Análisis de Determinantes y Predicción de Rendimiento de los Estudiantes en Pruebas Nacionales Saber

Daniel Hernando Pájaro Romero

Universidad Tecnológica de Bolívar, UTB

Cartagena, Colombia

danielpajaro@hotmail.com

John Javier Hernández Araujo

Universidad Tecnológica de Bolívar, UTB

Cartagena, Colombia

johnj.hernandez.araujo@gmail.com

***Abstract* — This document presents the study of students´ performance in Colombia base on the results of the mandatory national test Saber 11°, taking into account**   
**several of its socioeconomic and geographical characteristics, and prediction of possible performance of future students using an Machine Learning algorithm.**

***Keywords — PRUEBAS SABER 11, Machine Learning***

***Resumen* — Este documento presenta el estudio del desempeño de los estudiantes en Colombia en los resultados de las pruebas obligatorias nacionales Saber 11°, teniendo en cuenta varias de sus características socioeconómicas y geográficas, y la realización de predicción del posible desempeño de futuros estudiantes utilizando un algoritmo de Machine Learning.**

***Palabras clave — PRUEBAS SABER 11, Machine Learning***

# **Introducción** .

Las Pruebas Saber son evaluaciones realizadas por el Instituto Colombiano para la Evaluación de la Educación - ICFES, que deben rendir todos los alumnos en Colombia al finalizar el curso lectivo, de acuerdo a las pruebas establecidas. Consiste en una prueba de carácter nacional en la que todos los estudiantes del país que cursen los niveles exigidos por las pruebas, deberán presentarse a realizar la prueba. Es una vía para evaluar periódicamente la educación en Colombia. La modalidad de la evaluación consiste en responder preguntas de opción múltiple y, con un cambia reciente en 2018, preguntas de libre respuesta. Las pruebas SABER 11° son las pruebas específicas para la evaluación del nivel de la Educación Media, que a partir del año 2014 se alinea con las evaluaciones de la Educación Básica para proporcionar información a la comunidad educativa en el desarrollo de las competencias básicas que debe desarrollar un estudiante durante el paso por la vida escolar. Esta se realiza a los estudiantes que están en su último grado escolar (grado 11°) antes de su graduación.

El ICFES lleva registro de todos los datos de los alumnos que han presentado las pruebas, como las condiciones socioeconómicas, instituciones educativas, cantidad de familiares, condiciones de vida, etc. Así como los resultados obtenidos por cada uno de ellos en las pruebas.

A continuación, se evidencia el análisis de la relación de cada uno de estos datos con el rendimiento de los estudiantes en las pruebas, así como se pretende realizar una clasificación de los resultados de futuros estudiantes al realizar la prueba partiendo de estos datos ya establecidos, utilizando algoritmos de Machine Learning; kmeans para la creación de grupos de interés y posteriormente knn para la clasificación de estos.

# **Objetivos** .

1. Crear un modelo de clasificación que permita segmentar el desempeño global de los estudiantes en las pruebas Nacionales saber 11.
2. Identificar cuáles son las características que representan los grupos de interés, partiendo de los elementos centrales de cada uno de los mismos.

# **Investigaciones y estudios previos**

* 1. *El logro académico y el efecto de colegios en las pruebas ICFES en Antioquia.*

*Piedad Patricia Restrepo y Mauricio Alviar (2004)*

Resumen: esta investigación muestra que, a través de un análisis de estudio de los resultados de la prueba del año 2004 y de cómo se relacionan los colegios participantes y las condiciones de los estudiantes con ellos y si son algunos de los factores que incidieron en sus resultados de las pruebas. Al final, se presentan algunas propuestas para mejorar el diseño de este tipo de programas que permitan alcanzar los puntajes objetivos planteados, teniendo en cuenta la cobertura, calidad y equidad.

La metodología utilizada fue un análisis de regresión multivariado. contemplaron utilizar dos formas de abordarlo. De un lado, la desagregación, que consiste en llevar todas las variables al nivel del individuo, y la otra opción utilizada es la agregación, que consiste en promediar todas las variables al nivel de grupo o colegio. Ninguna de estas opciones resulta óptima para hallar los En el primer caso, se viola el supuesto de independencia de las observaciones al desconocer el fenómeno de anidación del proceso educativo. Los alumnos que pertenecen al mismo grupo o al mismo colegio, comparten ciertas características que influyen en el resultado, usualmente medido a través de los puntajes en pruebas estandarizadas, como proxy de las de la eficacia escolar. Al violar el supuesto de independencia, la estimación de los errores estándar no es óptima. En el segundo caso, la agregación produce la pérdida de información valiosa.

Los modelos jerárquicos o multinivel, se han convertido en una alternativa superior en los estudios de la eficacia escolar, y en general, en fenómenos de cualquier índole en los cuales se presente anidación; Esto es, los individuos pertenecen a grupos, y esos grupos a su vez pertenecen a unidades de mayor tamaño, y así, sucesivamente.

Llegan a la conclusión que resultados de los estudiantes antioqueños con respecto a la mayoría de departamentos del país no permiten augurar un futuro muy prometedor en términos de igualdad de oportunidades y bienestar para los antioqueños. Sin embargo, no se pudieron establecer cuáles son los mecanismos a través de los cuales opera la brecha entre logro medio de subregiones, pero lo atribuyen a 3 posibles causas:

* Falta de insumos escolares
* La pobreza
* El nivel de actividad
  1. *Determinants of the performance of the schools in Medellin in the high-school graduation-year test (ICFES)*

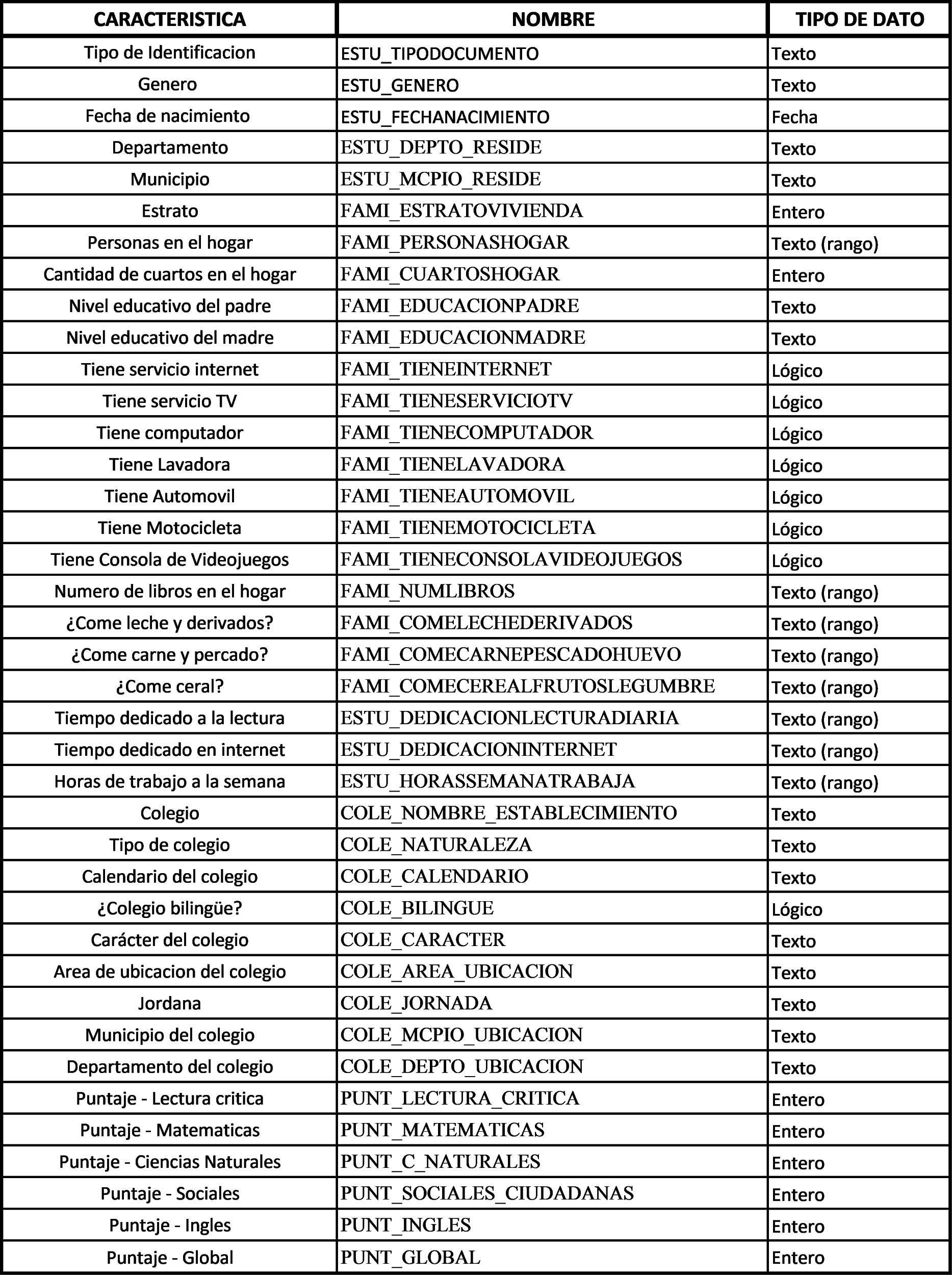
*David Tobón Orozco, Héctor Mauricio Posada Duque y Paul Ríos Gallego (2009).*

Este documento analiza la manera en que los procesos institucionales, la dirección de los colegios y las circunstancias del colegio y de sus estudiantes pueden afectar el rendimiento en la prueba del ICFES en el grado 11. Se aplicó un cuestionario que verifica las acciones e interacciones entre los actores que participan en el proceso educativo. Se utilizaron modelos de análisis factorial y modelos econométricos y jerárquicos para construir los índices y medir el impacto de los variables considerados. Se encontró que las variables de niveles de los colegios, incluida su administración, explican un alto porcentaje de la variabilidad en el rendimiento, mientras las variables construidas en asociación con las características del estudiante tienen un impacto bajo. El capital humano de los padres de familia y las condiciones iniciales del estudiante son mucho más importantes que el capital humano de los docentes. Contrario a lo esperado, no hay claridad sobre qué tipos de uso o de aumento de recursos de infraestructura podrían mejorar el rendimiento.

# **Datos de los estudiantes**

Los datos con los cuales se trabajará serán los de 20.000 estudiantes aleatorios que presentaron las pruebas en el segundo semestre de 2018. Estos datos fueron obtenidos de la base de datos oficial del Ministerio de Tecnologías de la Información y Telecomunicaciones (MinTIC), a través de su portal. Esta base de datos cuenta con la siguiente información de los estudiantes. Estos datos son los mostrados en la siguiente página (Tabla 1)

TABLA I. Datos obtenidos de los estudiantes

.

Fuente: Datos Abiertos Colombia, Resultados del examen Saber 11 para el segundo periodo del año 2018. sitio web: https://www.datos.gov.co/Educaci-n/Saber-11-2018-2/m2nt-jw2h

# **Depuración de los datos**

Debido a la cantidad de datos, se decidió eliminar características consideradas irrelevantes o redundantes a la hora de realizar el análisis. Así mismo se decidió eliminar variables cuyos valores eran incorrectos y/o incompletos. Las variables resultantes son 19, y son las siguientes:

* Genero del estudiante: transformada en variable binaria indicando si es genero masculino 1, o femenino 0.
* Edad
* Personas en el hogar
* Cuartos en el hogar
* Tiene computador
* Tiene internet
* Tiene lavadora
* Tiene motocicleta
* Tiene consola videojuegos
* Hrs de dedicación lectura: representada así, 0 No lee por cuenta propia,1 30 minutos o menos ,2 entre 30 a 60 minutos, 3 entre 1 a 2 horas, 4 más de 2 horas.
* Hrs de dedicación internet: 0 no navega en internet, 1 30 minutos o menos, 2 entre 30 y 60 minutos,3 entre 1 a 3 horas, 4 más de 3 horas.
* Estrato de vivienda: estrato económico de su vivienda desde 0 a 6.
* Tiene servicio TV
* Tiene automóvil
* Numero de libros: representada del 1 al 5 de la siguiente manera:

1 de 0 a 10 libros, 2 de 11 a 25,3 26 a 100 y 4 más de 100.

* Naturaleza del colegio estudio: Oficial o publico 1, de lo contrario 0.
* Colegio bilingüe
* Área de ubicación del colegio: binaria,1 si pertenece a área urbana.
* Puntaje global: puntaje final de las pruebas que va de 1 a 500.

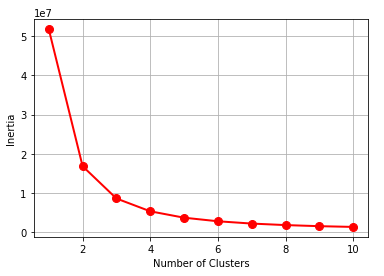
1. **METODOLOGIA**

**Clustering (Creación de grupos de interés).**

Para la aplicación de este modelo se procedió a realizar la agrupación de los datos a través del método de clustering conocido como Kmeans, este algoritmo crea un conjunto de grupos o clústeres a partir de las similitudes que se presentan entre los datos. El primer paso para aplicar este algoritmo consiste en identificar el número de clúster más apropiado para agrupar los datos; un método muy utilizado para dicha tarea es el grafico de codo o Elbow Plot, dicho método nos permite observar la varianza entre los errores generados por la aplicación del algoritmo de clustering para n números de clústers.

El número máximo de clústers escogidos para realizar dicha evaluación fue de 10, obteniendo así el siguiente gráfico.

Imagen I. Datos obtenidos del elbow en clustering.

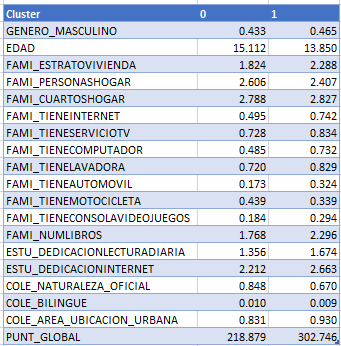


Como se observó en el grafico la variación entre los errores generados (Inercia) para los distintos números de clústers deja de ser notoria entre los valores 2 y 3, criterio comúnmente aceptado para la escogencia del número de clusters, por lo que se procederá a reanudar el clustering con 2 centroides.

Obtenidos los centroides, se procede a pasar como parámetro a la función kmeans para posteriormente hallar los centroides que son representativos de los grupos creados y el clusters al que pertenecen cada registro.

Para finalizar, es necesario realizar un minucioso análisis de los valores que se marcan con respecto a las variables previamente listadas.

Tabla 2. Representación de centroides obtenidos por clusters.



**Clasificación por medio de knn:**

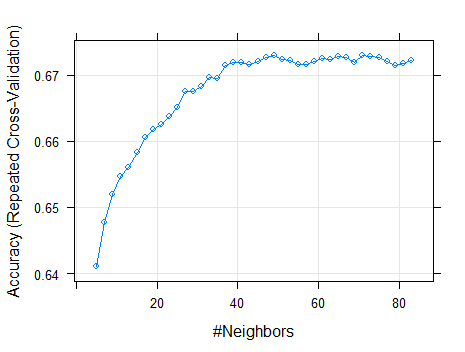
Una vez se tienen los clústeres a los que pertenece cada registro de la base de datos podemos usar un algoritmo de clasificación para construir un modelo que permita predecir el clúster al que pertenecería un estudiante con base al valor de sus variables.

Se usarán 2 algoritmos de clasificación y se usara el 80% de los datos para entrenamiento y el 20% para probar su precisión.

En primer lugar, se usará Knn o clasificación por medio de los vecinos más cercanos, este algoritmo calcula la distancia entre los distintos puntos e intenta predecir la clase a la que pertenece un dato teniendo como base la clase de los k vecinos más cercanos o semejantes.

La función knn recibe como parámetros el número de vecinos con los que se van a comparar, para hallar el número ideal de vecinos se usara la función “caret”, de la cual se obtuvo la siguiente gráfica a través del plot():

Imagen II. Grafica obtenida del modelo caret.



Como se observa en la gráfica el número ideal de vecinos para comparar es de aproximadamente 50 debido a que con los datos de entrenamiento este fue el que permitieron una mayor precisión, La función caret automáticamente escoge este numero optimo de vecinos para comparar y nos ofrece el modelo listo para ser usado y poner a prueba. Con este modelo se clasificará los datos, las etiquetas a utilizar como se dijo anteriormente, fueron los clústeres previamente creados. Con base al modelo desarrollado a partir de los datos de entrenamiento se intentará predecir la etiqueta de los datos de prueba y calcular la correspondiente precisión para verificar la eficacia del algoritmo para clasificar nuevos datos.

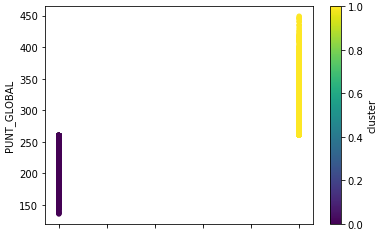
**Clasificación por medio de Regresión logística:**

Posteriormente se procedió a realizar un modelo de clasificación usando el algoritmo de regresión logística, dicho algoritmo es muy útil cuando se requiere clasificar datos entre dos posibles clases, para este caso cluster 1 y cluster 2, la librería Caret en el lenguaje R también ofrece la construcción de un modelo de regresión logística, por lo cual se construirá el modelo con dicho método, en su interior el modelo de regresión logística construye una línea o curva para separar los datos pertenecientes a la clase positiva y negativa y hace uso de la función sigmoide para convertir el resultado de los datos en el rango de 0 a 1 para luego clasificarlo en una de las dos clases.

1. **RESULTADOS Y ANALISIS DE DATOS:**

En cuanto al resumen de los datos que se especifican en los centroides de ambos clúster generados por el algoritmo de kmeans, se puede determinar el orden que maneja el modelo una vez se han localizado los clústers. Estos toman un valor especifico, basado en un resumen para cada variable respecto al valor central que las representa en un conjunto de datos, el cual parte de la previa clasificación de 2 grupos que manejan rangos de puntaje específicos en cuanto a las pruebas saber 11.

Imagen III. Esparcimiento de puntaje global



Una vez establecido el modelo de los centroides, es posible partir de la variable del puntaje global como punto de referencia para ver el comportamiento de las otras variables y proceder a su interpretación para notar como la proporción de cada una interviene en el resultado final, ya que como se puede observar en el grafico el puntaje global promedio para cada clúster tiene una diferencia notoria de 84 puntos.

Siguiendo el mismo orden de la escala gráfica, se pueden evaluar otras variables:

Imagen IV. Diagrama de barras apiladas modelado en Excel.

(ya tiene los datos del centroide en Python)

Del anterior diagrama se puede determinar el factor de incidencia del transporte en cada uno de los clústers, en el cluster 2 al que pertenecen estudiantes con una media de puntajes más alto se hace más notorio la posesión de automóviles los cuales requieren claramente de un poder adquisitivo mayor, mientras que las motocicletas no presentan una variación considerable

En caso tal, una variable que podría llamar bastante la atención de los analistas es ver como incide la clase de escuela en este tipo de ejercicios:

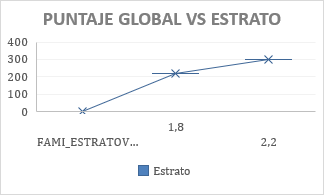
Imagen V. Diagrama de barras apiladas modelado en Excel.

(no esta actualizado con los nuevos datos de python)

Como se observa en la grafica al pasar al clúster numero 2(puntaje promedio mas alto) el numero de estudiantes pertenecientes a escuelas privadas o no oficiales aumenta, de la misma manera en la que el porcentaje de estudiantes en escuelas rurales que entran a esta clasificación empieza a disminuir, lo anterior puede sugerir que las condiciones de las instituciones educativas en zonas urbanas son mejores que las rurales, así como las instituciones privadas sobre las públicas u oficiales.

En el caso particular de los estratos sociales, también se puede convertir en un punto de comparación considerable en el estudio:

Imagen VI. Gráfico de tipo cotizaciones modelado en Excel.



Lo notorio es que a medida que crece el valor medio en los estratos, hay una tendencia a un mayor puntaje, lo que permite categorizar de tal manera los grupos formados en los clusters. Partiendo de esto, se podría deducir que esta es una variable bastante proporcional en el modelo del puntaje global.

Otros aspectos importantes para resaltar que se pueden observar en la tabla 2, caracterización de los centroides, es que el numero de horas dedicadas a internet, a la lectura, así como el numero de libros con los que cuenta la familia, tienen un valor promedio superior en el clúster número 2.

##### **análisis del modelo de clasificación**

La métrica escogida para determinar cuál era el modelo más eficiente fue la precisión o tasa de acierto de los algoritmos,

Al usar el modelo de Knn comparando con alrededor de 50 datos más próximos, se obtuvo una precisión de 67.525%, por otro lado, al realizar el modelo por medio de regresión logística la precisión obtenida fue de 70,2% es decir presento una mejora con respecto al algoritmo de Knn.

El modelo de regresión logística resulto más eficiente y este además proporciona un resumen de las variables más significativas para llegar a dicho modelo que para esta investigación fueron las siguientes variables:

El género del estudiante, la edad, Estrato socioeconómico, número de personas en el hogar, si se contaba con internet, computador, motocicleta y automóvil, el número de libros con los que su familia, las horas de dedicación a la lectura y el internet, así como la naturaleza oficial del colegio y la zona en la que el estudiante realizaba sus estudios. Confirmando así que si existen ciertos factores como el analizado en los centroides de los clúster que de cierta manera contribuyen al desempeño de los estudiantes en las pruebas de estado Saber 11°.

**CONCLUSIONES**

A partir de la investigación realizada se pudo identificar que en el desempeño de los estudiantes afectan diversas características socioeconómicas en las que viven y características propias de las instituciones educativas a las que pertenecen, denotando así que la calidad de la educación y la situación económica de los estudiantes si tienen un efecto directo en el resultado en las prueba Saber 11°.

Otro hallazgo importante de la investigación fue la construcción de un modelo de clasificación que permitiera conocer al grupo al que un estudiante pueda pertenecer, teniendo en cuenta algunas de las características expuestas anteriormente, dicho modelo logro clasificar con un 70% de precisión, y el uso de mas de 1 modelo de clasificación destaco que pueden existir modelos que predicen o clasifiquen de mejor manera que otros ciertos conjuntos de datos, al ajustarse mejor a los datos de estudio.

**REFERENCIAS**

* M. Barberan, "ICFES: ¿Qué son las Pruebas Saber?", Rankia,2018. [Online]. Sitio web: <https://www.rankia.co/blog/mejores-opiniones-colombia/4115573-icfes-que-son-pruebas-saber>.
* Tobón Orozco, D., Posada Duque, H. and Ríos Gallego, P. (2009). Determinants of the performance of the schools in Medellin in the high-school graduation-year test (ICFES). 38th ed. Medellín: Universidad de Antioquia.
* Restrepo, P. and Alviar, M. (2004). El logro académico y el efecto colegio en las pruebas ICFES en Antioquia. 1st ed. Medellín: Ecoe Ediciones