Factores que predicen la culminación de la formación a nivel de pregrado a partir de la prueba Saber 11 y el cuestionario socioeconómico.

María Fernanda Bonilla Naranjo
José John Fredy González Veloza
Fundación Universitaria los Libertadores

Resumen

El presente estudio pretende identificar los factores socioeconómicos que predicen los factores asociados a la brecha educativa en Colombia. Se utilizaron los resultados del cuestionario socioeconómico y las pruebas Saber 11 reportados por El Instituto para la Evaluación de la Educación (Icfes) de los periodos 2012-1 al 2016-2, y como variable respuesta el haber presentado el examen Saber Pro (bajo la hipótesis de que el presentarlo es indicador de estar culminando sus estudios universitarios y el no presentarlo por el contrario asume el hecho no haber accedido o haber desertado de este proceso formativo). Se utilizaron técnicas de aprendizaje automático supervisado, implementando el modelo Light Gradient Boosting Machine para el análisis de la información. Como principales resultados, se identificó que las principales variables sociodemográficas relacionadas con el hecho de terminar o estar culminando el proceso formativo a nivel de pregrado son: pertenecer al género femenino, tener acceso a internet en el hogar, estudiar en una institución educativa con jornada completa, y la formación profesional de la madre; de forma complementaria las variables que son mejores predictoras de NO estar en el proceso de finalización de educación superior a nivel de pregrado, están relacionadas con bajos ingresos familiares, encontrarse residiendo en zonas rurales, y no tener acceso a computador en el hogar.

Introducción

La brecha de educación en Colombia ha sido una problemática reportada en diferentes estudios, denunciando que la falta de acceso a una educación de calidad puede limitar las oportunidades de desarrollo personal, económico y perpetuar la desigualdad social; hay variables estudiadas previamente que evidencian aún más esta brecha, entidades como la UNESCO y el IESALC reportan que en Colombia existe hasta un 35% más de posibilidades para acceder a la educación superior para las zonas urbanas que si se reside en zonas rurales, superando la brecha de américa latina (2020), así mismo, otros estudios señalan como variables importantes en este fenómeno el nivel de pobreza por departamentos, el residir en zonas rurales y ser mujer(Rodríguez, 2018); por su parte el Banco Mundial en conjunto con la OCDE reportan como principales variables diferenciadoras del acceso a la educación superior la zona de ubicación de las instituciones (Rural, Urbana), la naturaleza del colegio (Privado o Público) y las condiciones socioeconómicas de las entidades privadas (OECD et al., 2013). Dentro de los estudios en los que se han utilizado modelos estadísticos para identificar las variables predictivas que den información acerca de esta problemática se encuentra: el estudio de Aislamiento Geográfico y Aprendizaje en las Escuelas Rurales, el cual busca factores causales a partir de la distancia geográfica que influyen en la educación, para esto hace uso de modelos de regresión discontinua; como principales variables explicativas relaciona la diferencia de insumos educativos, la formación de los docentes y asignación de cargos permanentes (Bonilla-Mejía & Londoño-Ortega, 2021). Por otro lado, la explicación de la brecha del rendimiento estudiantil rural y urbano de diferentes cuantiles de distribución en las pruebas PISA, señala que las diferencias del rendimiento en instituciones rurales y urbanas son significativas, principalmente dadas por las diferencias en las características

de las instituciones, en este estudio hicieron uso de regresión cuantil (Gomez-Gonzalez et al., 2021). Adicionalmente, otros estudios que tienen en cuenta las pruebas PISA y evalúan diferencias que contribuyen a la brecha educativa se encuentra el realizado por(Ramos et al., 2016). Finalmente, hay estudios que calculan la brecha en la calidad educativa en educación media y educación superior, como resultados principales se reporta que la brecha está asociada con las diferencias entre planteles (Celis Gálvez et al., 2012)**.**
Por lo anterior, el presente estudio busca desarrollar una herramienta por medio del uso de métodos estadísticos que permita identificar los factores socioeconómicos al momento de finalizar la formación secundaria que predicen estar cerca o haber culminado un proceso de formación universitaria, para esto se identificaron las variables socioeconómicas y puntajes obtenidos al momento de realizar la prueba saber 11 que están asociadas con el hecho de haber presentado el examen de Saber Pro. Se tomaron como referencia los resultados de las pruebas saber 11 de los periodos 2012_1 al 2016_2 y por medio del archivo ”Llave_Saber11_Saberpro.txt”se relacionó como dato pareado el código del examen Saber Pro para los estudiantes que lo hubieran presentado. Para el análisis se emplearon varios modelos supervisados con el fin de evaluar cuál de ellos ofrecía el mejor rendimiento, a partir de lo anterior, se determinó que el modelo Light Gradient boosting Machine (LGBM) logra la mejor comprensión de los factores subyacentes.

Métodos (Materiales y Métodos)

Programas y librerías utilizadas utilizados

Para el procesamiento de los datos y la implementación del modelo se utilizaron los lenguajes de programación R y Python, dentro de las principales librerías utilizadas para la implementación y evaluación del modelo, se usó Pycaret, Sweetviz.

Selección Inicial de Variables

Con el objetivo de obtener información de participantes que, además de haber presentado el examen Saber 11, también hubieran podido presentar el examen Saber Pro, se seleccionaron los resultados de las pruebas Saber 11 y el cuestionario socioeconómico correspondiente a los periodos 2012-1 al 2016-2 (en total 10 bases de datos) como variables predictoras. Estos datos fueron obtenidos del repositorio de datos del Icfes, adicionalmente, para contar con la información de datos pareados de las pruebas Saber 11 y Saber Pro se descargó la base “Llave_Saber11_Saberpro.txt” alojada en el apartado de cruces de este mismo repositorio

Para la selección inicial de las variables, se tuvieron en cuenta aquellas que fueran transversales en todos los cuestionarios descargados y como criterio de exclusión se retiraron las variables que tuvieran campos de respuesta abierta. En este apartado es importante señalar que a partir del año 2014 cambió la forma de calificación y los componentes que comprenden el examen Saber 11, sin embargo, las bases de datos de los años 2012 y 2013 cuentan con variables de recalificación que contienen la información de estos nuevos componentes (Icfes, 2014), estás últimas variables fueron las que se incluyeron para el posterior análisis. Teniendo en cuenta lo anterior, y a partir de la revisión de las bases se seleccionaron 40 variables y se unificó una sola base de datos.

En cuanto a la base que contiene la variable llave (códigos pareados del examen Saber 11 y Saber Pro), se realizaron algunas exploraciones en las que se halló la frecuencia del número de casos diferentes del examen Saber 11 que se presentaron para un mismo código de Saber Pro. Se presentaron 21.618 casos en los que el examen se presentó 2 veces, 2.052 ocasiones en las que el participante presentó el examen en 3 ocasiones y como dato atípico se presentó 1 caso en el que se encontraron 9 registros de examen Saber 11 para un mismo participante. Cabe señalar que esta base datos contiene un rango de tiempo mayor al seleccionado para el presente estudio, por lo que únicamente se llevo a la base definitiva los casos que estuvieran en el rango estipulado. Por último, en esta base se hizo la revisión de casos únicos con el código Saber Pro dejando para los casos con más de una aplicación en el Saber 11 la última aplicación realizada

Posteriormente, se unió a la base principal el código del examen Saber Pro, de esta manera los

casos que permanecían vacíos corresponden a los casos en los que no se ha presentado el examen Saber Pro, esta variable adquiere el nombre de “PRESENTO_SABER_PRO” y corresponde a la variable objetivo o a predecir del presente estudio. La base final contó con un total de 2’185.483 casos. A continuación, se describen las 40 variables seleccionadas en la primera fase.

Tabla 1. Variables iniciales del presente estudio

Variable	Descripción	Escala
PRESENTO_SABER_PRO	Reporta si el estudiante presentó el examen Saber Pro	Categórica
ESTU_CONSECUTIVO	Código del examen Saber 11	Categórica
ESTU_NACIONALIDAD	Nacionalidad del estudiante	Categórica
ESTU_GENERO	Sexo del estudiante (F o M)	Categórica
PERIODO	Periodo en el que presentó el examen Saber 11	Categórica
ESTU_ESTUDIANTE	Estudiante	Categórica
ESTU_PAIS_RESIDE	País de Residencia del estudiante	Categórica
ESTU_DEPTO_RESIDE	Departamento de Residencia del estudiante	Categórica
ESTU_COD_RESIDE_DEPTO	Código de Departamento de Residencia del estudiante	Categórica
ESTU_MCPIO_RESIDE	Municipio de Residencia del estudiante	Categórica
ESTU_COD_RESIDE_MCPIO	Código de Departamento de Residencia del estudiante	Categórica
ESTU_AREARESIDE	Área en la que reside el participante (Rural o Urbana)	Categórica
ESTU_VALORPENSIONCOLEGIO	Valor mensual de la pensión al momento de presentar el examen Saber 11	Ordinal
FAMI_EDUCACIONPADRE	Nivel educativo más alto alcanzado por el padre	Categórica
FAMI_EDUCACIONMADRE	Nivel educativo más alto alcanzado por la madre	Categórica
FAMI_OCUPACIONPADRE	Ocupación u oficio del padre	Categórica
FAMI_OCUPACIONMADRE	Ocupación u oficio de la madre	Categórica
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de su vivienda según recibo de energía eléctrica	Ordinal
FAMI_PERSONASHOGAR	Número de personas conforman el hogar donde vive el participante	Categórica
FAMI_CUARTOSHOGAR	Total de cuartos en los que duermen las personas del hogar	Categórica
FAMI_TIENEINTERNET	El hogar cuenta con servicio de internet	Categórica
FAMI_TIENECOMPUTADOR	El hogar cuenta con computador	Categórica
FAMI_INGRESOFMILIARMENSUAL	Total de ingreso mensual en SMLV	Ordinal
ESTU_TRABAJAACTUALMENTE	El estudiante trabaja actualmente	Categórica
COLE_GENERO	Indica el género de la población del establecimiento	Categórica
COLE_NATURALEZA	Colegio Oficial o No Oficial	Categórica
COLE_CALENDARIO	Calendario académico del establecimiento	Categórica
COLE_BILINGUE	El colegio es bilingüe	Categórica
COLE_CARACTER	Indica el carácter del establecimiento (Académico, Técnico, Técnico/Académico)	Categórica
COLE_AREA_UBICACION	Área de ubicación de la sede	Categórica
COLE_JORNADA	Jornada de la sede	Categórica
COLE_COD_MCPIO_UBICACION	Código de Municipio de ubicación del Colegio	Categórica
COLE_MCPIO_UBICACION	Municipio de ubicación del Colegio	Categórica
COLE_COD_DEPTO_UBICACION	Código de Departamento de ubicación del Colegio	Categórica

COLE_DEPTO_UBICACION	Departamento de ubicación del Colegio	Categórica
ESTU_PRIVADO_LIBERTAD	Respuesta del evaluado si actualmente se encuentra privado de la libertad	Categórica
PUNT_SOCIALES_CIUDADANAS	Puntaje en Ciencias Sociales	Intervalo
PUNT_INGLES	Puntaje inglés	Intervalo
PUNT_LECTURA_CRITICA	Puntaje Lectura Crítica	Intervalo
PUNT_MATEMATICAS	Puntaje Matemáticas	Intervalo
PUNT_C_NATURALES	Puntaje Ciencias Naturales	Intervalo

Selección de las variables a trabajar en el modelo

Se realizó la exploración inicial de los datos para identificar la distribución de las categorías de cada variable y la relación con la variable objetivo. A partir de la exploración de los datos, se eliminaron algunas variables que no son informativas para el modelo a desarrollar y se recategorizaron otras variables para disminuir el número de categorías a utilizar. A continuación, se realiza a descripción de las variables eliminadas

- **ESTU_CONSECUTIVO**: La variable identifica los casos únicos, pero no es informativa frente al modelo.
- **ESTU_NACIONALIDAD:** Eliminar debido a que cada categoría diferente a la Nacionalidad Colombiana representa menos del 1% de la muestra.
- **ESTU_ESTUDIANTE:** Esta variable sólo presenta una categoría no hay variabilidad por lo tanto no funcionará como variable predictora del modelo.
- **ESTU_PAIS_RESIDE:** Las categorías diferentes a Colombia son inferiores al 1% de los datos del modelo.
- **COLE_CALENDARIO:** Se elimina esta variable porque al revisar los porcentajes de la variable todos los colegios de calendario B son No oficiales, la variable no aporta información que no se pueda revisar con la variable **ESTU_PAIS_RESIDE.**
- **ESTU_PRIVADO_LIBERTAD** Se decide retirar la variable ya que la categoría "No" cuenta con menos del 1% de los datos.
- El siguiente conjunto de variables se eliminaron teniendo en cuenta que se reportan la misma información o tienen alto índice de correlación aumentando la colinealidad del modelo con la variable ESTU_RESIDE_DEPTO. Se decidió mantener la ubicación de residencia de los participantes a nivel departamental, por lo tanto, se eliminaron las variables **ESTU_COD_RESIDE_DEPTO, ESTU_MCPIO_RESIDE, ESTU_COD_RESIDE_MCPIO**. Así mismo, se prioriza la ubicación de residencia del participante, por lo tanto, se eliminaron las variables de ubicación de la institución educativa, **COLE_AREA_UBICACION, COLE_COD_MCPIO_UBICACION, COLE_MCPIO_UBICACION, COLE_COD_DEPTO_UBICACION, COLE_DEPTO_UBICACION.**

Adicionalmente, se agruparon las categorías de algunas variables que se consideran importantes para el modelo de predicción,

- **ESTU_DEPTO_RESIDE** En esta variable se recategorizaron los departamentos que representaban valores inferiores al 2% de la muestra unificándolos en la categoría "otros". Dentro de esta categoría se agrupan los siguientes departamentos (Sucre, Risaralda, La Guajira, Quindío, Casanare, Putumayo, Caquetá, Chocó, Arauca, Guaviare, Amazonas, San Andrés, Vichada, Vaupés, Guainía, y Extranjero).
- Para las variables **FAMI_OCUPACIONPADRE** y **FAMI_OCUPACIONMADRE** se unificaron las categorías (Profesional Independiente y Profesional independiente) ,

(Empleado de nivel directivo y Empleado con cargo como director o gerente general) (Empresario, Pensionado y Pequeño empresario como “otros”).

- En la variable **FAMI_PERSONASHOGAR** se redujo el número de categorías unificando de 8 personas en adelante como "más de 8".
- La variable **FAMI_CUARTOSHOGAR** se disminuyó el número de categorías unificando el número de cuartos por encima de 5 como 'más de 5".
- En la variable **ESTU_TRABAJAACTUALMENTE** se agruparon las variables que describen el tiempo que el estudiante trabaja como "Sí" dejando la variable dicotómica.

Elección del modelo

Para la implementación del modelo se seleccionó una muestra de 500.000 datos de la base depurada, con el 70% como datos de entrenamiento y el 30% de datos de testeo. Como herramienta para identificar principales variables que aportan al modelo se utilizó Feature selection, el cual indicó como principales variables predictoras del modelo:
los puntajes de la prueba Saber 11, categorías de la jornada del colegio, el género del participante, y el departamento, sin embargo, se decidió conservar las 27 variables como predictoras al momento de aplicar el mejor modelo.

Selección del Mejor Modelo

Se realizó la comparación entre diferentes modelos de clasificación con Machine Learning, como evaluadores del modelo se revisaron las siguientes métricas:

Accuracy: Porcentaje total de estudiantes que fueron clasificados de manera correcta.

Área bajo la curva (AUC): Es una medida de la capacidad del modelo para clasificar correctamente los datos, el valor de 1 en esta variable indica que el modelo clasifica de forma perfecta los datos y valores de 0.5 que la clasificación es completamente aleatoria.

Sensibilidad (Recall): El número de estudiantes que presentaron el Saber Pro, clasificados correctamente del total de participantes que presentaron este examen en los datos reales.

Precisión: El número de estudiantes que presentaron el Saber Pro, clasificados correctamente sobre el total de los estudiantes que el modelo predijo presentaron el examen Saber Pro.

F1 Score: Esta medida es la media armónica de la precisión y sensibilidad del modelo, buscando minimizar el posible engaño que generan altos puntajes en las métricas anteriores.

Coeficiente de correlación de Matthews (MCC) = Este coeficiente toma en cuenta tanto los verdaderos positivos (TP), como los verdaderos negativos (TN) del modelo, así como también los falsos positivos (FP), y falsos negativos (FN), y produce un valor que oscila entre -1 y +1, siendo 1 la clasificación perfecta, 0 una clasificación aleatoria y -1 y clasificación contraria, esto lo lleva a cabo, a partir de la siguiente fórmula, tomada de (Boughorbel et al., 2017)

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

Puede ser más informativo que el F1 score ya que considera las relaciones de desbalance existentes entre las cuatro categorías que conforman la matriz de confusión (verdaderos positivos, verdades ton negativos, falsos positivos y falsos negativos(Chicco & Jurman, 2020).

Tabla 2. Evaluadores de los distintos modelos aplicados

Model	Accuracy	AUC	Recall	Prec.	F1	MCC
Light Gradient Boosting Machine	0.82	0.83	0.37	0.62	0.46	0.38
Logistic Regression	0.82	0.83	0.33	0.63	0.44	0.37
Linear Discriminant Analysis	0.82	0.83	0.37	0.61	0.46	0.38
Gradient Boosting Classifier	0.82	0.83	0.33	0.62	0.43	0.36
Random Forest Classifier	0.82	0.82	0.33	0.62	0.43	0.36
Ada Boost Classifier	0.82	0.82	0.34	0.61	0.43	0.36
Naive Bayes	0.77	0.78	0.58	0.45	0.51	0.36
Quadratic Discriminant Analysis	0.74	0.76	0.58	0.43	0.48	0.34
K Neighbors Classifier	0.79	0.72	0.31	0.49	0.38	0.27
Decision Tree Classifier	0.75	0.62	0.41	0.39	0.40	0.24
Dummy Classifier	0.80	0.50	0.00	0.00	0.00	0.00
SVM - Linear Kernel	0.81	0.00	0.22	0.64	0.29	0.27
Ridge Classifier	0.82	0.00	0.29	0.64	0.40	0.34

Modelo Light Gradient Boostig Machine: Este modelo se basa en la construcción de múltiples pequeños árboles de decisión en secuencia, de forma que cada uno va corrigiendo los errores del anterior para mejorar la predicción, tiene ventajas sobre los otros métodos ya que emplea técnicas de optimización para poder manejar un gran volumen de datos, ya que selecciona durante el entrenamiento solo los ejemplos con un mayor gradiente (los más importantes), y agrupa características que usualmente se usan reduciendo el número de características que el modelo tiene que procesar (Ke et al., 2017).

Técnica de regularización utilizada.

A partir de las iteraciones del modelo , se analizó sobreajuste, con el fin de evaluar el desempeño de los modelos, tanto para predecir los datos de entrenamiento, como para generalizar estos resultados con otro conjunto de datos. Para esto, se utilizó como técnica de regularización del modelo la máxima profundidad de los árboles de decisión, la cual se ajustó a un valor de 3.

Resultados.

Los resultados presentados se dividen en dos partes, la primera responde al análisis descriptivo de las variables más importantes realizado con todos lo casos de la base de datos y el segundo a los resultados del modelo aplicado para la muestra seleccionada.

Análisis Descriptivo de las variables seleccionadas.

Se observó que el porcentaje de participantes que presentaron la prueba Saber Pro por departamento fue mayor en Bogotá (29%), Norte de Santander (24%), Atlántico y Boyacá (22% cada uno), y Santander (20%), en contraste con los departamentos de Antioquia (15%), Magdalena (14%), y Cauca (13 %).

Tabla 3. Distribución porcentual de participantes que presentaron el examen Saber Pro por Departamento

Departamento	N	Porcentaje de casos	N de personas que presentaron el Saber Pro	Porcentaje de Personas que presentaron el Saber Pro por Departamento
Bogotá	376,120	17%	108,811	29%
Antioquia	290,539	13%	44,843	15%
Otro	229,877	11%	41,873	18%
Valle del Cauca	197,079	9%	38,834	20%
Cundinamarca	139,448	6%	29,597	21%
Atlántico	114,671	5%	25,523	22%
Santander	103,372	5%	21,395	21%
Bolívar	94,024	4%	15,899	17%
Córdoba	71,240	3%	11,857	17%
Nariño	67,232	3%	13,114	20%
Boyacá	66,258	3%	14,618	22%
Tolima	65,812	3%	13,291	20%
Norte de Santander	58,986	3%	14,367	24%
Magdalena	58,599	3%	8,177	14%
Cauca	57,115	3%	7,503	13%
Huila	52,240	2%	10,378	20%
Cesar	47,643	2%	8,864	19%
Vacíos	92,336	4%	18,399	19%

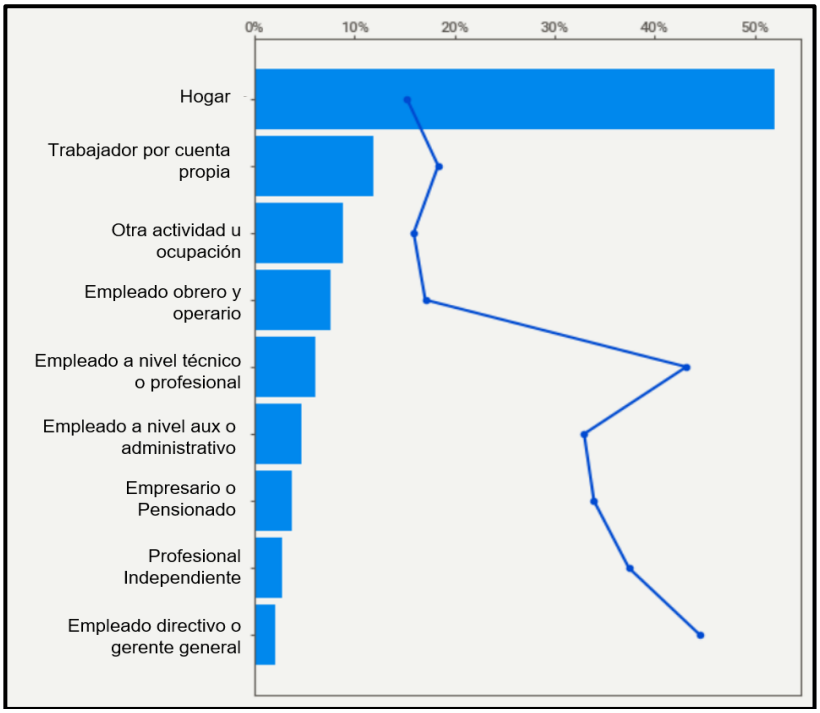
En zonas urbanas el porcentaje es del (23%), en comparación de las zonas rurales en las que únicamente es del (10%).

Con relación al valor de la pensión del colegio, en las instituciones en las que pagan pensiones con valores inferiores a los $87.000 el porcentaje de estudiantes que posteriormente presentan el examen Saber Pro es menor (10%) que en los colegios en los que no se paga pensión (16%), sin embargo, estos a su vez tienen porcentaje más bajo en comparación con lo colegios que pagan más de $87.000 (28% - 59%).

Con respecto a la formación educativa de los padres, en los casos en los que cualquiera de los padres completó el nivel de pregrado en su formación académica, el porcentaje de estudiantes que presentaron el saber Pro es del (48% al 49%) y este porcentaje puede aumentar hasta 61% o 63% si alguno de los padres culminó una formación a nivel de postgrado. Por el contrario, disminuyen a medida que los padres han tenido menos recorrido en el proceso académico, de esta manera en los hogares en los que alguno de sus padres no hubiera tenido ningún tipo de estudio, el porcentaje de estudiantes que culmina o está en el proceso de finalización del pregrado es tan solo del 5 a 8%.

El 52% de las madres se enfoca en realizar exclusivamente el trabajo doméstico en comparación con el 1% de los padres. Sin embargo, solo el 15% de los participantes que tienen madres dedicadas exclusivamente al hogar ha terminado o están en proceso de culminar la educación superior. No obstante, cuando la madre tiene empleo como técnico, profesional, directiva o gerente, este porcentaje aumenta considerablemente a un rango del 43% al 45%.

Gráfico 1. Distribución porcentual de los participantes que presentaron el examen Saber Pro según la ocupación de la madre.



Nota: Las barras representan el porcentaje de casos en los que la madre presentó esa ocupación y la línea grafica el porcentaje de participantes que presentaron el Saber Pro dentro de cada categoría.

En cuanto al estrato, se observa una relación directa entre estas variables, donde a medida que el estrato aumenta, también lo hace el porcentaje de estudiantes que presentan el examen Saber Pro. Este porcentaje es del 10% para los estudiantes de estrato 1, del 33% para los estudiantes de Estrato 3 y del 63% en los estudiantes de Estrato 6. Sin embargo, es importante destacar que solo el Estrato 1 representa el 42% de los casos totales.

Con referencia al acceso a internet, del porcentaje de personas que cuentan con este servicio en sus hogares el 29% presentó el examen Saber Pro, a diferencia de los hogares que no cuentan con este servicio, en los que solo el 11% logran culminar el proceso de educación superior a nivel pregrado. De igual forma ocurre con el acceso a computador en los hogares, en los que cuentan con computador el 28% culmina, a diferencia de los hogares que no poseen computador, en los que únicamente el 9% presenta el examen Saber Pro.

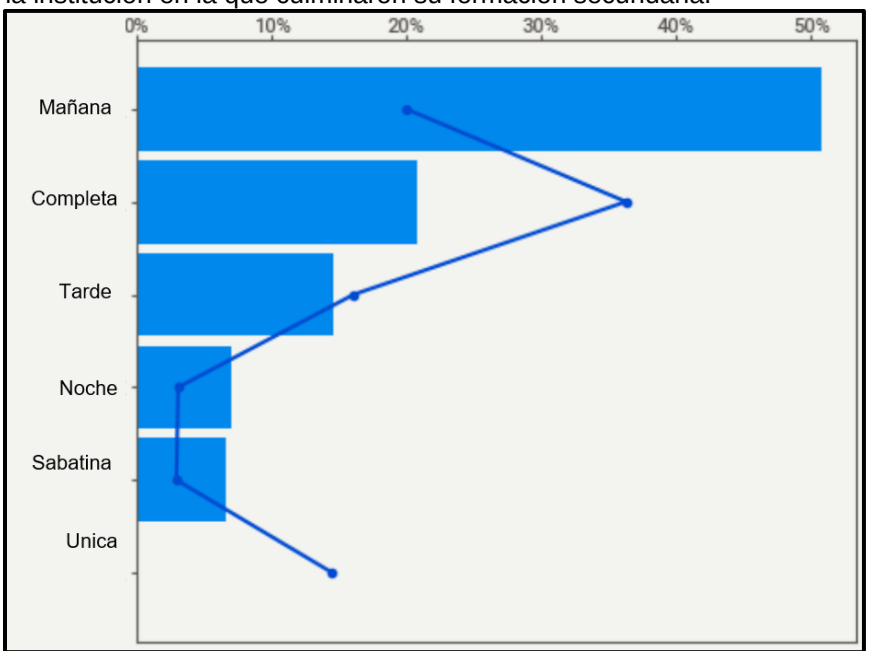
Referente a los ingresos en el hogar, a medida en que aumenta el ingreso se evidencia un mayor porcentaje de personas culminando la universidad, de esta forma, de los hogares que cuentan con ingresos inferiores a 1 salario mínimo legal vigente (SMLV) solo el 9% consiguen llegar al examen Saber Pro, este porcentaje aumenta al 42% en los hogares con ingresos entre 3 y 5 SMLV, y llega hasta a un 64% en los hogares con ingresos superiores a 10 SMLV.

El 10% de los estudiantes se encontraban laborando al momento de presentar el examen Saber 11, de estos casos únicamente el 8% presentó la prueba Saber Pro, a diferencia de los estudiantes que no se encontraban laborando, en este último grupo el 22% culminaron o están culminando el pregrado.

Por otro lado, a pesar de que los estudiantes de colegios oficiales representan el 71% del total de los datos, de este grupo únicamente el 16% presentó el examen Saber Pro, a diferencia de los estudiantes de colegios no oficiales, en el que este porcentaje sube al 32%.

En cuanto a la jornada de la institución en la que el participante estudió, el mayor porcentaje de personas que presentaron el examen Saber Pro (36%) estudiaron en instituciones con jornada completa, este porcentaje disminuye drásticamente en las instituciones con jornadas nocturnas o sabatinas, en las que para cada caso únicamente el 3% de los participantes para cada grupo culminaron o se encuentran en proceso de finalización del pregrado.

Gráfico 2. Distribución porcentual de los participantes que presentaron el examen Saber Pro según la jornada de la institución en la que culminaron su formación secundaria.



Nota: Las barras representan el porcentaje de casos de estudiantes según la jornada de la institución en la que culminaron la educación secundaria, y la línea señala el porcentaje de participantes que presentaron el Saber Pro dentro de cada categoría.

Los puntajes de la prueba en cada una de las competencias presentaron el mismo comportamiento asociado a que entre mayores puntajes se obtienen mayor es la posibilidad de culminar los estudios de educación superior a nivel de pregrado

Resultados frente al modelo

A partir de la regularización del modelo, los evaluadores reportan los siguientes valores

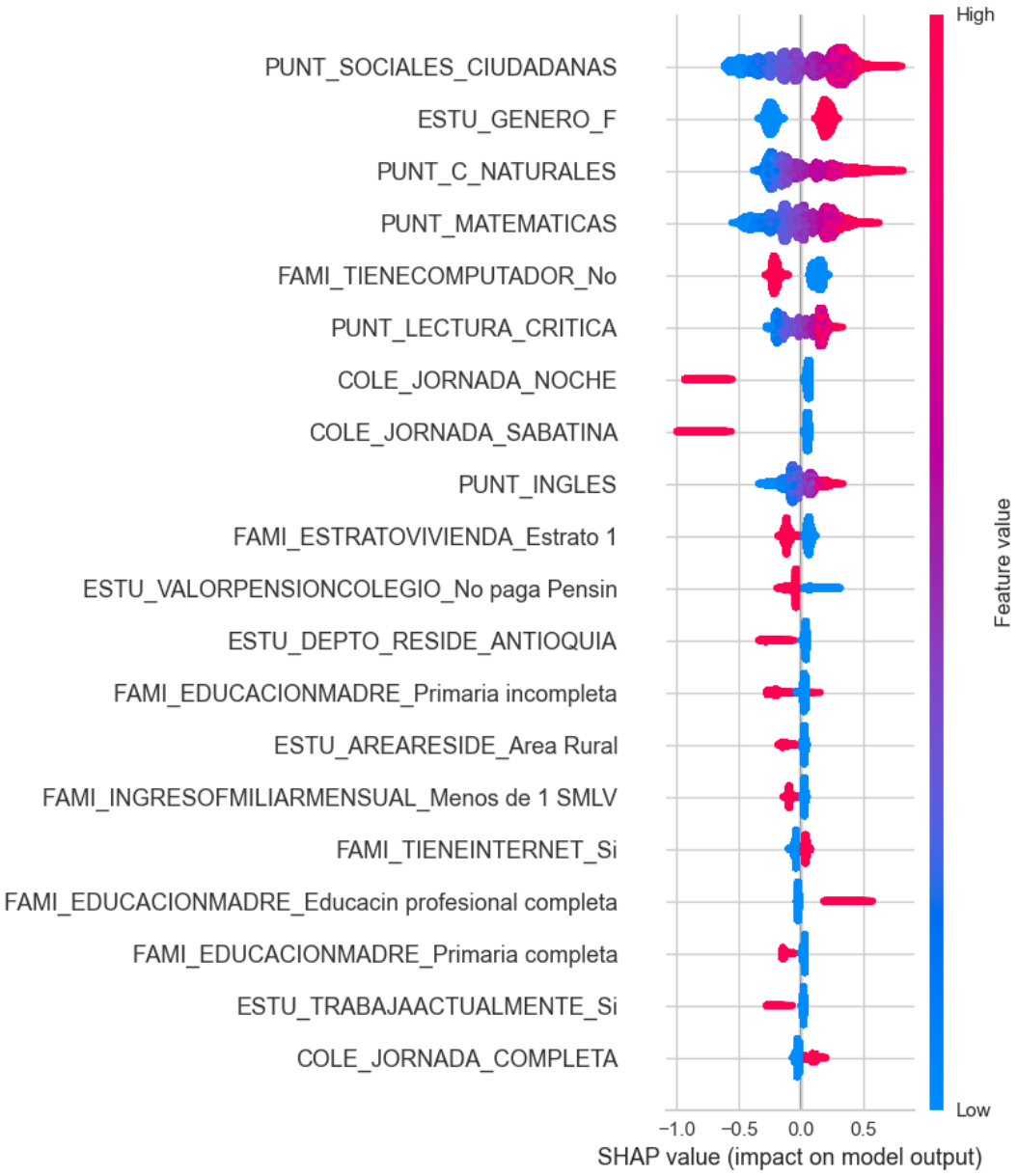
Tabla 4. Evaluadores del mejor modelo

Model	Accuracy	AUC	Recall	Prec.	F1	MCC
Light Gradient Boosting Machine	0.82	0.83	0.37	0.62	0.46	0.38

El modelo de clasificación analizado ha demostrado tener un rendimiento aceptable, a pesar de que presenta un buen comportamiento en la tarea de clasificación reflejado en el AUC, presenta problemas de sensibilidad. En particular, se identifica que solo clasificó correctamente el 37% de los casos positivos (Recall). Por lo tanto, es importante tener en cuenta esta limitación al interpretar los resultados del modelo. Para esto se sugiere en estudios posteriores ajustar el umbral de decisión y/o explorar el ajuste de otros hiperparámetros distintos a la profundidad de los árboles de decisión.

A continuación, se relacionan las principales variables que predicen que un participante culmine o se encuentre en proceso de finalización de la educación superior de acuerdo con el modelo aplicado.

Gráfico 3. Principales variables predictoras de que el participante culmine el proceso de formación superior a nivel de pregrado (Saber Pro)



Discusión de Resultados

A partir del grafico 3 se observa que los puntajes de la prueba son principales predictores de que el estudiante culmine o se encuentre en el proceso de finalización de la educación superior, de esta manera el tener puntajes altos en la prueba Saber 11 aumenta la probabilidad de culminar la educación superior, sin embargo, es de resaltar que los componentes que mejor predicen son Sociales y Ciudadanas, Ciencias Naturales y Matemáticas.

En coherencia con lo registrado en los resultados descriptivos, el estudiar en jornada nocturna o sabatina, es predictor de que el estudiante no culmine un proceso de educación superior a nivel de

pregrado, como principal hipótesis de este fenómeno puede deberse a que la cantidad horaria dispuesta para la enseñanza en estas modalidades es menor que la de otras jornadas, esto se complementa con la última variable predictiva presentada en el gráfico que señala como predictor de presentar el examen Saber Pro el hecho de haber estudiado en un colegio en jornada completa.

Por otro lado, la educación de la madre es primordial como predictor, es así como en los casos en los que ella haya llegado a una formación primaria completa o incompleta como máximo nivel educativo en menos probable que el estudiante culmine o se encuentre a puertas de terminar su formación de pregrado, por el contrario, en los casos en el que la formación de la madre es de pregrado completo hay mayor posibilidad de que el participante culmine la educación superior.

Adicionalmente, dentro los factores relacionados en otros estudios, se destaca que el ser estudiante de zona rural o estar en una familia con ingresos inferiores a 1 SLMV son predictores de no culminar los estudios a nivel de pregrado. De igual forma, estar en un colegio en el que no se paga pensión (colegio oficial) es predictor de no presentar el examen Saber Pro. Estos resultados contribuyen a los estudios que sustentan que la brecha en la educación en su mayoría está soportada por desigualdades a nivel económico.

Como hallazgo se presenta que el pertenecer al género femenino es un predictor de culminar o encontrarse culminando el proceso formativo de educación superior, a pesar de que no hay una explicación aparente de este fenómeno al momento de culminar la educación secundaria, algunos estudios señalan que la deserción en la formación de educación superior es mayor en hombres que en mujeres(Katzkowicz & Arim, 2017; Isaza et al., 2016) y que el ingreso a la formación de educación superior ha sido mayor en mujeres que hombres a partir del 2022 (UNESCO, 2021)

Por lo que se refiere a resultados a nivel de Departamento, el estudiar en Antioquia disminuye la posibilidad de culminar o encontrarse finalizando la formación académica a nivel de pregrado. Por otro lado, la situación laboral de los estudiantes también es un factor importante, que refleja que aquellos que laboran tienen una menor probabilidad de presentar el examen Saber Pro que aquellos que no trabajan.

Los resultados demuestran que la disparidad en la educación superior se relaciona con factores socioeconómicos, esto se complementa de manera adecuada con lo obtenido en el análisis descriptivo, a pesar de que los resultados del modelo estuvieron principalmente asociados a los puntajes de la prueba, hay que mencionar que estos resultados están directamente asociados con variables como el estrato de la familia, nivel de formación del padre o la madre, y los ingresos familiares, en estos últimos se encuentran correlaciones ratio entre (0.35 y 0.52) dependiendo el componente, siendo mayor en el Puntaje de Inglés y menor en el Puntaje de Sociales y Ciudadanas.

El tener ingresos familiares bajos, dificultad para adquirir computador, residir en zona rural y encontrarse trabajando durante la formación secundaria son predictores de no terminar la educación superior, esto refleja que la problemática de la brecha educativa sigue estando en un problema de acceso económico o territorial, es por esto que las políticas públicas o proyectos de ley que busquen abordar esta problemática deben estar enfocadas en lograr disminuir las causas subyacentes de estos factores, como la promoción de opciones para acceder a la educación superior, mejoras en la educación pública y la resolución de los problemas en las instituciones educativas. Además , se deben fortalecer los programas universitarios con opciones de alojamiento y alimentación para ayudar a los estudiantes a mantenerse en este proceso formativo durante más tiempo. Se deben implementar programas de financiación que faciliten el acceso a la educación superior sin crear una brecha económica después de graduarse debido a las altas deudas.

Adicionalmente sería interesante realizar estudios acerca de las expectativas de futuro y condiciones socioeconómicas de los estudiantes con madres dedicadas exclusivamente al trabajo en el hogar, con el fin de identificar posibles explicaciones al comportamiento de los datos actuales, ya sea que este se dé porque los estudiantes tengan que asumir responsabilidades

económicas a temprana edad o porque dentro de sus expectativas a futuro no sea prioridad el continuar la formación académica a un nivel de pregrado.

Limitaciones

Para el presente estudio no se realizaron pruebas de especificidad del modelo, adicionalmente el modelo no contempla variables durante el proceso de formación superior, como la deserción estudiantil u otras que influyan en la culminación del pregrado, sería importante en estudios posteriores poder incluirlas.

Conclusiones.

- A nivel descriptivo, el estrato socioeconómico y el nivel de ingresos del hogar están relacionados con la tasa de finalización del pregrado, ya que a medida que aumenta el estrato y los ingresos del hogar, también lo hace a tasa de finalización del pregrado, esto a la vez es consistente con lo reportado por el modelo con la variable predictora.

- Dentro de las principales variables predictoras del modelo desarrollado se encuentran los puntajes de la prueba Saber 11, la jornada de la institución en la que culminó la educación secundaria, y la formación académica de la madre

- La disponibilidad de recursos tecnológicos en el hogar, como el acceso a internet y computador son predictores de la continuidad en la formación de educación superior. Los estudiantes que tienen acceso a estos recursos tienen una tasa de finalización del pregrado más alta que aquellos que no los tienen.

- El nivel de ingresos familiares y la zona en la que se reside el estudiante siguen siendo factores relacionados con la brecha en la educación superior

- Como resultados inesperados se presenta que el hecho de ser mujer predice el hecho de presentar el Saber pro, por el contrario, el ser residente de Antioquia disminuye la probabilidad de culminar o encontrarse finalizando el pregrado.

- Con respecto al modelo desarrollado es necesario hacer ajustes en el umbral de decisión o/y realizar nuevas pruebas a partir de los diferentes hiperparametros con el fin de mejorar la sensibilidad del modelo.

Referencias

Bonilla-Mejía, L., & Londoño-Ortega, E. (2021). Geographic Isolation and Learning in Rural Schools. *Borradores de Economía; No.1169*. https://doi.org/10.32468/be.1169

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, *12*(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

Celis Gálvez, M. T., Jaramillo Salazar, J. F., & Jimenez, O. (2012). *¿Cuál es la brecha de la calidad educativa en Colombia en la educación media y en la superior?* (pp. 67-98).

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Gomez-Gonzalez, J. E., RodrÃ-guez GÃ³mez, W., & RodrÃ-guez GÃ³mez, E. (2021). Explaining the Rural-Urban Student Performance Gap for Different Distribution Quantiles in Colombia. *Working Papers*, Article 74. https://ideas.repec.org//p/rie/riecdt/74.html

Icfes. (2014). *Diario Oficial No. 49.150 de 13 de mayo de 2014*. https://normograma.icfes.gov.co/docs/acuerdo_icfes_0023_2014.htm

Isaza, L. G., Lubert, C. D., & Montoya, D. M. (2016). Caracterización de la deserción estudiantil en la universidad de caldas el período 2009-2013. Análisis a partir del Sistema para la Prevención de la Deserción de la Educación Superior –Spadies. *Latinoamericana de Estudios Educativos*, *12*(1), Article 1.

Katzkowicz, N., & Arim, R. (2017). Trayectoria estudiantil: Determinantes de la deserción y culminación del ciclo educativo de estudiantes universitarios. *InterCambios: Dilemas y Transiciones de la Educación Superior*, *4*(2), 108-127.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

OECD, International Bank for Reconstruction and Development, & The World Bank. (2013). *Evaluaciones de Políticas Nacionales de Educación: La Educación Superior en Colombia*. OECD. https://doi.org/10.1787/9789264180710-es

Ramos, R., Duque, J. C., & Nieto, S. (2016). Decomposing the Rural-Urban Differential in Student Achievement in Colombia using PISA Microdata. *Estudios de Economía Aplicada*, *34*(2), 379-411.

Rodríguez, G. J. (2018). La persistencia de la inequidad y la desigualdad en la educación en Colombia. *PAPELES*, *10*(19), 26-39.

UNESCO. (2021). *Mujeres en la educación superior: ¿la ventaja femenina ha puesto fin a las desigualdades de género?* https://unesdoc.unesco.org/ark:/48223/pf0000377183