Análisis de Segmentación para estudiantes que realizaron las Pruebas Saber 11 en el año 2020

Federico Berbesi Vasquez
Carolina Patricia Castellano de la Cruz
Jose Ignacio Rengifo Villegas
Sandy Lorena Rojas Giraldo

Universidad de los Andes
Departamento de Ingeniería Industrial
Maestría en Inteligencia Analítica de Datos – (MIAD)

Bogotá D.C (Colombia)
25 de septiembre de 2024

1. Resumen

Este estudio analiza los datos de estudiantes que presentaron las pruebas Saber 11 en Colombia durante el año 2020 utilizando técnicas aprendizaje no supervisado. Se realizó el análisis de componentes principales (PCA) para reducir la dimensión de los datos y el algoritmo DBSCAN para la segmentación. El análisis identificó dos clusters estudiantes después de aplicar DBSCAN, ademas se removieron los outliers. Las columnas que más contribuyeron en la segmentación fueron la edad, el período de las pruebas, el departamento de residencia y el tipo de institución educativa. Los resultados muestran que existe una correlación considerable entre las características socioeconómicas y el rendimiento académico de los estudiantes. Estos hallazgos pueden ser útiles para políticas que busquen reducir la desigualdad en el sistema educativo colombiano.

2. Introducción

El examen Saber 11 es un test esencial para evaluar la calidad de la educación en el país. Esta prueba mide las competencias de los estudiantes en: matemáticas, ciencias naturales, lectura crítica, sociales y ciudadanas, e inglés. Los resultados obtenidos por los estudiantes no solo permiten su acceso a la educación superior, sino que también muestran las condiciones educativas y socioeconómicas de desigualdad.
Por esta razón es importante entender los factores que influyen en los resultados de los estudiantes, como características demográficas y socioeconómicas y el colegio al que asisten.
Mediante el análisis no supervisado es posible encontrar patrones en grandes bases de datos (complejos y ata dimensionalidad), por lo que estos resultados son un gran candidato para este tipo de análisis.
A través del uso de herramientas como PCA y DBSCAN, se busca identificar grupos de estudiantes con características comunes para ver como esto afecta su rendimiento en la prueba.
Se quieren encontrar posibles patrones de desigualdad en el acceso a oportunidades educativas y si ciertos estudiantes se encuentran en una posición de alta desventaja respecto a otros.
El objetivo de este estudio es ofrecer un análisis profundo del rendimiento académico de los estudiantes en Colombia mediante la implementación de técnicas de análisis de datos. Los hallazgos de este trabajo ayudan a comprender los factores que afectan al desempeño de los estudiantes y pueden servir de base para investigaciones futuras para implementar políticas educativas y mejorar la equidad en el acceso a oportunidades de aprendizaje en Colombia.
Para esto fue necesario realizar un análisis de la literatura disponible, que se encuentra a continuación:

Desde 1960, se han desarrollado modelos de pruebas que buscan evaluar la calidad de la enseñanza en competencias transversales. Organizaciones como la OCDE han investigado el concepto de Valor Agregado mediante pruebas estandarizadas en áreas como matemáticas, lectura, escritura y ciencias (OCDE, 2014).

Los primeros estudios sobre las calificaciones de estudiantes en instituciones de educación superior, que integran variables socioeconómicas para detectar relaciones y comportamientos inusuales entre alumnos, secciones y profesores, se remontan a 2007 con la investigación de Chue Gallardo, Barreno y Millones Rivalles. Este estudio utilizó técnicas estadísticas como análisis de correspondencias simples, análisis de clúster, análisis de covarianza, análisis discriminante y regresiones logísticas binaria, ordinal y nominal.

En 2010, Ibarra y Michalus analizaron el rendimiento académico de estudiantes mediante modelos de análisis multivariado. Con Regresión Logística, encontraron que el desempeño de los estudiantes de ingeniería de las cohortes 1999-2003 estaba fuertemente influido por el tipo de institución de nivel medio donde cursaron sus estudios, las calificaciones obtenidas y el número de asignaturas aprobadas en el primer año de carrera.

El estudio de Pérez-Pulido, Aguilar-Galvis, Orlandoni-Merli y Ramoni-Perazzi (2016) empleó un análisis de correspondencias múltiples, factorial y regresión cuantílica para diagnosticar a los estudiantes de la Universidad de Santander. Este diagnóstico se utilizó para implementar planes de mejoramiento académico enfocados en aquellos con bajos puntajes en las pruebas Saber 11. El perfil de mayor riesgo académico incluía estudiantes de estratos bajos, con ingresos familiares entre 1-2 SMLV, provenientes de familias numerosas, sin vivienda propia y de regiones fuera de Santander.

Ruiz Escorcia et al. (2017) aplicaron el primer Análisis de Componentes Principales (ACP) a los resultados de la prueba Saber 11, y encontraron que, en 2012, las instituciones no oficiales de Barranquilla obtuvieron, en su mayoría, mejores resultados en estas pruebas.

Madrid (2017) utilizó diversas técnicas estadísticas multivariadas, como Máquinas de Soporte Vectorial, Análisis Discriminante, K-vecinos más próximos y Regresión Logística, para clasificar a los estudiantes de la Universidad Nacional de Colombia, Sede Medellín, según su probabilidad de deserción.

En 2020, Cerón Benavides, Romero Ospina y Fonseca Gómez identificaron los factores clave en los resultados de las pruebas Saber Pro para estudiantes del programa de Ingeniería de Producción Acuícola de la Universidad de Nariño (2016-2019) mediante un análisis por componentes principales, junto con regresiones múltiples y cuantílicas. Factores como la comunicación escrita, el razonamiento cuantitativo y las competencias ciudadanas fueron esenciales para el modelo.

Chan & Galli (2020) exploraron el uso del lenguaje R y técnicas estadísticas multivariadas en la selección y análisis de investigaciones educativas a nivel superior.

Rodríguez Rodríguez y Hernández Cano (2021) emplearon el modelo Random Forest para identificar las variables más influyentes en los resultados de las Pruebas Saber 11 de 2017-2018, destacando que el acceso a internet, a un computador y la cantidad de libros en el hogar estaban estrechamente relacionadas con la información socioeconómica de los estudiantes.

En 2021, Acosta Solano et al. concluyeron que el modelo Logistic Model Tree (LMT) fue el más efectivo para predecir el rendimiento sobresaliente de los estudiantes en la región Caribe, usando los resultados de la prueba Saber 11 (2017-2019). Identificaron que la naturaleza de la escuela, el acceso a internet y la jornada escolar influyen significativamente en el desempeño de los estudiantes.

Rodríguez Manrique, Ruiz Escorcia y Cohen Manrique (2022) emplearon análisis de clúster y correspondencias para estudiar los resultados de las pruebas Saber Pro de 2016 en el Departamento de Sucre, identificando asociaciones significativas entre niveles y competencias genéricas.

Larrarte Torres (2022) utilizó modelos de minería de datos, tanto supervisados como no supervisados, para analizar los resultados de las pruebas Saber 11 en Cundinamarca (2017-2021), revelando tendencias consistentes a nivel nacional, departamental y municipal.

Ahumada (2023) clasificó a estudiantes según sus características socioeconómicas utilizando K-Means, y realizó pruebas de homogeneidad para identificar variables influyentes en los resultados de las pruebas Saber 11 en Antioquia.

Finalmente, Giraldo (2024) realizó un Análisis de Componentes Principales (ACP) sobre los datos académicos de un colegio en Medellín (2019-2021), identificando las estrategias pedagógicas y metodológicas que contribuyeron al éxito en las pruebas Saber 11 en cada grado. El enfoque descrito en el proyecto de segmentación y análisis de resultados de estudiantes que realizaron las Pruebas Saber 11 en 2020 presenta similitudes y diferencias con los estudios y metodologías mencionados previamente.

En el proyecto actual, la exploración y preparación de datos (identificación de valores faltantes y outliers) es similar a las metodologías previas, como las usadas por Chue Gallardo et al. (2007) y Ahumada (2023), quienes también lidiaron con datos complejos de estudiantes y aplicaron técnicas de preprocesamiento.

En el proyecto se usa DBSCAN para segmentar a los estudiantes. Esto es comparable con estudios como el de Madrid (2017), que utilizó algoritmos como Máquinas de Soporte Vectorial y K-vecinos más próximos para clasificar a los estudiantes según sus probabilidades de deserción. Ahumada (2023) también empleó K-means para clasificar estudiantes con base en sus características socioeconómicas.

En resumen, aunque hay similitudes en la utilización de técnicas estadísticas avanzadas, el enfoque del proyecto descrito destaca por su uso de tecnologías y métodos más recientes y sofisticados para el procesamiento, segmentación y análisis de los datos educativos, mientras que los estudios previos mencionados se basan más en técnicas tradicionales de análisis multivariado.

3. Materiales y Métodos

3.1. Base de Datos

Los datos utilizados para este estudio provienen de los resultados de las pruebas Saber 11 realizadas en Colombia en el año 2020, los cuales se obtuvieron a través del portal de datos abiertos del gobierno colombiano (https://datos.gov.co/). Este dataset incluye registros de estudiantes de los calendarios A y B, cubriendo tanto colegios públicos como privados. El conjunto de datos tiene más de 500000 registros con variables que describen no solo los resultados de las pruebas, sino también una amplia gama de características sociodemográficas y académicas de cada persona.

Entre las variables más relevantes se encuentran la nacionalidad, el género, la edad, la dedicación a actividades como la lectura o el uso de internet, la estructura del hogar, el nivel educativo de los padres, el acceso a recursos tecnológicos como internet o consolas de videojuegos, y el tipo de institución educativa.

Para hacer el análisis, se seleccionaron una serie de variables de interés que fueron limpiadas y transformadas para su uso adecuado en el análisis.

De manera general se presenta un resumen de las características y variables de la base de datos:

Estadísticas del Conjunto de Datos

Numero de observaciones	520307
Datos Faltantes	608930
Datos Faltantes (%)	1.4%
Registros Duplicados	0
Porcentaje Registros Duplicados (%)	0.0%

Tipos de Variables

Categóricas	**52**
Texto	7
Numérica	22

Luego de analizar las variables en términos de correlación y relevancia, se eligieron las siguientes para el desarrollo del proyecto:

Variable	Valores Faltantes	Porcentaje Valores Faltantes	Tipo Variables
ESTU_NACIONALIDAD	0	0	Categórica
ESTU_GENERO	9	0	Categórica
ESTU_FECHANACIMIENTO	0	0	Fecha
PERIODO	0	0	Numérica
ESTU_TIENEETNIA	3757	0.72	Categórica
ESTU_COD_RESIDE_DEPTO	766	0.15	Numérica
FAMI_ESTRATOVIVIENDA	18510	3.56	Categórica
FAMI_PERSONASHOGAR	16580	3.19	Categórica
FAMI_EDUCACIONPADRE	14539	2.79	Categórica
FAMI_EDUCACIONMADRE	15030	2.89	Categórica
FAMI_TIENEINTERNET	15194	2.92	Categórica
FAMI_TIENECONSOLAVIDEOJUEGOS	18929	3.64	Categórica
FAMI_NUMLIBROS	15606	3	Categórica
ESTU_DEDICACIONLECTURADIARIA	16013	3.08	Categórica
ESTU_DEDICACIONINTERNET	16179	3.11	Categórica
ESTU_HORASSEMANATRABAJA	18404	3.54	Categórica
COLE_CODIGO_ICFES	0	0	Numérica
COLE_GENERO	0	0	Categórica
COLE_NATURALEZA	0	0	Categórica
COLE_CALENDARIO	0	0	Categórica
COLE_CARACTER	15105	2.9	Categórica
COLE_AREA_UBICACION	0	0	Categórica
COLE_JORNADA	0	0	Categórica
COLE_COD_DEPTO_UBICACION	0	0	Numérica
COLE_DEPTO_UBICACION	0	0	Text
PUNT_GLOBAL	0	0	Numérica

Las estadísticas descriptivas de las variables se pueden consultar en el archivo estadisticasDescriptivasSaber2020.html ubicado en el repositorio del proyecto.

3.2. Limpieza de datos

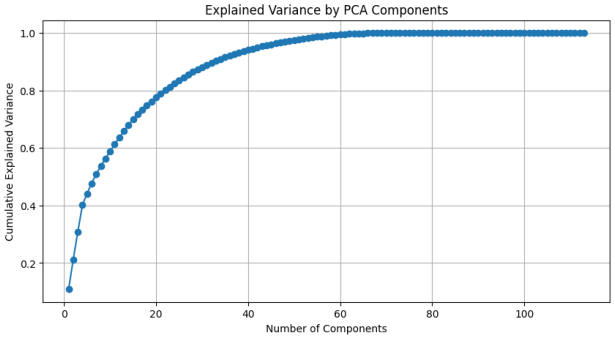
El primer paso que se realizó fue la limpieza de los datos. El conjunto de datos contenía valores faltantes en algunas variables por lo que se realizó una imputación de datos para solucionar este problema, se realizó de la siguiente forma para los dos tipos de variables:

- Variables categóricas: Para las variables categóricas se utilizó la moda para la imputación.
- Variables numéricas: En el caso de las variables numéricas se utilizó la media para la imputación.

Posteriormente, fue necesario transformar las variables categóricas en variables numéricas. Esto se llevó a cabo mediante el uso de One-Hot Encoding, que convierte las categorías en variables binarias, y permite que cada categoría se represente como una columna separada en el conjunto de datos.

Además de esto, dado que las variables numéricas presentaban diferentes escalas fue necesario aplicar una estandarización con el método de StandardScaler. Esto garantiza que en el PCA que todas las variables contribuyan de forma equitativa al análisis.

Después de hacerlo, ya se pudo seguir con el PCA para reducir la dimensionalidad del conjunto de datos manteniendo la mayor varianza posible. El PCA identificó 33 componentes principales que explicaban el 90% de la varianza total. Esto sirvió para reducir la complejidad del conjunto de datos y no perder una cantidad significativa de información.

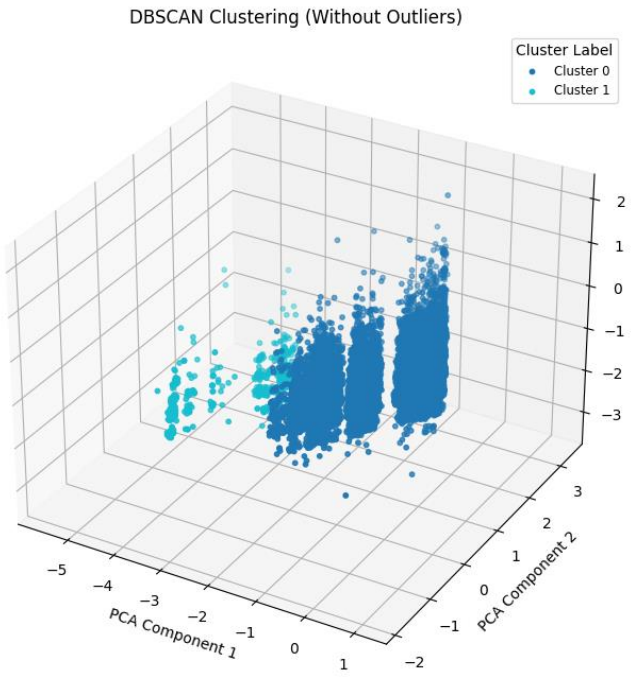


El número óptimo de componentes principales se determinó evaluando la varianza. Se hizo un gráfico del codo (que se encuentra a la izquierda), el cual, mostró que 33 componentes eran suficientes para capturar la mayor parte de la información importante de los datos. Se procedió a realizar la segmentación utilizando DBSCAN y se determinó que el valor óptimo de epsilon era 1.92 (con el método del NearestNeighbors), lo que permitió identificar clusteres de estudiantes con características similares, además de detectar outliers. Adicionalmente, el parámetro min_samples se fijó en 66, debido a que se trata del doble del número de las variables obtenidas en PCA.

Al realizar la ejecución de los algoritmos de NearestNeighbors y DBSCAN, dada la cantidad de datos, encontramos serios problemas de rendimiento que requerían mayor capacidad en Maquina (aun cuando el proceso fue ejecutado recursos con alto desempeño como es Collab de Google), por este motivo decidimos hacer un muestreo aleatorio de los datos en el cual obtuvimos una muestra correspondiente al 10% de los datos y con esta muestra se realizó la segmentación y análisis de NearestNeighbors para obtener un epsilon adecuado. Al finalizar se asignó cada label a la base de datos original, los labels solo representa el 10% de la base de datos por lo que los registros atípicos como los registros que no eran parte de la muestra se marcaron en el campo de Cluster con –1, nuestro análisis demográfico y de resultados basado en los segmentos obtenidos solo toma en cuenta los datos del muestreo.

4. Resultados y Discusión

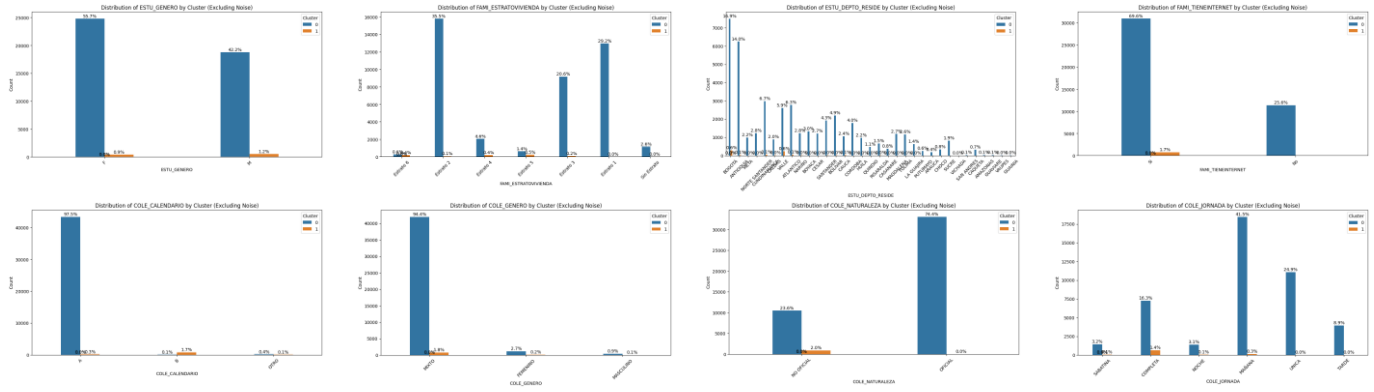
El análisis de segmentación mediante el algoritmo DBSCAN permitió identificar dos clústeres principales, (como se ve en el gráfico de la derecha), entre los estudiantes que participaron en las pruebas Saber 11 de 2020, junto con una gran cantidad de puntos de ruido (estudiantes que no se agruparon de forma clara en ningún cluster). En total, se asignaron 43,560 estudiantes al cluster 0, 914 al cluster 1 (el resto de los 53000 datos encontrados fue ruido). Esto se debe a que este algoritmo utiliza muchos recursos computacionales por lo que se decidió hacer que utilizara un porcentaje menor (53000 datos) aleatorio de los datos para reducir su costo. La alta proporción de puntos clasificados como ruido puede deberse a la alta heterogeneidad de los estudiantes y a la gran cantidad de datos. DBSCAN es un algoritmo que identifica grupos muy conectados y descarta los que están demasiado dispersos para formar un cluster. En este caso, los resultados sugieren que gran parte de la población estudiantil no forma grupos homogéneos definidos de forma clara por la amplia variabilidad en las características socioeconómicas y académicas que se pueden encontrar.



4.1. Distribución de variables por cluster

Al analizar la distribución de varias variables categóricas por cluster, se observaron diferencias notables entre los dos grupos principales de estudiantes (clusteres 0 y 1). Variables como el género, el estrato socioeconómico y el acceso a internet mostraron una distribución muy desigual entre los grupos.

Se generaron los siguientes gráficos de barras:



Estos gráficos muestran que el cluster 1 tenía una mayor proporción de estudiantes con acceso a internet, mientras que el cluster 0 incluía una mezcla más heterogénea de estudiantes sin internet. Esto sugiere que el acceso a internet es un factor clave en la diferenciación de los grupos por lo que repercute de gran manera en el rendimiento (Puntaje global). Los estudiantes del cluster 1 (que ademas son muchos menos) tendían a provenir de estratos socioeconómicos más altos en contraste con el cluster 0 que tenía una mayor proporción de estudiantes de estratos bajos y medios. Estos resultados reafirman que el acceso a recursos es un factor determinante en el rendimiento de cada estudiante en la prueba.

El análisis de la variable edad mostró una distribución desigual entre los clusteres. Los estudiantes del Clúster 1 tendían a ser mayores que los del cluster 0. Esto es relevante ya que podría indicar que los estudiantes que han repetido algún año o que han ingresado tardíamente al sistema educativo podrían estar concentrados en uno de los grupos. Las diferencias de edad también podrían estar relacionadas con otros factores socioeconómicos o con el tipo de colegio al que van. En cuanto al rendimiento académico medido por el puntaje global, se observó que el cluster 1 (la minoría de estudiantes) tendía a obtener puntajes más altos en comparación con el cluster 0. Los resultados mostraron que el cluster 1 no solo tenía una media más alta de puntajes, sino también una menor variabilidad en comparación con el cluster 0.

4.2. Discusión

Los resultados obtenidos con el análisis de DBSCAN confirman diferencias en el rendimiento académico de los estudiantes, correlacionadas con características sociodemográficas como el estrato socioeconómico, el acceso a recursos y el nivel educativo de los padres. Estas diferencias son consistentes con investigaciones previas que han mostrado la gran influencia del contexto socioeconómico en el rendimiento académico de los estudiantes en el país

Una de las limitaciones de este análisis es la gran cantidad de estudiantes clasificados como ruido por el algoritmo DBSCAN. Una posible mejora futura sería probar otros métodos de clustering que no dependan tanto de la densidad de los datos, como el algoritmo de K-means para ver si es posible mejor las relaciones, además de poder correr el algoritmo con todos los datos sin ocupar demasiados recursos computacionales.

5. Conclusiones

Este trabajo proporciona una visión detallada sobre cómo diversos factores sociodemográficos y académicos influyen en el rendimiento de los estudiantes colombianos. Utilizando el PCA y el algoritmo de clustering DBSCAN, se identificaron dos grupos principales de estudiantes, junto con una gran cantidad de puntos considerados como ruido, lo que refleja la heterogeneidad que existe entre los que presentan esta prueba.

Entre los hallazgos más significativos se destacan las diferencias entre los clusters en cuanto a estrato socioeconómico, acceso a internet y el nivel educativo de los padres, variables que resultaron muy importantes para explicar las disparidades en el rendimiento académico.

Los estudiantes del cluster 1, que tendían a tener mejores condiciones socioeconómicas y acceso a mejor educación obtuvieron puntajes considerablemente más altos en las pruebas Saber 11 en comparación con los estudiantes del clúster 0, que muestran mucha más variabilidad. Este análisis expone de nueva cuenta la inequidad en el sistema educativo colombiano. Las diferencias observadas entre los clusters nos hacen ver la necesidad de implementar estrategias para reducir estas brechas tan grandes, esto mejorando el acceso a recursos y oportunidades educativas para estudiantes que no las tienen.

6. Bibliografía

Acosta Solano, J., Lancheros Cuesta, D. J., Umaña Ibáñez, S. F., and Coronado Hernández, J. R. (2022). Predictive models assessment based on crisp-dm methodology for students performance in colombia saber 11° test.

Corporación Universitaria Rafael Núñez, Universidad Cooperativa de Colombia, Departamento de Productividad e Innovación Universidad de la Costa CUC, 198(10):512–517.

Ahumada Riaño, D. (2023). Técnicas de minería de datos para el análisis de pruebas SABER. Universidad Nacional de Colombia.

Cerón Benavides, S. M., Romero Ospina, M. F., and Fonseca Gómez, L. R. (2020). Análisis Estadístico Multivariado De Los Resultados En Las Pruebas Saber Pro Del Programa De Ingeniería En Producción Acuícola En La Universidad De Nariño 2016 – 2019. Fundación Universitaria Los Libertadores.

Chan, D. y Galli, M. G. (2020). Aplicación de técnicas estadísticas multivariadas con el lenguaje de programación R en investigaciones educativas del nivel superior. RAES, 12(20), pp. 123-136.

Chue Gallardo, J., Barreno, E., and Millones Rivalles, R. (2007). Sistema para el análisis estadístico con técnicas multivariadas del rendimiento académico de los estudiantes de una institución de enseñanza superior. Universidad Nacional Agraria La Molina, The University of San Martín de Porres y Universidad de Lima, 2(164).

Giraldo Salguero, I. A. (2024). Análisis del rendimiento académico y de los resultados en Pruebas Saber 11° usando Técnicas Estadísticas Multivariadas (TEM) como un insumo de autoevaluación de un colegio de Medellín (Doctoral dissertation, Universidad Nacional de Colombia).

Ibarra, M. d. C. and Michalus, J. C. (2010). Análisis del rendimiento académico mediante un modelo logit. Revista Ingeniería Industrial. Universidad Nacional de Misiones. Vol. 9. Núm. 2 : 47 – 56.

Larrarte Torres, C. A. (2022). Minería de Datos Educativos (EDM): análisis de los factores determinantes que influyeron el desempeño de las pruebas SABER en Cundinamarca (Colombia) entre 2017 a 2021 (Doctoral dissertation, Maestría en Gestión de Información).

Madrid Echeverry, J. I. (2017). Propuesta De Un Modelo Estadístico Para Caracterizar Y Predecir La Deserción Estudiantil Universitaria. Universidad Nacional de Colombia, Facultad de Minas, Medellin.

OECD (2014). «Education at a Glance 2014: OECD Indicators.,» OECD Publishing.

Pérez-Pulido, M.O., Aguilar-Galvis, F., Orlandoni-Merli, G., y Ramoni-Perazzi, J. (2016). Análisis estadístico de los resultados de las pruebas de estado para el ingreso a la educación superior en Colombia. Revista Científica, 27, 328-339. Doi: 10.14483/udistrital.jour.RC.2016.27.a3}

Rodríguez Manrique, J. A., Ruiz Escorcia, R. R., and Cohen Manrique, C. S. (2022). Análisis multivariado aplicado a la evaluación de competencias saber-pro en el departamento de sucre, colombia. Corporación Universitaria del Caribe (CECAR), 16(16):19–21.

Rodríguez Rodríguez, A. K., & Hernandez Cano, C. Y. (2021). Análisis de las Pruebas Saber 11, años 2017 y 2018, identificando las variables determinantes en los bajos resultados y en la brecha existente entre los estudiantes de colegios categoría A y A+ versus los D en el sector público. Retrieved from https://ciencia.lasalle.edu.co/ing_industrial/170

Ruiz Escorcia, R. R., Arevalo Medrano, J. B., Morillo, G. P., and Acosta-Human˜ez, P. B. (2016). Principal component analysis applied to the state colombian test icfes saber 11°. Universidad Nacional de las lomas, Corporación Universitaria Minuto de Dios, Universidad Simón Bolívar, 39(10):3.