Identificación de las variables más influyentes en el rendimiento académico de los estudiantes en las pruebas Saber Pro con modelos de machine learning

Darwin Raul Mercado Diaz

Asesor

Jose Laureano Cruz Cardozo

Universidad Nacional Abierta y A Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingenieria - ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Resumen

Las pruebas saber Pro significan una gran medición del conocimiento de los próximos profesionales en Colombia, dado que se realizan en los últimos semestres de las carreras profesionales de las instituciones educativas de educación superior. Al identificar las variables que afectan el rendimiento académico de los estudiantes en las pruebas Saber Pro, se podrán implementar nuevas herramientas de trabajo y estrategias que mejoren no solo los resultados, sino también la educación superior en Colombia. Por lo tanto, con este proyecto se busca identificar las variables que afectan el rendimiento académico estudiantil en instituciones de educación superior al momento del desarrollo de las pruebas de ICFES Saber Pro con el uso de modelos de Machine Learning. Hoy en día se tienen modelos de Machine Learning qué permiten predecir rendimientos académicos, los más comunes son los algoritmos de Random Forest y Gradient Boosting, los cuales permiten analizar grandes cantidades de datos y poder identificar las variables con más influencia. Luego de la aplicación de ambos algoritmos, se encontró que los factores socioeconómicos tienen una gran incidencia frente a el rendimiento académico de los estudiantes en las pruebas Saber Pro del ICFES y se llegó a recomendaciones desde el punto económico y aplicación de estrategias por parte de las instituciones de educación superior.

Palabras Clave: ICFES, rendimiento, modelos, algoritmo, influencia.

Abstract

The Saber Pro tests are a great measure of the knowledge of the next professionals in Colombia, since they are taken in the last semesters of professional careers in higher education institutions. By identifying the variables that affect the academic performance of students in the Saber Pro tests, it will be possible to implement new work tools and strategies that improve not only the results, but also higher education in Colombia. Therefore, this project seeks to identify the variables that affect student academic performance in higher education institutions at the time of the development of the ICFES Saber Pro tests with the use of Machine Learning models. Nowadays there are Machine Learning models that allow predicting academic performance, the most common are the Random Forest and Gradient Boosting algorithms, which allow analyzing large amounts of data and identifying the most influential variables. After the application of both algorithms, it was found that socioeconomic factors have a great impact on the academic performance of students in the ICFES Saber Pro tests and recommendations were made from the economic point of view and the application of strategies by higher education institutions.

Keywords: ICFES, performance, models, algorithm, influence.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Planteamiento del Problema

Las pruebas saber Pro significan una gran medición del conocimiento de los próximos profesionales en Colombia, dado que se realizan en los últimos semestres de las carreras profesionales de las instituciones educativas de educación superior. Poder identificar las áreas de conocimiento donde se tienen falencias y donde los estudiantes siempre tienden a obtener bajos puntajes en las pruebas Saber Pro. Esta identificación pueden darle a los entes gubernamentales y privados una analítica clara de qué se debe de mejorar en las fases educativas, para poder realizarlo se tiene un gran desafío que es poder comprender las áreas de conocimiento y las variables externas que pueden afectar este rendimiento académico, por lo tanto, estos factores como la calidad de la educación en cada una de las instituciones, el entorno socioeconómico y factores personales siempre van a impactar de gran manera en el rendimiento académico. Por lo tanto, se tiene una necesidad muy grande de utilizar modelos de Machine Learning para poder utilizar los modelos que se tienen para poder predecir las áreas de conocimiento y factores externos que tienen una afectación mayor, con esto poder brindar una estructura completa y concreta de en donde se debe de impactar en la educación superior.

Justificación

Este proyecto tiene una gran relevancia tanto en el ámbito académico como social. Al identificar las variables que afectan el rendimiento académico de los estudiantes en las pruebas Saber Pro, se podrán implementar nuevas herramientas de trabajo y estrategias que mejoren no solo los resultados, sino también la educación superior en Colombia. Demarchi Sánchez (2023) señala los factores claves que tienen lugar en el desempeño estudiantil en las pruebas Saber Pro.

Los modelos de Machine Learning supervisados y no supervisados permitirán desarrollar soluciones para mejorar el rendimiento académico en dichas pruebas. Con este proyecto las instituciones educativas como la UNAD podrán implementar nuevas herramientas y estrategias de educación basados en los factores tanto personales, externos y educativos a los estudiantes (Gómez & Martínez, 2022). Las predicciones obtenidas darán resultados cualitativos y cuantitativos que darán guías de los cambios necesarios en los entornos de educación a nivel nacional (Pérez & Rodríguez, 2021). Esto va a permitir tener una educación superior de mayor calidad y con un impacto social importante, dado que para las empresas esto aumentará el interés y contratación de estos estudiantes (López & Hernández, 2020), además, los sistemas de información educativa pueden guardar gran cantidad de datos que tengan impacto en la educación superior (Romero & Ventura, 2013).

Objetivos

Objetivo General

Identificar las variables que afectan el rendimiento académico estudiantil en instituciones de educación superior al momento del desarrollo de las pruebas de ICFES Saber Pro con el uso de modelos de Machine Learning.

Objetivos Específicos

Analizar la base de datos de las pruebas Saber Pro 2021 y 2022 publicadas por el ICFES para la selección de los patrones y variables que influyen sobre el rendimiento académico de los estudiantes.

Realizar la selección de los modelos de Machine Learning, teniendo en cuenta la cantidad de variables y sus categorías.

Evaluar los modelos de predicción e identificación de las variables esenciales influyentes en el rendimiento académico de los estudiantes.

Marco Conceptual y Teórico

Este proyecto de investigación se basa en el análisis de las pruebas Saber Pro, un examen estandarizado utilizado en Colombia para evaluar el rendimiento académico de los estudiantes en los últimos semestres de sus carreras profesionales. Según el ICFES (2022), este examen mide competencias básicas y específicas, proporcionando un indicador clave del estado de la educación superior en el país. La importancia de estas pruebas radica en su capacidad para reflejar la calidad de la educación recibida en las instituciones de educación superior, así como las competencias adquiridas por los estudiantes
(Pérez & Rodríguez, 2021; Gómez & Martínez, 2022).

En este contexto, es crucial considerar diversos factores que pueden influir en el rendimiento de los estudiantes, tales como la calidad educativa de las instituciones, el entorno socioeconómico, y aspectos personales (Gómez & Martínez, 2022). Estos factores pueden ser cuantificados y analizados utilizando técnicas de Machine Learning, lo que permite identificar patrones y tendencias en los datos educativos (López & Hernández, 2020). La aplicación de estos modelos no solo ayuda a predecir el rendimiento, sino también a proponer mejoras en las estructuras educativas, apuntando a un impacto positivo en la educación superior y en los futuros profesionales del país (Pérez & Rodríguez, 2021). Las técnicas de Machine Learning utilizadas son arboles de decisión y Random Forest, que ayudan en la identificación de patrones y las relaciones en los datos que se tienen (Alonso & Fernández, 2019). Dichas técnicas se aplican a bases de datos de gran tamaño para obtener predicciones lo más exactas posibles.

Evaluación Educativa: La evaluación educativa es un proceso sistemático que busca medir y analizar el rendimiento de los estudiantes en diversas áreas del conocimiento. Las

pruebas Saber Pro son un ejemplo de evaluación educativa a gran escala, diseñada para evaluar las competencias adquiridas por los estudiantes en la educación superior. Esta evaluación puede ser formativa y sumativa, y la integración de ambas es esencial para una evaluación óptima, según (Looney, 2011). Cuando se combina evaluaciones formativas y sumativas fortalece el análisis en el rendimiento académico, que es fundamental a la hora de realizar observaciones, mejoras y cambios en los procesos de educación (Pérez & Rodríguez, 2021).

De la misma manera, el uso de modelos de Machine Learning en la educación ha ganado relevancia en los últimos años, debido a su capacidad para manejar grandes volúmenes de datos y descubrir patrones complejos que pueden no ser evidentes a simple vista. Estos modelos pueden ser supervisados o no supervisados, dependiendo de la naturaleza del problema y los datos disponibles. (Romero & Ventura, 2013) han destacado el potencial de estas técnicas para mejorar los sistemas de información educativa, permitiendo la personalización y optimización de los procesos de enseñanza y aprendizaje.

Así mismo, los factores socioeconómicos, como el nivel de ingresos familiares, el acceso a recursos educativos, y el contexto cultural, tienen un impacto significativo en el rendimiento académico. Además, factores internos como la motivación, el estilo de aprendizaje, y la calidad de la enseñanza recibida también juegan un papel crucial en los resultados obtenidos en las pruebas Saber Pro (EcoRfan, 2023). Las disparidades en el acceso a recursos educativos y las diferencias en el entorno escolar pueden explicar gran parte de las variaciones en el rendimiento académico de los estudiantes. (Coleman, 2019)

Hoy en día se tienen modelos de Machine Learning qué permiten predecir rendimientos académicos, uno de los más comunes es el algoritmo de Random Forest, el cual el que más se usa para estos proyectos de investigación, siendo una técnica de aprendizaje supervisado y por su

alta probabilidad de predicción. (Forero-Corba & Negre Bennasar, 2024) Así mismo, se tiene el algoritmo Gradient Boosting que es una técnica de Machine Learning que se ha utilizado para problemas de clasificación, en casos como la predicción del rendimiento académico ha sido aplicado de manera exitosa, permitiendo conocer los factores relevantes en dicho desempeño de los estudiantes. (Calva et al., 2021)

Metodología

Recolección de Datos

En la base de datos obtenida de la página de datos abiertos Colombia, inicialmente se realizó un filtro de la base de datos de ICFES saber pro, dicho filtro se indicó que se arrojaran únicamente los datos de los años 2021 y 2022. De este filtro se exportó un archivo tipo CSV con 57 columnas que contienen y 346658 filas, dichos datos se pueden observar en la Tabla 1, de las cuales se tomaron los datos más esenciales para el desarrollo de los modelos de machine Learning.

Tabla 1

Base de datos ICFES 2021-2022

Nombre de la Variable	Descripción de la Variable
PERIODO	Periodo académico
ESTU_CONSECUTIVO	Código identificación
ESTU_TIPODOCUMENTO	Tipo de documento
ESTU_PAIS_RESIDE	País de residencia
ESTU_COD_RESIDE_DEPTO	Código departamento
ESTU_DEPTO_RESIDE	Departamento
ESTU_COD_RESIDE_MCPIO	Código municipio
ESTU_MCPIO_RESIDE	Municipio
ESTU_CODDANE_COLE_TERMINO	Código DANE colegio
ESTU_COD_COLE_MCPIO_TERMINO	Código colegio bachiller
ESTU_COD_DEPTO_PRESENTACION	Código departamento presentación
INST_COD_INSTITUCION	Código institución
INST_NOMBRE_INSTITUCION	Nombre de institución
INST_CARACTER_ACADEMICO	Tipo de institución
ESTU_NUCLEO_PREGRADO	Nombre de estudio
ESTU_INST_DEPARTAMENTO	Departamento institución
ESTU_INST_CODMUNICIPIO	Código municipio institución
ESTU_INST_MUNICIPIO	Municipio de institución
ESTU_PRGM_ACADEMICO	Programa académico
ESTU_PRGM_DEPARTAMENTO	Departamento programa académico
ESTU_PRGM_CODMUNICIPIO	Código municipio programa académico
ESTU_PRGM_MUNICIPIO	Municipio programa académico
ESTU_NIVEL_PRGM_ACADEMICO	Nivel programa académico

ESTU_METODO_PRGM	Método del programa académico
ESTU_VALORMATRICULAUNIVERSIDAD	Valor de matricula
ESTU_DEPTO_PRESENTACION	Departamento de residencia
ESTU_COD_MCPIO_PRESENTACION	Código municipio de presentación
ESTU_MCPIO_PRESENTACION	Municipio de presentación
ESTU_PAGOMATRICULABECA	Estudiante beca
ESTU_PAGOMATRICULACREDITO	Pago matricula crédito
ESTU_HORASSEMANATRABAJA	Estudiante horas de trabajo
ESTU_SNIES_PRGMACADEMICO	SNIES programa académico
ESTU_PRIVADO_LIBERTAD	Estudiante privado de libertad
ESTU_NACIONALIDAD	Nacionalidad
ESTU_ESTUDIANTE	Estudiante
ESTU_GENERO	Genero
ESTU_COLE_TERMINO	Donde termino colegio
ESTU_PAGOMATRICULAPADRES	Pago de matrícula los padres
ESTU_ESTADOINVESTIGACION	Tipo de investigación
ESTU_FECHANACIMIENTO	Fecha de nacimiento
ESTU_PAGOMATRICULAPROPIO	Pago matricula propios medios
ESTU_TIPODOCUMENTOSB11	Tipo de documento b11
FAMI_EDUCACIONPADRE	Educación padre
FAMI_TIENEAUTOMOVIL	Tiene automóvil
FAMI_TIENELAVADORA	Tiene lavadora
FAMI_ESTRATOVIVIENDA	Estrato vivienda
FAMI_TIENECOMPUTADOR	Tiene computador
FAMI_TIENEINTERNET	Tiene internet
FAMI_EDUCACIONMADRE	Educación madre
INST_ORIGEN	Institución de origen
MOD_RAZONA_CUANTITAT_PUNT	Módulo razonamiento
MOD_COMUNI_ESCRITA_PUNT	Módulo comunicación escrita
MOD_COMUNI_ESCRITA_DESEM	Módulo comunicación escrita desempeño
MOD_INGLES_DESEM	Módulo ingles desempeño
MOD_LECTURA_CRITICA_PUNT	Módulo de lectura critica
MOD_INGLES_PUNT	Módulo ingles puntaje
MOD_COMPETEN_CIUDADA_PUNT	Módulo competencia ciudadana

Nota. La base de datos contiene los datos socioeconómicos, municipio de residencia, departamento, lugar de presentación de la prueba, información sobre sus padres, información sobre horas de trabajo semanal, el tipo de pago de matrícula y los puntajes de cada uno de los módulos, en total la dimensión de esta base de datos era de 57 columnas que contienen 346658 filas. *Fuente.* Datos Abiertos Colombia.

Los datos seleccionados a partir de un análisis de las variables más importantes al momento de desarrollar el modelo de machine Learning se muestran en la tabla 2.

Tabla 2

Datos que se utilizaron en los modelos de machine learning

Nombre de la variable	Descripción de la variable
ESTU_VALORMATRICULAUNIVERSIDAD	Valor de matricula
ESTU_PAGOMATRICULABECA	Estudiante beca
ESTU_PAGOMATRICULACREDITO	Pago matricula crédito
ESTU_HORASSEMANATRABAJA	Estudiante horas de trabajo
ESTU_GENERO	Genero
ESTU_PAGOMATRICULAPADRES	Pago de matrícula los padres
ESTU_PAGOMATRICULAPROPIO	Pago matricula propios medios
FAMI_ESTRATOVIVIENDA	Estrato vivienda
FAMI_TIENECOMPUTADOR	Tiene computador
FAMI_TIENEINTERNET	Tiene internet
INST_ORIGEN	Tipo de institución
MOD_RAZONA_CUANTITAT_PUNT	Módulo razonamiento
MOD_COMUNI_ESCRITA_PUNT	Módulo comunicación escrita
MOD_INGLES_PUNT	Módulo ingles puntaje
MOD_LECTURA_CRITICA_PUNT	Módulo de lectura critica
MOD_COMPETEN_CIUDADA_PUNT	Módulo competencia ciudadana

Nota. La tabla contiene los datos socioeconómicos, pago de matrícula, los módulos de competencia, institución de origen y valor de matrícula, dichos datos en total su dimensionalidad fue de 17 columnas de datos donde cada uno de estos datos contienen 346658 filas. *Fuente.* Datos Abiertos Colombia.

Modelos de Machine Learning

El modelo de Machine Learning Random Forest y Gradient Boosting se desarrollaron utilizaron la herramienta Python, estos fueron los que se utilizaron para el desarrollo del proyecto, los cuales se implementaron luego del proceso de segmentación, transformación, limpieza y codificación de las variables.

Para poder evaluar el funcionamiento de los modelos de Machine Learning, inicialmente en el código propuesto se realizó una transformación exhaustiva de cada una de las variables que no tenía un valor numérico establecido, dichas transformaciones permiten ingresar los datos de manera más fácil y rápido en el modelo de entrenamiento y poder así empezar a realizar las evaluaciones de su rendimiento.

Para realizar algunas las transformaciones para variables con valores tipo Si o No, Masculino o Femenino, se utilizó la función de pandas lambda la cual nos permite clasificar de manera binaria una variable y poder convertir estas categorías en forma numérica. Para poder incluir los nuevos valores se le dio una nueva columna a estas variables donde se utilizó la función Lambda de Pandas. También se utilizó la función get dummies para darle estructura a las variables categóricas con más de 2 categorías, por ejemplo, los estratos socioeconómicos se separaron en 7, luego se realizó el promedio de los puntajes de cada módulo para poder obtener el puntaje global que será nuestra variable de comparación para desarrollar el modelo de Random Forest.

Luego se utilizaron las librerías de sklearn donde se importó el modelo a utilizar que en este caso será Random Forest Classifier y Gradient Boosting, pero antes de esto se categorizó la variable Puntaje global en categorías para que pudiera ingresar en nuestro modelo, dicha variable tomará entradas bajas, medio y alto, a partir del puntaje global obtenido de los estudiantes. También se seleccionó las variables que irán en “X” eliminando puntaje global que será la variable objetivo en “y”. Luego se dividen los datos para poder empezar a realizar el entrenamiento del modelo.

Al finalizar la implementación de ambos algoritmos se observó las importancias de cada una de las variable utilizando la función forest.feature_importances.

Evaluación de Modelos

Después del desarrollo de los modelos, se emplearon técnicas de validación cruzada para asegurar la exactitud de los modelos, así como las métricas de rendimiento como la precisión, sensibilidad, especificidad, recall, y el F1-score para evaluar su efectividad.

Cronograma de Actividades

Tabla 3

Cronograma de las actividades que se van a desarrollar

Cronograma				
Actividad	**Mes 1**	**Mes 2**	**Mes 3**	**Mes 4**
Selección de variables principales	X			
Limpieza de los datos	X			
Clasificación de los datos		X		
Selección de modelos de machine learning		X		
Evaluación de modelos de machine learning		X		
Interpretación de resultados			X	X
Creación de recomendaciones				X

Nota. Esta tabla muestra el cronograma de actividades a desarrollar.

Recursos Necesarios

Tabla 4

Recursos necesarios

Recursos necesarios		
Recurso	Descripción	Presupuesto ($)
Equipo Humano	Científico de datos y analista de datos	$ 3000000
Equipos y Software	Computador Personal, Office, Visual Studio Code	$ 0
Viajes y Salidas de Campo	No aplica	$ 0
Materiales y suministros	Conexión a internet	$ 700000
Bibliografía	No aplica	$ 0
Total		$ 3700000

Nota. Esta tabla muestra los recursos necesarios para el desarrollo del proyecto.

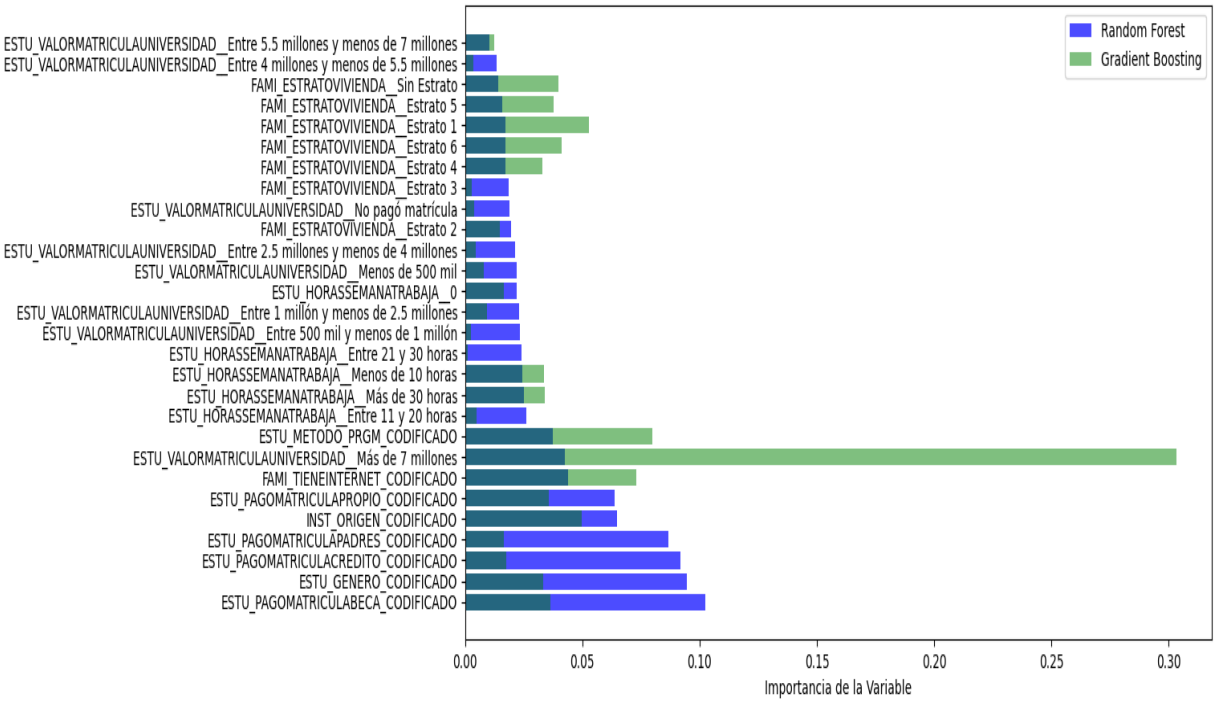
Análisis de Resultados

Identificación de Las Variables Con Más Impacto en el Puntaje Global de Las Pruebas Saber Pro

Los modelos desarrollados permiten observar claramente que los factores socioeconómicos, estrato social, el costo de la matricula, tienen una relevancia muy grande frente a la obtención de un buen resultado, los factores como las horas de trabajo semanales, institución de origen y el no tener horas de trabajo a la semana, muestran claramente que son variables esenciales a la hora de obtener excelentes pruebas. La figura 1 nos muestra los diferentes impactos gráficamente de las variables mencionadas con el uso del modelo Random Forest y el Gradient Boosting.

Figura 1

Importancia de las variables en el modelo random forest



Análisis de la Precisión de Los Modelos

La evaluación de los modelos Random Forest y Gradient Boosting, muestran métricas de rendimiento precisión, recall y f1-score. Ambos modelos tuvieron una buena precisión en las predicciones a nivel general, el modelo se sesga un poco para las clases altas y bajas, dado que el común de los resultados de los estudiantes en las pruebas saber pro se ubicó en el Medio que son valores comunes de estas pruebas globales. En la figura 2 y la figura 3 se muestran las diferentes métricas que se mencionaron anteriormente.

Figura 2

Matriz de confusión modelo random forest

Reporte de Clasificación (Precisión, Recall, F1-score):

	precision	recall	f1-score	support
Alta	0.12	0.01	0.01	1218
Baja	0.14	0.01	0.02	2246
Media	0.95	1.00	0.97	59982
accuracy			0.94	63446
macro avg	0.40	0.34	0.33	63446
weighted avg	0.90	0.94	0.92	63446

Precisión Global del Modelo: 0.94

Figura 3

Matriz de confusión modelo gradient boosting

	precision	recall	f1-score	support
Alta	0.00	0.00	0.00	1218
Baja	0.31	0.00	0.00	2246
Media	0.95	1.00	0.97	59982
accuracy			0.95	63446
macro avg	0.42	0.33	0.33	63446
weighted avg	0.90	0.95	0.92	63446

Conclusiones

Las variables estrato y valor de matrícula de la universidad, tienen una gran determinación frente a el puntaje global, lo que indica que las condiciones socioeconómicas influyen de manera directa los resultados de las pruebas Saber Pro, adicionalmente los estudiantes que tienen una carga laboral baja y los que no trabajan, tienden a lograr mejores resultados, lo que indica que la dedicación de estudio ayuda en el rendimiento académico de estas pruebas. Los modelos capturaron una buena cantidad de relación de las variables, pero se debe de tener en cuenta que para los puntajes bajos este modelo no las determino con tanta importancia, pero se pudo observar que ambos algoritmos presentan una precisión alta en la clase de calificaciones media, lo que indica que estos están ajustados correctamente para los puntajes medios.

Los modelos presentan ciertas dificultades a la hora de obtener las clases alta y baja, lo que indica que se puede realizar ciertas mejoras en estos, las métricas obtenidas como recall y f1-score indican que se debe de equilibrar las clases para no tener tanto sesgo entre ellas, pero se pudo ver que ambos predicen de manera correcta la clase Media, lo que indica que estos pueden predecir los estudiantes con desempeño promedio, también estos tienen una predicción inadecuada para las clases baja y alta, lo que indica que no les dieron relevancia a estos dos tipos de clase. Los dos muestran una alta fiabilidad en las predicciones de la clase Media, lo cual es un buen indicio para la mayoría de los estudiantes, pero la falta de precisión para las clases alta y baja sugiere que no están captando bien las variaciones entre estudiantes con valores extremos.

Recomendaciones

La implementación de técnicas como Deep Learning podrían ayudar a explorar patrones más complejos, también el desarrollo de visualizaciones más elaboradas para la muestra de los factores que influyen sobre el rendimiento académico. A partir de los resultados obtenidos se recomienda a las instituciones educativas que implementen programas para apoyar a los estudiantes con mayor riesgo de tener un puntaje bajo, adicionalmente se debe de tener programas de ayuda económica más activa, dado que se pudo observar que los factores socioeconómicos afectan en el rendimiento académico. Plantear mejoras en las acciones para preparar a los estudiantes para dichas pruebas, en especial los que deben de trabajar y estudiar en el mismo tiempo, dado que el tiempo que tienen para dedicar a este tipo de actividades es limitado.

Referencias

Alférez, G., Esteban, O., & Clausen, B. E. (2022). Automated machine learning pipeline for geochemical analysis. *Earth Science Informatics, 15*(3), 1683–1698. https://doi.org/10.1007/s12145-022-00821-8

Alonso, J. A., & Fernández, R. (2019). Aplicación de técnicas de machine learning en la educación superior. *Journal of Advanced Education, 14*(1), 88-105. https://doi.org/10.4567/jae.2019.14105

Calva, K., Flores, M., Porras, H., & Cabezas-Martínez, A. (2021). Modelo de predicción del rendimiento académico para el curso de nivelación de la Escuela Politécnica Nacional a partir de un modelo de aprendizaje supervisado. *Latin-American Journal of Computing, 8*(2), 61–70. https://lajc.epn.edu.ec/index.php/LAJC/article/download/264/159

Celestine, I., Kashif, B. A., Atharva, P., R., S., Moy, C. J., Swetha, P., & Ohyun, J. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health.* https://doi.org/10.3389/fpubh.2020.00357

Coleman, J. S. (2019). *Equality of educational opportunity* (Reprint ed.). Routledge.

Demarchi Sánchez, G. D. (2023). Factores que intervienen en el desempeño de los estudiantes en las Pruebas de Estado Saber Pro. *Ágora U.S.B., 23*(2), 490-502. https://doi.org/10.21500/16578031.62241

EcoRfan. (2023). Factores internos y externos que influyen en el aprendizaje. *Handbook of Education, Vol. III.*https://www.ecorfan.org/handbooks/Educacion%20T-III/ARTICULO%204.pdf

Forero-Corba, W., & Negre Bennasar, F. (2024). Técnicas y aplicaciones del machine learning e inteligencia artificial en educación: Una revisión sistemática. *RIED-Revista*

Iberoamericana de Educación a Distancia, 27(1), 209-253. https://doi.org/10.5944/ried.27.1.37491

Gómez, J. A., & Martínez, L. P. (2022). Factores asociados al rendimiento académico en estudiantes universitarios en Colombia. *Revista de Educación Superior, 15*(3), 45-60. https://doi.org/10.1234/revista.2022.4560

ICFES. (2022). Descripción de las competencias evaluadas en las pruebas Saber Pro. *Instituto Colombiano para la Evaluación de la Educación.* https://www.icfes.gov.co

Looney, J. W. (2011). Integrating formative and summative assessment: Progress toward a seamless system? *OECD Education Working Papers, 58.* https://doi.org/10.1787/5kghx3kbl734-en

López, C. R., & Hernández, D. M. (2020). Uso de modelos de machine learning para mejorar el rendimiento académico en pruebas estandarizadas. *Computers & Education, 10*(4), 789-805. https://doi.org/10.1016/j.compedu.2020.789

NumFOCUS. (2024). *Pandas documentation.* https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html

Pérez, M. T., & Rodríguez, S. (2021). Impacto de la motivación en el rendimiento académico de los estudiantes de secundaria. *Journal of Educational Research, 12*(2), 134-150. https://doi.org/10.5678/jer.2021.12134

Romero, C., & Ventura, S. (2013). Data Mining in Education. *WIREs Data Mining and Knowledge Discovery, 3*(1), 12–27. https://doi.org/10.1002/widm.1075