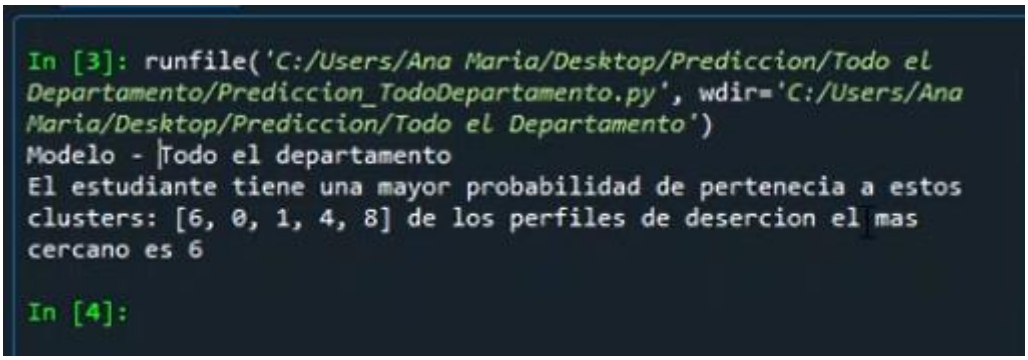


Figura 5.

Repositorio del proyecto

https://github.com/AnaGiraldoMarin/Proyecto_Desercion.git