Variables socioeconómicas que inciden en el desempeño en las Pruebas Saber 11: Una aproximación usando Redes Neuronales

Autor

Julián David Rojas Rojas

Maestría Analítica Aplicada

Trabajo de Profundización

Directora de Tesis

Sulma Paola Vera Monroy

Profesora Facultad de Ingeniería

Co Director de Tesis

Felix Mohr

Profesor Facultad de Ingeniería

2023

PÁGINA DE ACEPTACIÓN

Yosimar Díaz Monterroza

Jurado 1

Carlos Augusto Mera Acosta

Jurado 2

Jairo Fernando Fernández Romero

Jurado 3

Directora de Tesis

Sulma Paola Vera Monroy

Co Director de Tesis

Felix Mohr

Chía, Julio del 2023

CONTENIDO

LISTA TABLAS Y GRÁFICAS

RESUMEN

Este trabajo presenta los resultados de un proyecto de aprendizaje de máquinas realizada en la base de datos "Saber 11 - 2019". El conjunto de datos comprende más de 500,000 observaciones, que abarcan más de 80 atributos. Entre estas variables se destacan atributos como "ESTU_GENERO" (género del estudiante) y "FAMI_EDUCACIONPADRE" (nivel de educación del padre). El objetivo de este estudio es observar cuáles son las variables socioeconómicas que más inciden en el desempeño de los estudiantes en las pruebas Saber 11, 2019.

La etapa de preprocesamiento de datos implica el manejo de valores faltantes y la codificación de variables categóricas utilizando mapeos y técnicas de embeddings. Las variables numéricas se normalizan utilizando técnicas de estandarización. Luego se construye un modelo de red neuronal utilizando PyTorch, que incorpora capas de embeddings para variables categóricas, capas totalmente conectadas, funciones de activación (ReLU) y capas de dropout para disminuir los riesgos de sobreajuste.

Los hiperparámetros se ajustan para garantizar un rendimiento óptimo. La tasa de aprendizaje se establece en 0,0003, el weight decay es de 0,0003 y la tasa de dropout es 0,1. El modelo se entrena para 1000 épocas utilizando un tamaño de lote de 128. Se emplean el optimizador de Adam y la función de pérdida de error cuadrático medio (MSE), con recorte de gradiente y técnicas de early stopping para mejorar la estabilidad del entrenamiento. Las métricas de evaluación, incluido la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE), se calculan para evaluar el rendimiento del modelo.

Además del entrenamiento y la evaluación del modelo, esta tesis introduce el concepto de Permutation Feature Importance (PFI) para medir la importancia de cada atributo. Al permutar los valores de los atributos individuales y observar el impacto en la pérdida de validación, se determina la importancia relativa de cada característica.

Los resultados muestran que las variables más importantes para predecir el desempeño de los estudiantes en las Pruebas Saber 11, 2019, son la jornada escolar (-2.0848), seguido por el departamento donde se encuentra la institución educativa (-1.8912), el nivel socioeconómico de los evaluados (-0,8611), el sexo del estudiante (-0,8495), y el tiempo dedicado a la lectura diaria (-0,6721). Por otro lado, la red neuronal óptima encontrada para este trabajo tiene una función de pérdida de 1429.2660, RMSE de 37.8051 y un MAE de 30.3248 para el conjunto de validación en la iteración 40 dentro del ciclo de entrenamiento de la red neuronal. Finalmente, este trabajo marca el inicio de un proceso investigativo para mejorar la educación del país. Estos esfuerzos buscan desarrollar estrategias que fomenten la apropiación del conocimiento y la formación a nivel superior, promoviendo así un ambiente educativo más sólido y enriquecedor.

Palabras clave: Saber 11, supervised machine learning, deep learning, red neuronal, Pytorch, embedding, modelo predictivo, hiperparámetros, Permutation Feature Importance (PFI), atributos socioeconómicos.

ABSTRACT

This work presents the findings of a machine learning project conducted on the "Saber 11 - 2019" dataset. The dataset comprises over 500,000 observations, encompassing 80 features. Among these features are notable inputs such as "ESTU_GENERO" (student gender), "FAMI_EDUCACIONPADRE"(father's education level),and "ESTU_HORASSEMANATRABAJA" (student's weekly working hours). The objective of this study is to observe which socioeconomical attributes impact the most the student's performance in the Saber 11 standardized test, for the year 2019.

The data preprocessing stage involves handling missing values and encoding categorical variables using mappings and embeddings. Numerical variables are normalized using standard scaling techniques. A custom neural network model is then constructed using PyTorch, incorporating embedding layers for categorical variables, fully connected layers, activation functions (ReLU), and dropout layers to mitigate overfitting risks.

The hyperparameters are tuned to ensure an optimal performance. The learning rate is set to 0.0003, weight decay is 0.0003, and the dropout rate is 0.1. The model is trained for 1000 epochs using a batch size of 128. The Adam optimizer and mean squared error loss function are employed, with gradient clipping and early stopping techniques to enhance training stability. Evaluation metrics, including root mean squared error (RMSE) and mean absolute error (MAE), are calculated to assess the model's performance.

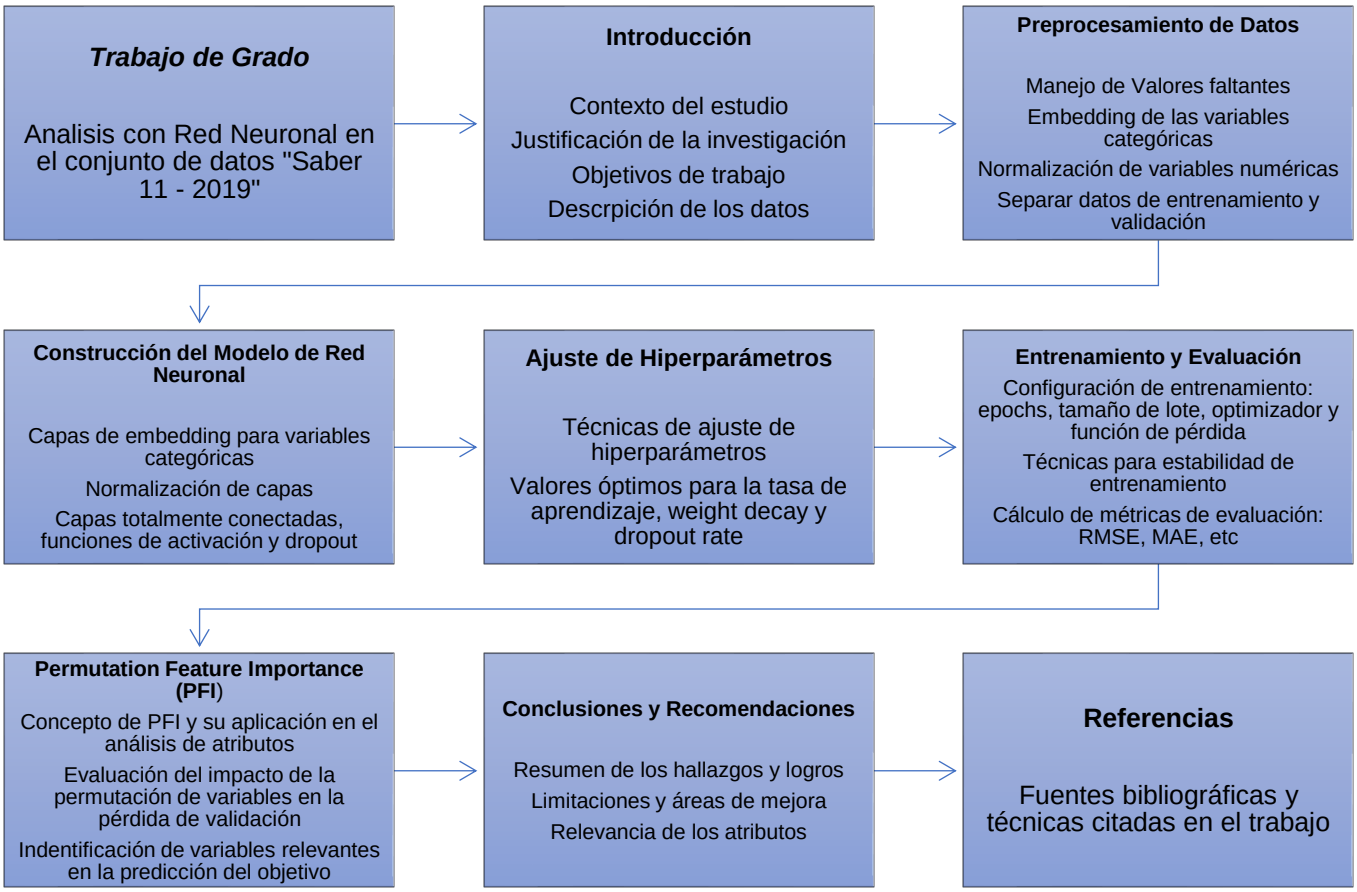
In addition to model training and evaluation, this thesis introduces the concept of Permutation Feature Importance (PFI) to measure the significance of each feature. By permuting the values of individual features and observing the impact on validation loss, the relative importance of each feature is determined.

The results show that the most important features in order to predict the performance of the student's standardized exam results is the school schedule (-2.0848), followed by the department where the educational institution is located (-1.8912), the socioeconomic level of the evaluated person (-0.8611), the student's sex (-0.8495), and the time dedicated to the daily reading (-0.6721). On the other hand, the optimal neural network found for this work has a loss function of 1429.2660, RMSE of 37.8051 and a MAE of 30.3248 for the validation set in the 40th iteration within the training loop of the neural network. Lastly, this work marks the beginning of a research process to improve the education system in the country. These efforts aim to develop strategies that foster the acquisition of knowledge and higher-level education, thus promoting a more solid and enriching educational environment.

Key words: Saber 11, supervised machine learning, deep learning, neural network, Pytorch, embedding, predictive model, hyperparameters, Permutation Feature Importance (PFI), socioeconomic features.

Resumen Gráfico

Trabajo de Grado



INTRODUCCIÓN

La educación es un derecho fundamental y es necesario en cualquier sociedad para reducir las desigualdades, impactar la calidad de vida de los individuos y proveer las bases para un desarrollo sostenible. Adicionalmente, la educación establece criterios y pautas de cómo relacionarnos con los demás en una sociedad, y promueve el desarrollo intelectual y emocional de los individuos. El acceso a la educación básica, media y secundaria ha contribuido significativamente a la mejora de diversos indicadores a nivel mundial. Un importante ejemplo que demuestra el progreso en el acceso educativo es la tasa de alfabetización, la cual refleja el nivel de acceso a la educación media. De acuerdo con un reporte emitido por la OECD (2014), *How was life? Global Well-being,* solamente el 12% de la población mundial podía leer y escribir en 1820; esto incremento a 86% en el 2016. La tasa de alfabetización es un indicador que se puede utilizar como un instrumento proxy para medir eficacia del sistema educativo y predecir la calidad de la futura mano de obra (The World Bank, 2022).

Sin embargo, es importante resaltar que el acceso a la educación no es suficiente para promover movilidad social y de este modo tener un impacto en el crecimiento de una región. Por este motivo, tiene que haber políticas nacionales que no solo se enfoquen en el acceso a educación sino también abogar por que las instituciones públicas educativas garanticen **calidad educativa.** En el 2015, la ONU estableció una agenda para el desarrollo sostenible con el objetivo de erradicar la pobreza, proteger el planeta y asegurar la prosperidad de todos. Una de las metas de sostenibilidad de las Naciones Unidas (la número 4), es garantizar una educación inclusiva, equitativa y de calidad y promover oportunidades de aprendizaje durante toda la vida para todos (Naciones Unidas, 2015b). La educación le da al individuo oportunidades para ser libre dentro de la sociedad y, asimismo, fomenta el desarrollo de la cultura, la economía, el deporte, la investigación, la infraestructura, entre otros.

En abril del 2020 cerca de 1600 millones de niños y jóvenes en el mundo estaban afuera de la escuela (Naciones Unidas, 2015b). Esta cifra es alarmante ya que, según un informe de las Naciones Unidas (2015a), la educación de calidad es primordial para salir del ciclo de pobreza y reducir desigualdades:

> Cuando las personas acceden a educación de calidad, pueden escapar del ciclo de pobreza. Por consiguiente, La educación contribuye a reducir las desigualdades y a lograr la igualdad de género. También empodera a las personas de todo el mundo para que lleven una vida más saludable y sostenible. La educación también es fundamental para fomentar tolerancia entre las personas y contribuye a crear sociedades más pacíficas. (p.1)

Colombia es uno de los países más desiguales de Latinoamérica (Valora Analitik, 2021), y esto se ve reflejado en la gran disparidad salarial o en los altos niveles de informalidad en el mercado. Esto está relacionado con la falta de oportunidades asociado con un pobre sistema educativo. La población, en general, tiene acceso a la educación, pero la calidad educativa, sobre todo en colegios rurales, es pésima.

En el mundo existen varias pruebas estandarizadas para medir la calidad de enseñanza dentro de las instituciones educativas. En Colombia, el examen de Estado Saber 11, son unas pruebas aplicadas por el Instituto Colombiano para la evaluación de la educación (ICFES), diseñadas para evaluar las competencias que deben estar alineadas al currículo y son referidas a criterios (Rivas y Scasso, 2017).

Según el Ministerio de Educación Nacional (2016):

> Los resultados de estas evaluaciones y el análisis de los factores asociados que inciden en los desempeños de los estudiantes permiten que los establecimientos educativos, las secretarias de educación, el Ministerio de Educación Nacional y la sociedad en general, identifiquen las destrezas, habilidades y valores que los estudiantes colombianos desarrollan durante la trayectoria escolar, independientemente de su procedencia, condiciones sociales, económicas o culturales, con lo cual, se puedan definir planes de mejoramiento en sus respectivos ámbitos de actuación.

Lo anterior, descrito por el Ministerio de Educación Nacional, es algo contradictorio ya que los factores que inciden en el desempeño de los estudiantes pueden estar asociados a sus condiciones sociales, económicas y culturales. La realidad es que las destrezas, habilidades y valores que los estudiantes colombianos adquieren no son independientes de su procedencia. Como consecuencia, lo que se quiere hacer en este estudio es identificar qué factores socioeconómicos inciden en el desempeño de diferentes estudiantes de undécimo grado en distintas instituciones educativas en Colombia, para una base de datos con más de 500,000 estudiantes que presentaron la prueba Saber 11 durante el segundo semestre del 2019.

JUSTFICACIÓN

Parte del objetivo final de la educación es promover la **movilidad social de los individuos**. La movilidad social la definiremos como un aumento equitativo de las oportunidades de las personas en temas tales como salud, educación, e ingreso a lo largo de la vida de una persona en particular y entre generaciones (undp, 2015). Y para que esto se dé, se **debe garantizar calidad en la educación**. Partiendo de esa base, se puede plantear la pregunta sobre qué factores influyen realmente en la existencia de un aprendizaje sostenible y pertinente dentro de las instituciones educativas. Zhang et al. (2021) denotan la importancia de la educación media como crucial ya que proporciona las bases para el crecimiento futuro de la juventud a adultos bien preparados y resilientes para afrontar diferentes adversidades que se presentan en la vida. Ellos también aprecian que son pocos los estudios que se han hecho en cuanto al desempeño académico de los estudiantes en zonas rurales. Por otro lado, dentro del contexto nacional, Ortega et al. (2021) muestran que mientras haya más apoyo familiar en el proceso educativo de los estudiantes, la probabilidad de obtener un resultado promedio del 50% es 4.650 más alto. Por lo tanto, este estudio proporciona información relevante sobre la importancia de estudiar estas variables socioeconómicas en Cordoba, Colombia.

En buena medida, la educación debe promover espacios de dialogo, debate y escucha, los cuales son elementos básicos para vivir en una democracia pluralista, igualitaria y en paz. Para profundizar un poco más en este tema, cabe mencionar que la UNESCO (Flotts et al., 2016) concibe la educación como un medio de transformación de las personas y comunidades entendiendo que el sujeto educado se transforma en espíritus de paz, armonía y reconciliación con el otro (Ortega et al., 2021). Por eso, para erradicar la pobreza y enfocarse en construir un desarrollo sostenible hay que adoptar políticas educativas orientadas en la dignidad humana y la protección de derechos humanos.

Asimismo, la educación debe fomentar un ambiente donde el individuo desarrolle un criterio reflexivo y crítico, halle sus pasiones y adquiera los conocimientos prácticos y teóricos necesarios para convivir en sociedad. Ahora, con relación a este último, hay ciertos saberes fundamentales que son esenciales observar y monitorear ya que son aptitudes transcendentales que le dan al individuo mejores herramientas para enfrentar distintos retos que se presentan a lo largo de la vida. De ahí la importancia de identificar qué factores sociales y económicos están influenciando el desempeño académico de los estudiantes en las pruebas Saber 11, ya que los resultados de estas pruebas dan ciertos indicios de cómo los estudiantes se están apropiando de estas herramientas fundamentales (tales como las matemáticas, la escritura, la lectura, la inferencia, entre otros) para la mejor tomar decisiones pertinentes al resolver problemas en un contexto determinado.

Al poder diagnosticar qué variables socioeconómicas influyen en el desempeño de los estudiantes en los diferentes colegios del país, se podrán adoptar políticas educativas que incidan en el sistema educativo colombiano y, de esta manera, hacer más competitivo a Colombia en términos regionales e internacionales con mano de obra mejor preparada y calificada. La UNESCO le da cierto énfasis a este eje cuando habla sobre la relevancia de la calidad educativa ya que debe responder a las necesidades de la sociedad. En este sentido, Ortega et al. (2021) resaltan un factor interesante y es el papel del sector educativo como respuesta a los retos de la globalización y su papel en la innovación y el desarrollo de un país. Álvarez-López y Matarranz (2020) muestran cómo existen ciertas corrientes teóricas que soportan la idea de que la escuela, docentes y estudiantes son partes de un sistema de intercambio comercial que se rigen bajo criterios del mercado y los principios económicos de la eficacia y eficiencia de los recursos. Solo de esta forma, según lo descrito anteriormente, habrá una oportunidad para que muchas personas puedan salir de la pobreza extrema, de la informalidad laboral y promover movilidad social.

Los resultados serán insumo para tomar decisiones en el sector público y privado ya que los llevaría a focalizar intervenciones en el sistema educativo de una manera más coordinada y descentralizada.

Con todo lo anterior se puede notar un fuerte contraste sobre el énfasis de la educación. Más allá del debate acerca de cómo se deben abarcar la educación y analizar las pruebas de Estado para los estudiantes que van a ingresar a la educación superior, hay que notar la importancia de evaluar competencias, como dice Rivas y Scasso (2017), lo cual implica concebir el conocimiento como un conjunto de disposiciones y capacidades en acción. En otras palabras, según Rivas y Scasso (2017), solo si los estudiantes aprenden de tal manera que

puedan ser capaces de usar el conocimiento para salir de su destino social se cumplirá el derecho pleno de la educación. Nos permite una comprensión integral de realidades y la toma de decisiones acertada de los gobiernos centrales frente a los procesos educativos (Ortega et al., 2021).

MARCO TEÓRICO

ESTADO DEL ARTE

Zhang et al. (2021), trataron una muestra similar en cuanto a población (muestra de 93 estudiantes en educación media): tomaron los resultados de las pruebas académicas entre estudiantes de zonas rurales. Este estudio buscó comparar los logros académicos de los estudiantes de dos grados (11 y 12 grado) en diferentes áreas del conocimiento como chino, inglés, Biología y física, con el objetivo de observar si hubo una manifestación efectiva de la calidad educativa en China suroccidental. Parte de su investigación también se enfocaba en ver si los logros académicos de los estudiantes variaban de acuerdo con atributos tales como el género y el grado en que cursaban.

En este aspecto los hallazgos son interesantes: Zhang et al. (2021) encontraron que, en la providencia de Guizhou, en China, no había una diferencia sustancial entre en los resultados generales de los estudiantes de grado 11 y 12. De los factores que los autores encontraron para explicar este resultado se resalta la falta de infraestructura adecuada, la falta de profesores cualificados y la escasez de recursos educativos. Estos tres factores restringen el desempeño académico de los estudiantes y cohíbe a que haya un desarrollo educativo inclusivo (Zhang et al., 2021). Otra interpretación que encontraron los autores para explicar los resultados es el contexto familiar de los estudiantes. El éxito académico puede estar influenciado por qué tan involucrado se encuentran los padres en las actividades escolares de sus hijos, la educación de los padres y el estrato socioeconómico de la familia. Por otro lado, el estudio encontró que hay una brecha significativa entre los resultados de los hombres y las mujeres, sobre todo en áreas tales como matemáticas, física y biología. Zhang et al. (2021) le atribuyen esto a realidad del contexto rural donde las mujeres tienen que ocuparse más por tareas del hogar. Además, en regiones rurales de China, todavía conservan actitudes tradicionales sobre los roles de género y sus responsabilidades especificas dentro de la comunidad.

Ortega et al. (2021) examinaron las diferentes formas de acompañamiento de la familia en el proceso educativo de sus hijos a partir de la variabilidad de los promedios de la prueba de Estado 2016 obtenidos por 249 instituciones educativas oficiales en 27 municipios en Córdoba, Colombia. Ortega et al. (2021) hallaron que no hay una relación lineal significativa entre el grado de acompañamiento familiar y la variabilidad de resultados obtenidos por las instituciones según los resultados de las pruebas SABER 11 del 2016. Sin embargo, hubo evidencia que, entre mayor acompañamiento de la familia en el proceso educativo, la probabilidad de obtener un puntaje promedio entre los 50% más alto es 4.650 mayor. Vivir en zona urbana, los municipios entre 6 y 10 instituciones educativas, las amas de casa y familias con acceso a

computador fueron las variables explicativas con mayor influencia en el acompañamiento familia. Otra conclusión importante de Ortega et al. (2021) es que hay poca pertinencia de las Pruebas Saber 11 respecto a los aprendizajes adquiridos y los tipos de evaluación aplicados en las escuelas a sus estudiantes.

Blanco (2015) realizó un análisis del desempeño académico en el Examen de Estado para el ingreso a la Educación Superior. Con una muestra de 11,329 estudiantes del Departamento del Cesar, que presentaron la prueba Saber 11 2012-2, halló que el entorno socioeconómico del estudiante incide en los resultados de desempeño académico, encontrándose que, a mayor nivel socioeconómico del estudiante y su familia, mayor es el puntaje en los resultados de la prueba. También encontró que los grupos con alta población de la zona rural tienden a presentar menores niveles de desempeño, y son grupos en los cuales se presenta que el nivel de escolaridad de los padres llega solo a un nivel de escolaridad de básica primaria y en pocos casos de secundaria. Blanco (2015), a su vez, evidenció que los estudiantes cuyos padres cuentan con estudios de postgrado tienen mayores posibilidades de obtener mayores resultados en la prueba Saber 11.

Timarán-Pereira et al. (2019) aplicaron la minería de datos educativa para descubrir factores asociados al desempeño académico, de los estudiantes colombianos de educación media que presentaron las pruebas Saber 11 entre los años 2015 y 2016. Usaron datos validos sobre los factores socioeconómicos, académicos e institucionales correspondientes a 1.061.680 estudiantes, donde se **escogió el atributo puntaje global como clase**. Con los resultados obtenidos en un modelo de árbol de decisión, se pudo observar que este clasifica correctamente a 711.116 instancias, que corresponde a un porcentaje de precisión de 67%, mientras que 350.664 instancias fueron clasificadas incorrectamente, lo cual corresponde a un porcentaje del 33%.

Timarán-Pereira et al. (2019) encontraron que los atributos con mayor ganancia de información que forman parte de los patrones descubiertos, asociados con el buen desempeño académico en las pruebas Saber 11 son: el estrato socioeconómico (medio o alto), jornada de estudio en la mañana o completa, el índice TIC regular y la edad menor que 18 años. Por otro lado, los atributos con mayor ganancia de información que forman parte de los patrones descubiertos, a un bajo desempeño en estas pruebas son: el estrato socioeconómico bajo, el índice TIC bajo y el nivel de SISBEN 1.

En otro estudio similar, Timarán et al. (2019) usaron la misma metodología con la misma base de datos, pero esta vez utilizando el puntaje de matemáticas como atributo respuesta. Este estudio arrojó que los atributos con mayor ganancia de información que forman parte de los patrones descubiertos, se destacan el estrato socioeconómico, la jornada de estudio, el índice TIC, la edad y el sexo de los estudiantes como factores importantes asociados al buen o bajo desempeño académico de los estudiantes en la prueba de matemáticas.

Ávila et al. (2021) usaron un modelo de machine learning empleando algoritmos de aprendizaje de maquina supervisado para analizar los resultados de la prueba Saber 11 en Colombia desde el periodo 2017–1 hasta el periodo 2021-1, interpretando las diferentes variables socioeconómicas para hallar relación con los puntajes obtenidos por los estudiantes, tanto en

colegios calendarios A y B. Consecuentemente, implementaron modelos de minería de datos y análisis estadísticos enfocados a la prueba Saber 11. Ávila et al. (2021) encontraron que los atributos más influyentes son: la educación de los padres, el estrato, la edad, el número de libros que tiene la familia, el tiempo de lectura diario, e incluso la jornada del colegio. El género también es importante resaltar ya que los hombres suelen tener un puntaje más alto que el de la mujer. Asimismo, Ávila et al. (2021) mostraron que los estudiantes con mayor acceso a internet, y con acceso a medios tecnológicos como computadores, mejoran los resultados obtenidos en la prueba Saber.

García-González y Skrita (2019) buscaron predecir el desempeño académico de los estudiantes que presentaron el examen de Estado de 2016 para acceder a la educación superior a partir de las observaciones y características familiares propias del estudiante. Usando arboles de decisión, García-González y Skrita (2019) mostraron que las variables familiares que mejor predicen los resultados académicos son (en orden): el nivel educativo de la madre, el estrato socioeconómico de la vivienda, el número de libros, el nivel educativo del padre y el poseer computador en la vivienda.

BASES TEÓRICAS

Para aplicar con éxito un modelo predictivo que permita observar qué variables socioeconómicas inciden en el desempeño de los estudiantes de Undécimo grado a la hora de presentar la Prueba de Estado 11, es necesario emplear conceptos relacionados con el aprendizaje de máquinas. En este proceso se utilizaron diferentes parámetros e hiperparámetros dentro de una red neuronal después de preprocesar la data.

En la revisión de literatura se observa que Zhang et al. (2021) emplearon un análisis multivariado de covarianza (MANCOVA). Un análisis descriptivo fue usado para ver desviación estándar y media de los dos grupos con base en los resultados en cada uno de los componentes de la prueba. Test de Shapiro-Wilk de normalidad mostró que los datos se distribuían normalmente para los resultados en chino e inglés, pero no para los datos en física y biología. Para asegurar que las condiciones de un Análisis de Varianza ANOVA para comparar los resultados de los dos grupos, se normalizaron los datos por medio de una transformación de rangos.

Ortega et al. (2021) analizaron los datos con el método de **análisis de componentes principales**. En su estudio usaron una encuesta para la recolección de datos que se estructuró en seis bloques. El primer bloque correspondía a 5 preguntas informativas sobre la institución educativa y su ubicación; el segundo bloque era integrado por 6 ítems acerca de los aspectos sociodemográficos de la muestra; el bloque 3 contaba con 2 preguntas sobre la situación ocupacional del representante del hogar; el bloque 4 preguntaba sobre el nivel económico del padre de familia; el bloque 5 tenía 8 ítems acerca de las características físicas y servicios de la vivienda; y finalmente el bloque 6 era el referente a la participación y acompañamiento escolar. Ortega et al. (2021) procedieron a crear medidas prescriptivas usando test de *Chi cuadrado, V de Cramer, Correlaciones (Spearman)*. Usaron pruebas de hipótesis para comparar grupos de

interés abordados, y finalmente para los datos sobre participación y acompañamientos familiar (bloque 6), se aplicó el análisis de componentes principales.

Blanco (2015), Timarán-Pereira et al. (2019), Ávila et al. (2021) implantaron técnicas de minería de datos para establecer relaciones entre variables socioeconómicas y los resultados de la prueba SABER 11. Blanco (2015) usó un Cross-Industry Standard Process for Data Mining, conocida como CRISP-DM, como metodología para el proyecto de minería de datos. Posteriormente, conformó el conjunto de datos para la aplicación de un algoritmo *K-means* que posibilitó el agrupamiento sin supervisión del conjunto de datos para obtener el modelo descriptivo. Timarán-Pereira et al. (2019) emplearon un enfoque similar: una metodología CRISP-DM para procesar la data, y los resultados fueron obtenidos utilizando un modelo de clasificación basado en arboles de decisión.

Similarmente, Ávila et al. (2021) utilizaron los árboles de decisión para generar reglas de clasificación en dos categorías: “Por encima de la media” o “Por debajo de la media”. Adicionalmente, usaron CRISP-DM y un algoritmo de regresión SGD Regressor para ver el impacto de las variables socioeconómicas en el puntaje global de los estudiantes. Por último, García-González y Skrita (2019) emplearon igualmente un árbol de decisión para la clasificación y la predicción de los resultados académicos en las pruebas Saber 11-2016, con base en las características familiares de los estudiantes.

EL CONTEXTO

En este apartado se realiza una descripción del contexto de las pruebas de estado, a nivel micro y macro, y qué pasa con las pruebas masivas a nivel internacional, Latinoamérica y en Colombia. Ahora bien, ¿por qué la importancia de las pruebas Saber 11 en Colombia? Las pruebas estandarizadas son un insumo útil para que las instituciones educativas mejoren su calidad (Demarchi, 2020). Adicionalmente, son una forma de hacerle seguimiento a la educación que están recibiendo los estudiantes en los estamentos educativos y además evalúan la capacidad que tiene el estudiante para responder a situaciones en su propio contexto, con base en los conocimientos adquiridos (Demarchi, 2020). En Colombia, una de las dimensiones de calidad consideradas por el Ministerio de Educación Nacional es el desempeño estudiantil, el cual ha cobrado gran importancia para el estudio de factores socioeconómicos asociados con el estudiante y su desempeño cuando ingresan a la educación superior (González, 2019). De aquí que se tenga un especial interés por el análisis y procesamiento de datos de la prueba de Estado con los estudiantes de undécimo grado en Colombia, para el estudio de variables socioeconómicas asociado con el estudiante.

En 1968 en Colombia, apareció el sistema de evaluación como una manera de verificar cómo se desarrollaban los currículos en los estamentos educativos (Demarchi, 2020). De la ley 1324 del 2009, se fijan criterios para un examen obligatorio para todos los niveles de formación académica. En esta Ley, el Estado toma la función de inspección y vigilancia de la educación mediante la aplicación de exámenes y otras pruebas externas, con el propósito de medir el nivel de cumplimiento de los objetivos y buscar el mejoramiento continuo de la calidad de la

educación. Para lograr esto, el Estado busca que las evaluaciones se rijan en función de los siguientes principios:

1. Participación: busca promover la participación creciente de la comunidad educativa en el diseño de los instrumentos y estrategias de evaluación.
2. Equidad: supone reconocer desigualdades existentes en los contextos de aprendizaje y asumir un compromiso proactivo por garantizar la igualdad de oportunidades para acceder a una educación de calidad.
3. Descentralización: promueve la formación del recurso humano en el nivel territorial y local.
4. Cualitativa: promueve la realización de ejercicios cualitativos, de forma paralela a las pruebas de carácter cuantitativo, que contribuyan a la construcción de explicaciones de los resultados en materia de calidad.
5. Pertenencia: se debe valorar de manera integral los contenidos académicos, los requerimientos del mercado laboral y la formación humanística del estudiante.
6. Relevancia: busca evaluar el grado de asimilación de un conjunto básico de conocimientos que sean exigibles no solo en el contexto nacional, sino en el contexto internacional, de tal manera que un estudiante pueda desempeñarse en un ámbito global competitivo (*ley 1324 del 2009*)

El Sistema Nacional de Evaluación Estandarizada (SNEE) tiene conformado cuatro pruebas Saber: la primera se aplica a los grados tercero, quinto y noveno, en la educación básica; la segunda se aplica a los grados undécimo de la educación media y es conocida como la prueba Saber 11; La tercera es Saber TyT que se aplica a las personas que se gradúan de niveles técnicos y tecnólogos. Y finalmente tenemos la prueba Saber Pro, la cual la presentan estudiantes en sus semestres académicos de educación superior. Hoy en día, desde el Ministerio de Educación Nacional, en cabeza del ICFES (Instituto Colombiano para la Evaluación de la Educación) se dirige y coordina el diseño, producción y aplicación de las pruebas, y el procesamiento y análisis de los resultados del Examen.

Ahora, es importante saber qué pasa en Latinoamérica y cómo funcionan estas pruebas en la región. Para los países latinoamericanos, la década de los noventa fue una época marcada por cambios en ámbitos económicos y sociales. Al adaptar un esquema neoliberal en varios países de Latinoamérica, se crearon nuevas exigencias en términos sociales, y esto incluyó cambios en el sector educativo (Demarchi, 2020). Con la llegada de nuevo conocimiento, se hace necesario reconocer las capacidades de las personas para desenvolverse profesionalmente y desarrollar una tarea especifica. Como consecuencia a esto, surge un modelo de educación por competencias y, junto a esta, nacen distintas entidades evaluadoras para medir las competencias obtenidas por los estudiantes en su etapa formativa. Para dar algunos ejemplos en Latinoamérica, Demarchi (2020) muestra cómo se fueron implementando gradualmente los sistemas de evaluación en la región. En México se creó el Instituto Nacional para la Evaluación Educativa (INEE), en Chile se estableció el Sistema Nacional de Evaluación de Resultados de Aprendizaje (SIMCE), Argentina fundó el Sistema Nacional de Evaluación de Calidad (SINEC), Brasil estableció el Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) y para el caso colombiano se formó el Sistema de Evaluación de la Calidad de la Educación Primaria y Secundaria.

Todos los sistemas mencionados tienen el objetivo de hacer seguimiento a los procesos educativos de los estudiantes en los diferentes niveles de formación. Lo que se quiere con las pruebas estandarizadas en este punto es informar a la sociedad y a las instituciones educativas sobre el estado de la educación del país (Demarchi, 2020). Cada país utiliza estas pruebas según un criterio diferente y con objetivos un poco distintos. Mientras en países de América latina y Norteamérica se utilizan estas pruebas estandarizadas como un indicador para hacer un ranking de las instituciones educativas de peor a mejor calificadas, en países como en Dinamarca y en Holanda los utilizan como un criterio de reflexión sobre la educación impartida por los docentes y buscar diferentes posibilidades pedagógicas que permitan el desarrollo y las capacidades de los estudiantes (Demarchi, 2020).

Teniendo en cuenta lo anterior, es imperativo tener un sistema evaluativo que valore de forma transversal el estado educativo de varios países a nivel internacional. Por esto, en el año 1950, se estableció una agencia de evaluación donde psicólogos, sociólogos y otros profesionales sociales de la UNSECO habilitan discusiones sobre las dificultades en la evaluación escolar y estudiantil. En consecuencia, se crea la Asociación Internacional para la Evaluación del Logro Educativo (IEA) con el objetivo de comparar internacionalmente los rendimientos educativos y comprender la eficiencia y efectos de las políticas relacionadas con el sistema educativo (Demarchi, 2020).

Para comprender un poco mejor cómo funciona, se describirá a continuación cuáles son las pruebas internacionales más destacadas, así como las competencias que evalúa y a los cursos en los cuales se enfoca dicha evaluación.

1. **TIMSS (Estudios sobre las tendencias en ciencias y matemáticas):** fue establecida en 1995 donde se evalúa las competencias adquiridas por los estudiantes en Matemáticas y Ciencias. Esta fue unas de las primeras pruebas aplicadas trasnacionalmente y son aplicadas cada cuatro años. Las pruebas TIMSS evalúa también contenidos dentro del currículo en tres niéveles: currículo pretendido (planificación educativa de las instituciones), currículo aplicado (enseñanza real que tuvo el estudiante en el aula) y currículo logrado (logro del estudiante en el aprendizaje de cada materia impartida). Esta prueba se aplica en Colombia, México, Argentina y Chile, a los estudiantes de los grados 3°, 4° a 7°, y 8° a 11°.

2. **LLECE (Laboratorio para la Evaluación de la Calidad de la Educación en América Latina y el Caribe):** se creó entre los años 1995 y 1997 en Santiago, Chile. Esta prueba tiene como objetivo:

> *Producir información sobre logros de aprendizaje y evaluación de sistemas educativos y sus componentes, analizando los factores a dichos avances. Constituirse y validarse en tanto foro de reflexión, debate e intercambio de nuevos enfoques y aproximaciones en evaluación de la calidad educativa. Contribuir a fortalecer las capacidades locales de las unidades de evaluación* (Dirección de Gestión y Evaluación de la Calidad (DGEC), 2019).

El primer estudio de laboratorio contó con la participación de 13 países entes ellos Argentina, Bolivia, Brasil y Colombia, y evalúa a los estudiantes de 3° y 4°.

3. **ICCS (Estudio Internacional de Educación Cívica y ciudadana):** se creó en el año 1999, y a diferencia de las demás pruebas, esta busca conocer las competencias, conocimientos o habilidades relacionadas con los temas de civismo, ciudadanía, democracia, identidad nacional, cohesión social y diversidad. La primera prueba fue aplicada en 1971 y la segunda fue aplicada en 1999 con el nombre de Estudio sobre Educación Cívica (CIVED), y medía principalmente dos aspectos: los conocimientos adquiridos en las instituciones educativas y el aprendizaje aplicado en contextos cotidianos. Esta prueba tuvo éxito en varios países europeos y latinoamericanos y se continúa aplicando en varias instituciones educativas. Se aplica en Colombia y Chile a estudiantes de 14 años.

4. **PISA (Programa Internacional para la Evaluación de Estudiantes):** esta es una prueba promovida por la OCDE. Fue creada en 1997 y se aplicó por primera vez en año 2000. Las competencias evaluadas son Matemáticas, Lectura y Ciencias y es aplicada cada tres años. Más de ochenta países han participado incluyendo Canadá, China Irlanda, Singapur y Japón, y a nivel latinoamericano, Argentina, Brasil, Colombia, Chile, entre otros. Esta prueba se aplica a los estudiantes con 15 años ya que están próximos a finalizar su etapa de escolarización y, por ende, se busca conocer sus conocimientos y destrezas que los adolescentes adquirieron durante su etapa de formación educativa, para desenvolverse de manera útil en la sociedad, según la OCDE. Los resultados ayudan a países a reflexionar, cuestionar y analizar su sistema educativo como es el caso de México o Chile que ha encontrado una diferencia significativa entre el desempeño de los estudiantes del área rural frente a los estudiantes de áreas urbanas (Demarchi, 2020)

5. **PIRLS (Progreso Internacional Estudio de Alfabetización en Lectura):** esta prueba es aplicada cada cinco años empezando su ciclo en el año 2001. PIRLS busca conocer las competencias relacionadas con la comprensión lectura de los estudiantes en los primeros años escolares (9 años). Además de Colombia y Honduras, han participado varios otros países de diferentes continentes como Europa, Asia y algunos territorios africanos, y del continente americano han participado países como Estados Unidos y Canadá.

Con relación a lo que se va a analizar en este trabajo, Miranda Schleicher (2009) encontraron que los resultados de las pruebas estandarizas se encuentran asociadas a factores económicos, familiares y contextuales, lo cual nos dice que la responsabilidad de formación no solo recae en la institución educativa, sino que es un proceso de construcción en conjunto.

El uso de la analítica de datos en las diferentes instituciones educativas ha tenido un crecimiento notable durante los últimos años para encontrar patrones ocultos en el proceso de aprendizaje de los estudiantes (Ávila et al., 2021). Con base en Ávila et al. (2021), los objetivos

del “Learning Analytics” es ayudar a los profesores y asesores a determinar qué estudiantes pueden estar en riesgo y quienes enfrenta dificultades en su carrera académica.

MARCO LEGAL (DECRETO)

Dentro del contexto legal, se puede extraer más detalles sobre la importancia de la evaluación realizada mediante los exámenes de Estado. Además de ley 1324 del 2009, mencionada anteriormente, El Ministerio de Educación Nacional, mediante el decreto 869 del 2010, reglamenta específicamente el Examen de Estado de la Educación Media, ICFES-SABER 11. Según el Ministerio de Educación Nacional, el propósito principal de estas pruebas SABER es: “contribuir al mejoramiento de la calidad de la educación colombiana mediante la realización de evaluaciones aplicadas periódicamente para monitorear el desarrollo de las competencias básicas en los estudiantes de la educación básica, como seguimiento de la calidad del sistema educativo” (Ministerio de Educación Nacional, 2016).

El decreto 869 del 2010 define El Examen de Estado de la Educación Media, ICFES-SABER 11°, que aplica el Instituto Colombiano para la Evaluación de la Educación (ICFES), como un instrumento estandarizado para la evaluación externa, que conjuntamente con los exámenes que se aplican en los grados 5°, 9° y al finalizar el pregrado, hace parte de los instrumentos que conforman el Sistema Nacional de Evaluación. Los objetivos de esta evaluación son:

A. Comprobar el grado de desarrollo de las competencias de los estudiantes que están por finalizar el grado undécimo de la educación media.
B. Proporcionar elementos al estudiante para la realización de su autoevaluación y el desarrollo de su proyecto de vida.
C. Proporcionar a las instituciones educativas información pertinente sobre las competencias de los aspirantes a ingresar a programas de educación superior, así como sobre las de quienes son admitidos, que sirva como base para el diseño de programas de nivelación académica y prevención de la deserción en este nivel.
D. Monitorear la calidad de la educación de los establecimientos educativos del país, con fundamento en los estándares básicos de competencias y los referentes de calidad emitidos por el Ministerio de Educación Nacional.
E. Proporcionar información para el establecimiento de indicadores de valor agregado, tanto de la educación media como de la educación superior.
F. Servir como fuente de información para la construcción de indicadores de calidad de la educación, así como para el ejercicio de la inspección y vigilancia del servicio público educativo.
G. Proporcionar información a los establecimientos educativos que ofrecen educación media para el ejercicio de la autoevaluación y para que realicen la consolidación o reorientación de sus prácticas pedagógicas.
H. Ofrecer información que sirva como referente estratégico para el establecimiento de políticas educativas nacionales, territoriales e institucionales.

El examen de Estado de la Educación Media está compuesto por pruebas, y su número y componentes son determinados por el ICFES mediante un acuerdo de su junta. Además, el

calendario de aplicación será determinado por el ICFES de acuerdo con el reporte sobre la población que cumpla el requisito para presentar el Examen establecido en este decreto (Artículo 2°. Estructura y organización). Los resultados obtenidos en el Examen de Estado de la educación media tendrán una vigencia indefinida (Artículo 7°. Vigencia de los resultados). Tiene un carácter periódico el cual posibilita “valorar cuales han sido los avances en un determinado lapso y establecer el impacto de programas y acciones específicas de mejoramiento” (Ministerio de Educación Nacional, 2016).

DESCRIPCIÓN DEL PROBLEMA

DELIMITACIÓN DEL PROBLEMA

La base de datos utilizada, “Saber 11, 2019”, contiene 82 atributos, de los cuales 63 son características socioeconómicas de los estudiantes, en diferentes instituciones educativas del país. Para construir el modelo predictivo, se emplearon 36 variables categóricas y numéricas con el objetivo de predecir el puntaje global dentro del mismo modelo. El puntaje global es el resultado de la sumatoria de las 5 pruebas evaluadas por la prueba Saber 11 (Lectura Crítica, Matemáticas, Competencias Ciudadanas, Ciencias Naturales e Inglés). Se incorporaron la mayor cantidad de atributos posibles para mejorar la capacidad predictiva del modelo en relación con el desempeño de los estudiantes en el Examen de Estado. Esto con el propósito de realizar un análisis más profundo de los resultados del aprendizaje de los estudiantes.

Por lo tanto, para entender el puntaje global, es necesario saber en qué consisten las cinco pruebas evaluadas por el examen Saber 11. El ICFES se interesa por evaluar un número determinado de competencias en cada una de las cinco pruebas, que son primordiales para el desarrollo académico de los estudiantes y el desarrollo de competencias profesionales o especificas (Demarchi, 2020). Los saberes que evalúa el ICFES, que serán parte integral en el foco del análisis, son: Lenguaje o Lectura Crítica, Matemáticas, Competencias Ciudadanas, Ciencias Naturales e Inglés. Una breve descripción aparece a continuación, en la Tabla 1:

Tabla 1

Competencias Pruebas Saber 11

Lectura Crítica	Evalúa las competencias necesarias para comprender, interpretar y evaluar textos que pueden encontrarse en la vida cotidiana y en los ámbitos académicos no especializados (Instituto Colombiano para la Evaluación de la Educación (ICFES), 2021, p.27). Desde esta competencia se espera que el estudiante cuente con las capacidades para tomar posturas críticas frente a esta clase de textos.

Matemáticas	Evalúa competencias para enfrentar situaciones que pueden resolverse con el uso de algunas herramientas matemáticas. En la prueba se evalúan tres competencias que recogen los elementos centrales de los procesos que se describen en los estándares básicos de competencias: interpretación y representación; formulación y ejecución; y argumentación (Instituto Colombiano para la Evaluación de la Educación (ICFES), 2021, p.33).
Competencias Ciudadanes	Evalúa los conocimientos y habilidades que le permiten al estudiante comprender el mundo social y establecer esa comprensión como referente para su ejercicio como ciudadanos. Asimismo, se evalúa la habilidad para analizar distintos eventos, argumentos, posturas, conceptos, modelos, dimensiones y contextos, así como la capacidad para reflexionar y emitir juicios críticos sobre estos (Instituto Colombiano para la Evaluación de la Educación (ICFES), 2021, p.41).
Ciencias Naturales	Evalúa la capacidad de comprender y usar nociones, conceptos, y teorías de las ciencias naturales en la solución de problemas. De igual manera, evalúa la habilidad de explicar fenómenos de la naturaleza basado en observaciones, patrones y conceptos propios del conocimiento científico (Instituto Colombiano para la Evaluación de la Educación (ICFES), 2021, p.47).
Inglés	Evalúa la competencia para comunicarse efectivamente en inglés, teniendo en cuenta el Marco Común Europeo, que clasifica a los evaluados en 5 niveles de desempeño: Pre-A1, A1, A2, B1, B2 (Instituto Colombiano para la Evaluación de la Educación (ICFES), 2021).

El **puntaje global,** que es el atributo objetivo que se evaluó, es numérico el cual refleja el puntaje total obtenido por el estudiante en la prueba.

PROBLEMA Y PREGUNTA DE INVESTIGACIÓN

Teniendo en cuenta lo mencionado anteriormente sobre las pruebas de Estado, uno de los objetivos principales de las pruebas Saber 11 es proporcionar a las instituciones educativas información pertinente sobre las competencias de los aspirantes a ingresar a programas de educación superior, así como sobre las de quienes son admitidos (Ministerio de Educación Nacional, 2016). Entrar a la educación superior, ya sea universitaria, técnica o tecnológica, tiene un impacto en el individuo, ya que le ayuda a desarrollar competencias específicas que le

permite a la persona cambiar su contexto social y así generar **movilidad social**. Además, son un indicador de qué tan eficaz es el sistema educativo.

Por otro lado, para que el individuo tenga la capacidad de resolver distintos problemas en diferentes contextos de la vida cotidiana, de una manera eficiente y ética, hay ciertos saberes y aptitudes fundamentales que son esenciales desarrollar. De ahí la importancia de ver qué tanto el estudiante se está apropiando de estas competencias (tales como el razonamiento numérico, la escritura, la lectura, la inferencia, el relacionamiento con otros, etc), dentro de su entorno educativo, para la mejor toma de decisiones. Las pruebas estandarizadas, de esta forma, son un indicador que permite observar si el estudiante, próximo a graduarse, tiene las habilidades necesarias para afrontar ciertos desafíos dentro de la sociedad.

Siguiendo este orden de ideas, es imperativo ver si el contexto del estudiante le está permitiendo adquirir las competencias que va a necesitar para desenvolverse correctamente dentro de la sociedad y relacionarse con otros. Con base en lo anterior, lo que se quiere responder en este trabajo es la **siguiente pregunta**: **¿Cuáles son las variables socioeconómicas que tienen o presentan una mayor influencia en el desempeño de los estudiantes de la Educación Media en las pruebas de Estado, ICFES, 2019 utilizando redes neuronales como técnica o metodología de análisis?** Para dar respuesta a esta pregunta, se implementó un modelo predictivo de aprendizaje de máquinas supervisado. Consecuentemente, se realizó un análisis de importancia de atributos, y finalmente se observaron cuáles variables socioeconómicas inciden en el resultado de los estudiantes en la prueba SABER 11 para el segundo semestre del año 2019.

OBJETIVO GENERAL

El principal objetivo del presente estudio es **identificar las variables socioeconómicas que inciden en el desempeño de los estudiantes de bachillerato en las pruebas Saber 11 usando Redes Neuronales,** tomando como referencia la información recolectada en el segundo semestre del año 2019.

Para llevar a cabo este objetivo, se proponen los siguientes **tres objetivos específicos:**

1. Crear una red neuronal en PyTorch que pueda identificar la importancia de cada atributo para el conjunto de datos Saber_11_2019, mediante técnicas de análisis de importancia de atributos.
2. Evaluar el comportamiento de la red neuronal, mirando el RMSE (*Root Mean Squared Error*) y el MAE (*Mean Absolute Error*). También se observará el comportamiento de las pérdidas de validación y entrenamiento conjuntamente para evaluar el modelo.
3. Identificar qué variables socioeconómicas son las que más importan en la tarea predictiva del modelo, por medio de un PFI (*Permutation Feature Importance*).

METODOLOGÍA

Este estudio utilizó una red neuronal en Pytorch para crear un modelo predictivo, con algoritmos de aprendizaje de máquinas supervisado. La idea es que esta red neuronal entrene las variables del cuestionario socioeconómico de los estudiantes, aplicado durante las pruebas Saber 11, para predecir el desempeño de un estudiante de bachillerato, reflejado a través de su puntaje global en Examen de Estado. Por este motivo, se quiso ver cómo inciden las diferentes variables socioeconómicas a la hora de realizar un ejercicio de predicción. Estas variables incluyen: Sexo del estudiante, área donde se ubica el colegio del estudiante, educación de padres, estrato socioeconómico, jornada del colegio en el que estudia el estudiante, si tienen acceso a internet y profesión de los padres.

RECOLECCIÓN DE LOS DATOS

En cada una de las pruebas Saber incluyen cuestionarios que permiten recolectar información acerca del entorno de los estudiantes, tales como antecedentes escolares, competencias socioemocionales y características socioeconómicas y culturales (Instituto Colombiano para la Evaluación de la Educación (ICFES) y Ministerio de Educación Nacional, 2019). Según el Instituto Colombiano para la Evaluación de la Educación (ICFES) y Ministerio de Educación Nacional (2019):

> Estas variables, conocidas como factores asociados al aprendizaje, son relevantes en el estudio de la calidad de la educación, ya que tienen influencia sobre el logro educativo. En particular, se ha documentado ampliamente que el nivel socioeconómico tiene una alta incidencia sobre el desempeño académico de los estudiantes. (p.2)

El cuestionario socioeconómico, que se responde durante la prueba:

> Permite obtener información sobre los procesos de enseñanza y aprendizaje de los estudiantes, así como como explicar los resultados del examen. La base obtenida indaga por características del núcleo familiar (composición, situación laboral y educativo), características del hogar (dotación de bienes dentro de la vivienda, estrato socioeconómico, disponibilidad de conexión a internet y servicio de televisión por cable) y el tiempo dedicado por la familia al entrenamiento. La información recopilada en el cuestionario socioeconómico tiene propósitos académicos, de investigación y de política pública, por tal razón, las respuestas dadas son de carácter confidencial y no afectan el resultado de los evaluados. (Instituto Colombiano para la Evaluación de la Educación (ICFES), 2021, p.12)

Estos son los datos con los que se trabajaron en este estudio.

DESCRIPCIÓN Y VISUALIZACIÓN DE LOS DATOS

La base trabajada se obtuvo de datos abiertos por parte del gobierno nacional. De aquí se consiguieron las estadísticas complementarias para explicar por qué un estudiante no obtiene buenos resultados las pruebas ICFES para ingresar a la educación superior, o para el desarrollo de su proyecto de vida (Ministerio de Educación Nacional, 2016). La base de datos usada, la del segundo semestre del 2019, fue la más completa de disponibles hasta entonces. Esta base de datos cuenta con 546,212 observaciones y 82 atributos, con las variables socioeconómicas de cada estudiante. De los 82 atributos 53 son categóricas y los 29 sobrantes son variables numéricas. Para evitar que haya información duplicada, se utilizó en pandas la función de *duplicated(),* la cual determina si hay alguna observación duplicada en la base estudiada. No hay observaciones repetidas.

El objetivo de esta parte es hacer ejercicios de Visualización y Estadística Descriptiva, para obtener una representación visual de los patrones y tendencias de los datos, y así tener una idea sobre el comportamiento de los atributos específicos que se utilizaron durante este estudio. Por este motivo, no se realizaron pruebas de significancia estadística. En la Tabla 2, a continuación, muestra un resumen de las variables socioeconómicas que más influyen en el desempeño de los estudiantes en la Prueba de Estado, con base en los estudios de los autores mencionados en la sección Marco Teórico:

Tabla 2

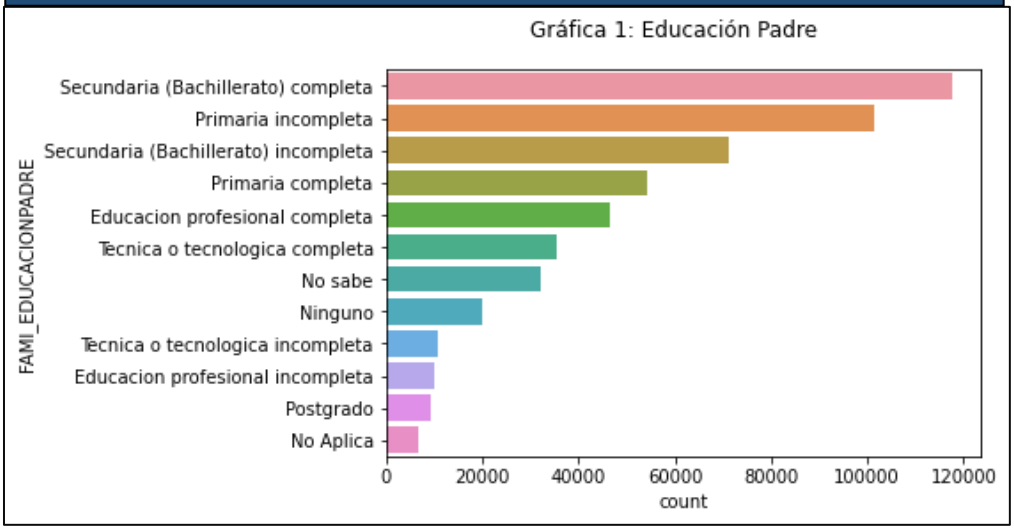
Referencias Socioeconómicas

Blanco (2015)	Timarán-Pereira et al. (2019)	Ávila et al. (2021)	García-González y Skrita (2019)
- Estrato - Ubicación de colegio - Educación padres	- Estrato socioeconómico - Jornada de estudio en la mañana o completa - Índice TIC regular - Sexo de los estudiantes	- La educación de los padres - El estrato - La edad - El número de libros que tiene la familia - Tiempo de lectura diario - Jornada del colegio	- Nivel educativo de la madre - Estrato socioeconómico de la vivienda - Número de libros - Nivel educativo del padre - Poseer computador en la vivienda

La educación de los padres es una de las variables que más se repite entre los autores. En la base de datos de este estudio, Saber_11_2019, de los padres de los estudiantes que presentaron las pruebas de Estado Saber 11 en el segundo semestre del año 2019, 117,488 tienen la secundaria (Bachillerato) completa, lo cual equivale a un 21.5% de la data trabajada. Le siguen padres con primaria incompleta que son 101,387 con un 18.5%. En tercer lugar, se encuentran los padres con secundaria (Bachillerato) incompleta, lo cuales son 71,295 padres

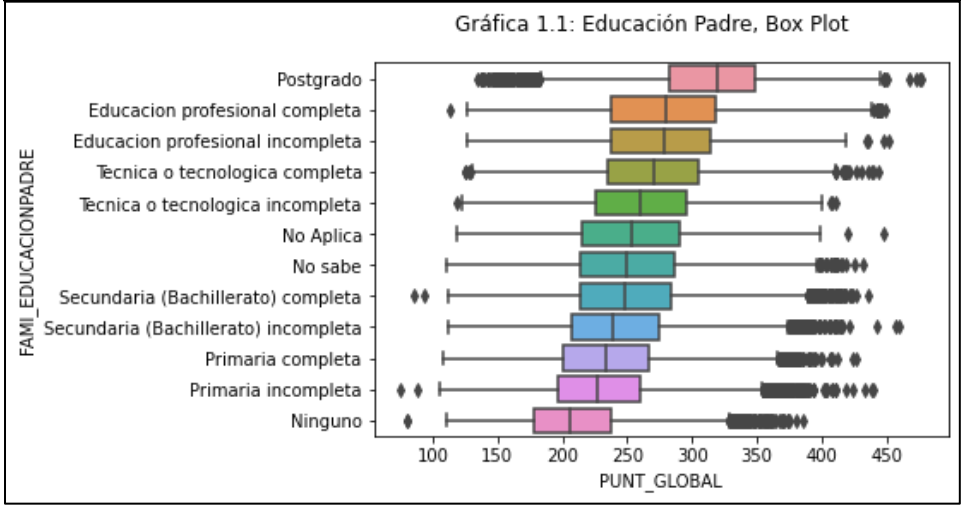
de toda la data, lo cual equivale a un 13.05%. También se puede notar que los padres con una educación profesional completa o postgrado tienen una baja representación en esta base con 8.4% y un 1.7% respectivamente. Lo anterior puede verse evidenciado en la **Gráfica 1**.

Gráfica 1: Educación del Padre



Ahora, para ver una relación más directa entre la profesión del padre con los resultados de los estudiantes en las pruebas Saber 11, es preciso observar la **Gráfica 1.1**. Los resultados arrojan ciertos datos previsibles: los estudiantes que tienen padres con postgrado se desempeñan mejor en las pruebas Saber 11, obteniendo un puntaje global promedio de 311.81. Le siguen padres con educación profesional completa, cuyos hijos obtienen un puntaje promedio 277.45 y en tercer lugar están los padres con educación profesional incompleta (los estudiantes con estos padres tienen un Puntaje Promedio de 274.75).

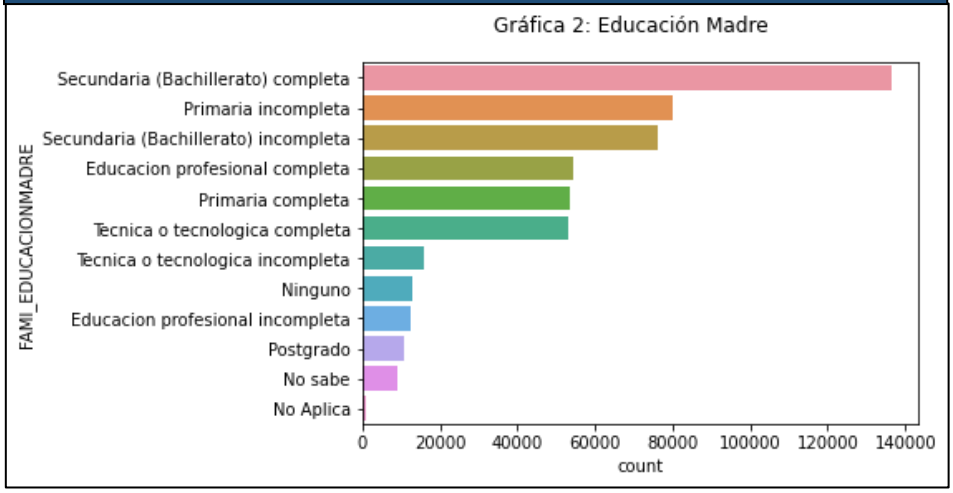
Gráfica 1.1: Educación del Padre, Box Plot



Los estudiantes con el más bajo desempeño son aquellos que tienen padres con primaria completa e incompleta con un puntaje promedio de 235.54 y 230.49 respectivamente. Si los padres no tienen ninguna educación, el estudiante obtiene un puntaje promedio de 211.36.

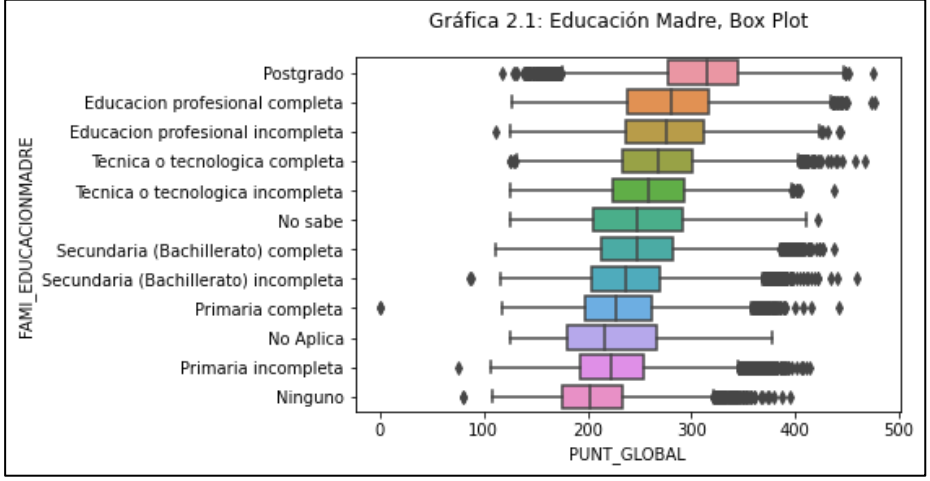
Pasando a la educación de la madre **Gráfica 2,** se ven resultados iguales a los del padre: dentro de la base de datos trabajada en este estudio, de las madres de los estudiantes que presentaron las pruebas Saber 2019-II, 136,286 tienen educación Secundaria (Bachillerato) completa, el cual equivale al 24.9%, 80,062 de las madres tienen la primeria completa (14.6%), y 76,353 tienen Secundaria (Bachillerato) incompleto (13.9%). Por otro lado, las madres con postgrado son las que menos representatividad tienen en los datos, siendo apenas 10,805 (lo cual equivale al 1.9%) y las madres con Educación profesional incompleta siendo 12,384 (2.26%).

Gráfica 2: Educación Madre



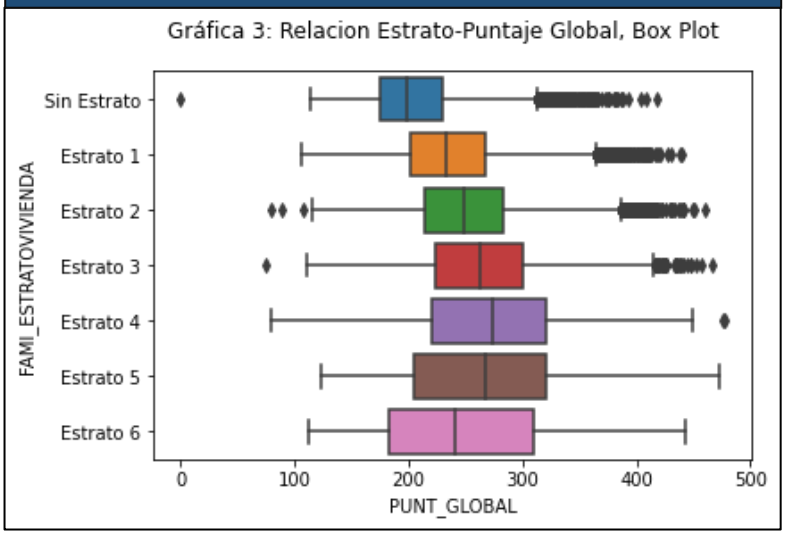
En la **Gráfica 2.1**, se puede notar la tendencia que se halló con respecto a la educación del padre: Los estudiantes cuyas madres tienen estudios en postgrado, obtuvieron el mejor desempeño, con un puntaje global promedio de 307.67, seguido por las madres con educación profesional completa (277.33), y seguido por educación profesional incompleta (273.74). Si la madre no tiene educación, los estudiantes sacan en promedio 207.8 en el puntaje global de las pruebas de Estado, y si la madre tiene una primaria incompleta, obtienen un resultado promedio de 225.83. Se observa aquí que la educación de los padres sí puede tener un impacto en el desempeño de los estudiantes en las pruebas de Estado Saber 11**.**

Gráfica 2.1: Educación Madre, Box Plot



Con respecto al estrato socioeconómico, se pueden observar dos aspectos interesantes. En primera instancia, se nota que los estudiantes que tuvieron un mejor desempeño en las pruebas Saber 11 son de estrato 4, con un puntaje promedio de 270. Le siguen los estudiantes de estrato 5 con promedio en el puntaje global de 264.73, seguido por los de estrato 3 con un promedio de 261,94. Los estudiantes de estrato 1 y sin estrato son los que peor desempeño tienen, con un desempeño del 236 y 206.43 respectivamente. Esto se puede observar en la **Gráfica 3**.

Gráfica 3: Relación Estrato-Puntaje, Box Plot



En un segundo aspecto, es importante notar que el desempeño de los estudiantes de cada estrato cambia dependiendo la naturaleza del establecimiento educativo. Por ejemplo, los estudiantes de estrato 4 que obtuvieron un puntaje más alto en las pruebas Saber 11, se encuentran en colegios no oficiales, y tuvieron un puntaje promedio de 296, mientras que los matriculados en instituciones oficiales obtuvieron un puntaje promedio de 235. Y esto se cumple para todos los estratos, a excepción del estrato 1. Para todos los demás casos, en la medida que va aumentando el estrato socioeconómico del estudiante, la diferencia entre los puntajes de las instituciones oficiales y no oficiales va siendo mayor, con el puntaje del colegio no oficial siempre siendo más grande. Esto se puede observar en la tabla 3.

Por lo tanto, observando la tabla 3, estar matriculados en instituciones oficiales y no oficinales puede tener una gran incidencia en el desempeño de los estudiantes.

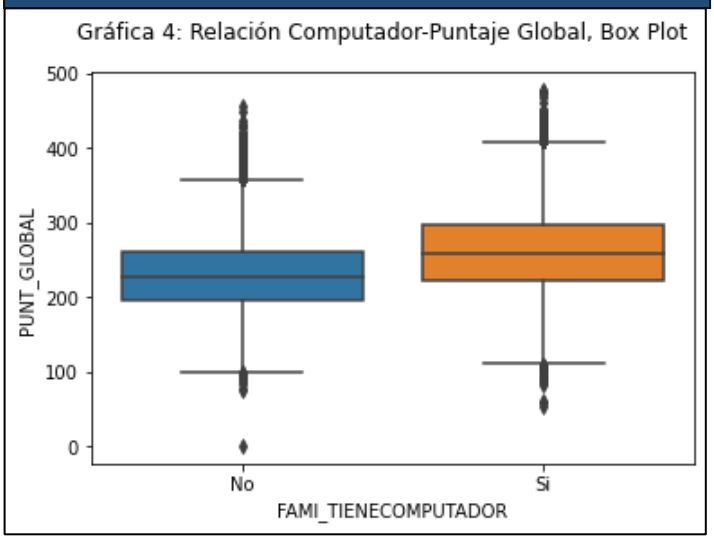
Tabla 3

Promedio Puntaje según Estrato y Naturaleza del Establecimiento

FAMI_ESTRATOVIVIENDA	NO OFICIAL	OFICIAL
Estrato 1	231.9	236.5
Estrato 2	255.8	248.3
Estrato 3	276.1	253.0
Estrato 4	294.0	235.0
Estrato 5	296.1	216.1
Estrato 6	293.1	201.4
Sin Estrato	214.3	205.5

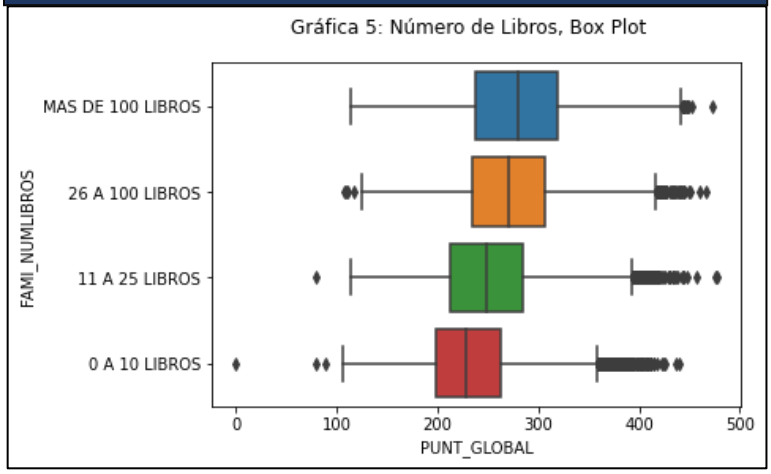
Por último, se puede ver la influencia de tres variables adicionales en el desempeño de los estudiantes: la disponibilidad de un computador en el hogar, la cantidad de libros y la jornada de la sede educativa. En la **Gráfica 4**. se observa cómo influye tener un computador en el hogar en el desempeño de los estudiantes en las pruebas Saber 11. Según el Gráfico que describe la relación entre poseer computador y los resultados de las Pruebas de Estado, se nota explícitamente que tener computador (García-González y Skrita, 2019) aumenta el puntaje promedio: para los estudiantes que tienen computador es de 259.83, mientras los que no tienen computador en su hogar obtienen un promedio de 230.10. Adicionalmente el mínimo de los estudiantes con computador (54) es más alto que los estudiantes que no tienen (0), y el máximo también es mayor (477 en comparación al 457).

Gráfica 4: Relación Computador-Puntaje



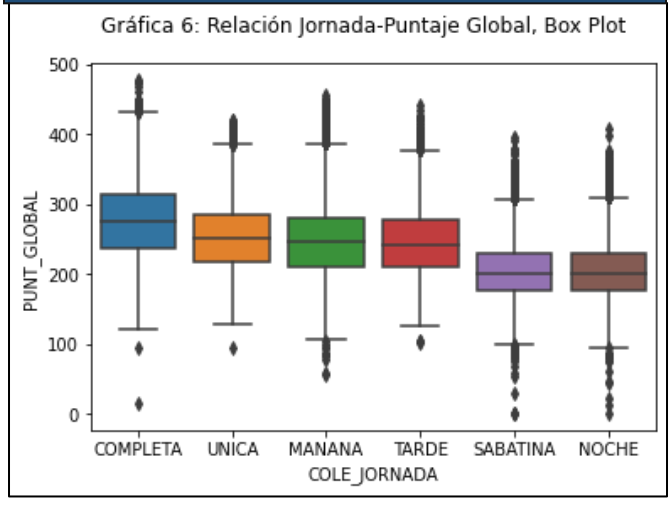
Por otro lado, el número de libros que el estudiante tiene el hogar también influye en el desempeño del estudiante (Ávila et al., 2021; García-González y Skrita, 2019). Como se observa en la **Gráfica 5**, tener más cantidad de libros aumenta el desempeño de los estudiantes en las pruebas. Se puede ver, en orden ascendente, cómo cambia la media de los resultados Globales con respecto a la cantidad de libros que el estudiante posee en su hogar: los estudiantes que tienen 0 a 10 libros obtienen un puntaje promedio de 231,63; los estudiantes con 11 a 25 libros tienen un resultado promedio de 249.65; le siguen los estudiantes que tienen entre 26 y 100 libros, quienes alcanzan un puntaje global promedio de 270.10; finalmente los que mejor desempeño promedio tienen en las pruebas Saber 11, son los estudiantes que tienen 100 o más libros, con un desempeño promedio en el puntaje Global de 276.49.

Gráfica 5: Número de libros, Box Plot



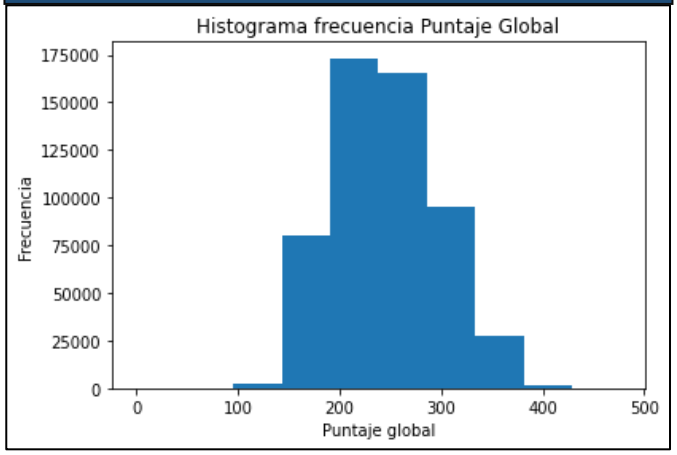
En la **Gráfica 6**, se observa cómo influye la jornada del establecimiento educativo sobre el desempeño de los estudiantes en las pruebas Saber 11 (Ávila et al., 2021; Timarán-Pereira et al., 2019). Las instituciones educativas con jornada completa obtienen, en promedio, los mejores resultados en el puntaje global de las pruebas, con un desempeño promedio de 274.60. Le sigue los establecimientos educativos con jornada única, los cuales tienen un puntaje promedio de 251.92. La diferencia entre la jornada única y completa es que en la primera se busca concentrar todas las horas de estudio en un solo turno diario para el desarrollo de las áreas obligatorias y fundamentales, mientras que la segunda busca ampliar el tiempo de la jornada escolar con el propósito de brindarle más oportunidades a los estudiantes en el marco de la formación integral (Portal Educativo Red Académica, 2022). Le siguen la jornada de la mañana (con un puntaje promedio de 246.83), tarde (244.68), sabatina (205.22), y finalmente, los que obtienen los peores resultados son los estudiantes que van al colegio por la noche, con un desempeño promedio de 204.95.

Gráfica 6: Relación Jornada-Puntaje Global



Finalmente, en la **Gráfica 7**, se observa la distribución de frecuencias del puntaje global de los estudiantes de undécimo para el segundo semestre del año 2019:

Gráfica 7: Histograma Puntaje Global



SELECCIÓN DE LOS DATOS

Se realizó una revisión cuidadosa de los atributos que fueron utilizados para la creación del modelo predictivo. Un criterio importante fue la cantidad de valores únicos en cada atributo (ya sea porque hay muchos o casi no tiene) y su pronosticada importancia en la variable respuesta (Tabla 2). En la Tabla 8 (anexo), se encuentran en detalle los atributos preliminares, que posteriormente se utilizaron para el análisis de la incidencia en el desempeño de los estudiantes en las pruebas Saber 11-2019. Estas incluyen 35 variables socioeconómicas como variables independiente o predictoras, donde 32 de ellas son categóricas y 2 son numéricas. Se estableció como variable respuesta el Puntaje Global del estudiante, que es un atributo continuo numérico. Dentro de las variables explicativas, se incluyeron: género del estudiante, grupo étnico, nivel educativo más alto alcanzado por el padre, jornada de la sede y estrato socioeconómico de la vivienda.

CORRELACIÓN ENTRE LAS VARIABLES

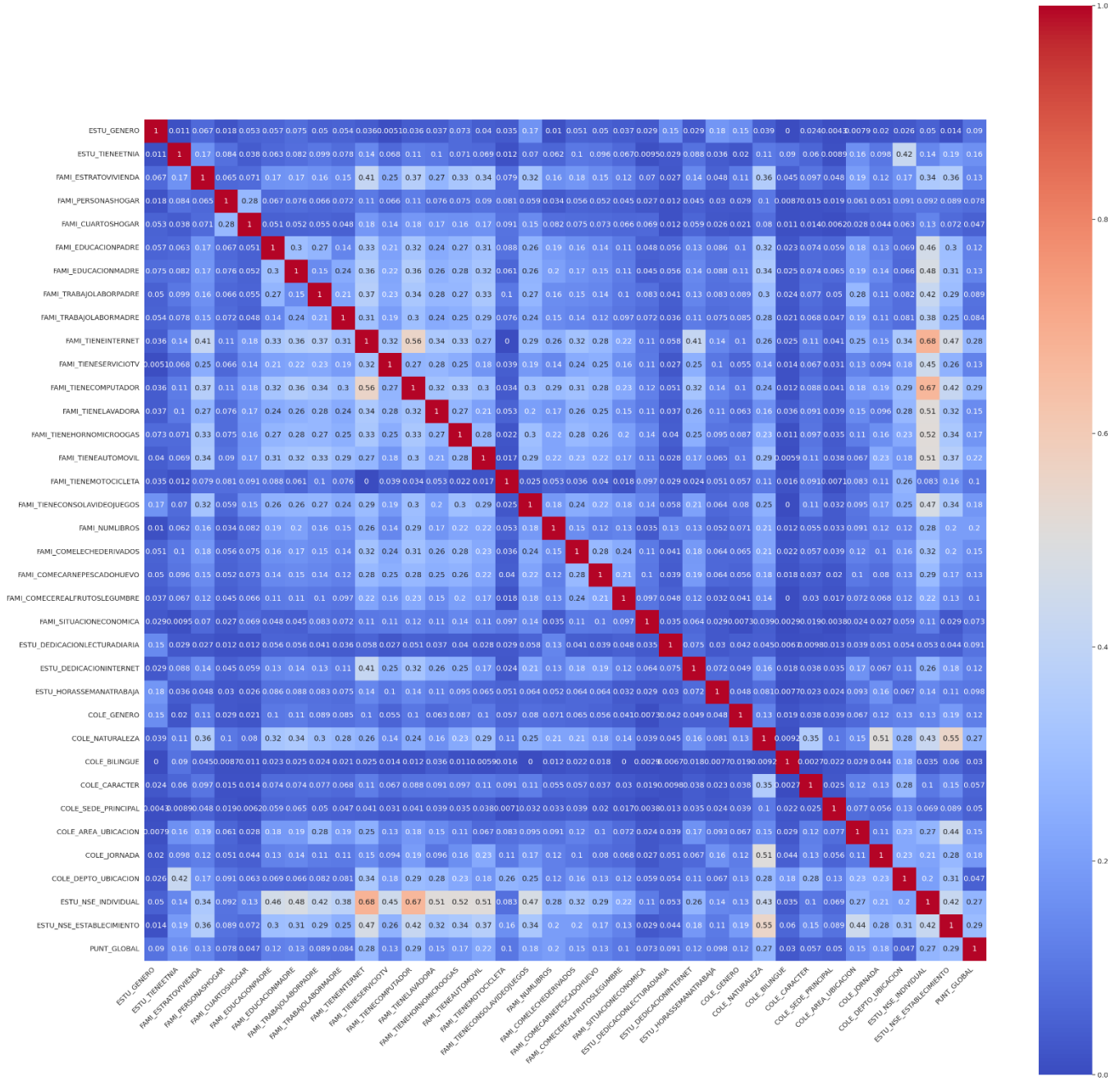
En el momento de trabajar con redes neuronales, usando Deep Learning en Pytroch, no es estrictamente necesario hacer pruebas de correlación entre las variables de entrada (input features). Una de la ventajas de utilizar modelos de Deep Learning, como las redes neuronales, es que tienen la habilidad de automáticamente aprender patrones y relaciones complejas en la data, incluso capturando las relaciones entre las variables correlacionadas. Sin embargo, puede, aun así, ser beneficioso analizar la correlación entre las variables de entrada, como parte del proceso de análisis exploratorio de los datos. Entender la relación entre las variables puede dar indicios de cómo se relacionan los datos, y ayudar a tomar decisiones informadas durante es el proceso de selección de variables.

Un análisis de correlación puede ayudar también a identificar variables altamente correlacionadas, que puedan no contribuir significativamente a la capacidad predictiva del modelo. Se cogieron las 36 variables seleccionas y se realizó una matriz de correlación, con un mapa de calor, para observar su relación y ver si cuál es la correlación que existe entre ellas. La **Gráfica 8** ayuda a entender la relación entre las variables socioeconómicas que se usaron en modelo predictivo. En el mapa de calor de observa que ninguna de la variables está altamente correlacionada: la máxima correlación se da entre la variable ESTU_NSE_INDIVIDUAL (Nivel Socioeconómico del Evaluado) y FAMI_TIENEINTERNET (si el hogar del evaluado cuenta con servicio o conexión a internet) con un nivel de correlación de 0.68, la cual se consideraría como una correlación moderadamente alta. La baja correlación entre las variables estudiadas puede beneficiar el análisis ya que indica que las variables son independientes y no presentan multicolinealidad.

Para crear el mapa de calor en la **Gráfica 8**, se trataron los variables categóricas. Se utilizó la función *factorize()* para asignar un valor numérico para cada categoría en una columna. El *Dataframe* resultante tendrá la misma estructura, pero con los valores categóricos reemplazados por etiquetas numéricas. Con este nuevo Dataframe se calculó la matriz de correlación: la función cramers_v() calcula la correlación V de Cramér entre dos variables en función de su tabla de contingencia. La V de Cramér es una medida de asociación entre dos variables categóricas. Se basa en la estadística chi-cuadrado y se utiliza para cuantificar la fuerza de la relación entre las variables.

Por otro lado, la estadística de chi-cuadrado, es una medida utilizada para determinar si existe una relación significativa entre dos variables categóricas. En el contexto de la correlación de V de Cramér, se creó una tabla de contingencia para mostrar la distribución de frecuencias de las ocurrencias conjuntas de dos variables. Con la tabla de contingencia, se calcula el chi cuadrado usando chi2_contingency() de scipy.stats. Por último, se obtiene el cuadrado φ^2, que normaliza la estadística de chi cuadrado entre 0 y 1, para luego calcular la Gráfica *V de Cramér,* que se observa a continuación:

Gráfica 8: Cramér V Correlation Matrix, Variables Socioeconómicas



RED NEURONAL CON DEEP LEARNING

Una Red Neuronal usando Deep Learning usa grandes cantidades de datos para aproximar funciones complejas, cuyas entradas (inputs) y salidas (outputs) están muy separadas, como una imagen de entrada y, como salida, una línea de texto describiendo la entrada; o un guion escrito como entrada y una voz que suena natural recitando el guion como salida (Stevens et al., 2020). En este trabajo, se observó cómo una red neuronal usando Deep Learning toma

características de entrada de una base de datos tabular, que son tanto categóricas como numéricas, para predecir una salida, que en este caso es una variable objetivo-continua. Se explicó cada componente de la red neuronal en detalle, incluido los *embeddings* para representar los atributos categóricos en la base, capas completamente conectadas con normalizadas, función de activación ReLU, y *dropouts* como método de regularización para evitar *overfitting*, el uso del error cuadrático medio y el optimizador Adam para entrenar el modelo.

Ahora, el modelo fue entrenado usando técnicas de aprendizaje de máquinas supervisado. Para poder encontrar un modelo que mejor se ajuste a nuestros datos, se usó ***Deep learning***, y más específicamente, para su implementación, se utilizaron herramientas dentro de Python como PyTorch (Stevens et al., 2020) para entrenar y validar el modelo.

Importar librerías.

Primero, se importaron las siguientes paquetes y librerías para trabajar con la base de datos y crear así la red neuronal del modelo de regresión usando PyTorch y scikit-learn:

- Pandas: es una librería en Python usada para la manipulación y el análisis de los datos.
- Numpy: es una librería en Python para la computación numérica.
- Torch: es una librería en Python usadas para crear y entrenar redes neuronales.
- Skelearn.preprocessing: es una librería en scikit-learn usada para el preprocesamiento de los datos y escalar atributos.
- Torch.utils.data: es una librería en PyTorch usada para trabajar base de datos y data loaders.
- Sklearn.model_selection: es una librería en scikit-learn usada para la selección y evaluación del modelo.
- Sklearn.metrics: es una librería en scikit-learn usada para evaluar el rendimiento del modelo.

Juntas, estas librerías proporcionan un poderoso conjunto de herramientas para construir y evaluar modelos de Aprendizaje de Maquinas. Al importar estas librerías, el código configura una infraestructura necesaria para construir y evaluar el modelo de regresión usando PyTorch y scikit-learn. Específicamente, este código importa las funciones necesarias para manejar los datos, escalar las características, construir y entrenar las redes neuronales, y evaluar el rendimiento del modelo. Antes de subir la base de datos a GoogleColab, se hizo una limpieza, quitando los caracteres especiales, doble espacio, tildes, y demás.

Preprocesamiento de los datos.

Después de cargar los datos, fue necesario preprocesar los datos, y esto regularmente se hace por medio de un *imputer* para quitar los datos nulos y un *one hot encoder* para binarizar las variables categóricas. Para los datos faltantes, se omitieron usando dropna(): aunque los datos omitidos puedan tener información relevante para la predicción del modelo, se van a descartar

por simplicidad del modelo y mejorar el entrenamiento de los datos para la tarea predictiva de la red neuronal. Esto dejó **390,254 observaciones** para trabajar.

Para la red neuronal creada en este trabajo, fue necesario hacer un preprocesamiento distinto de los datos, tanto los categóricos como los numéricos. Como se vio anteriormente, la mayor dificultad se centró en cómo procesar estos datos categóricos ya que son la gran mayoría de atributos en nuestra base, y además existen distintas herramientas en Pytorch o Sklearn para tratar dichas variables. Al efectuar esto correctamente, se aseguró una implementación más fácil en un ambiente de producción.

Para hacer un buen entrenamiento del modelo, y de esta forma predecir eficientemente la variable respuesta, se podría, primero, tomar todas nuestras variables categóricas y binarizarlas por medio de un *One hot encoder*. Sin embargo, pueden existir ciertas complicaciones usando el *one-hot encoding* que se tratarán en la siguiente sección.

Alternativa al *One-hot encoding*: *embedding*.

Ya que la mayoría de los atributos que se utilizaron son categóricos, tal y como se describió en la sección de “Selección de datos”, el *one-hot encoding* puede volver la red neuronal más compleja y difícil de entrenar ya que cada categoría dentro de los atributos categóricos se representa como un atributo binario separado. Se puede observar esto en el siguiente ejemplo: si se tiene un atributo categórico con tres clases (A, B y C), se crean tres atributos binarios (A, B y C), con un valor de 1 en una de ellas y para cada punto de datos, dependiendo de la categoría a la que pertenezca. Esto puede crear una gran cantidad de variables de entrada (*input*), especialmente si hay muchas categorías en el atributo, tal y como es el caso de este estudio, y esto, consecuentemente, puede hacer que la red neuronal sea más compleja y difícil de entrenar.

Por este motivo, se utilizaron *embeddings*, donde cada categoría está representada como un vector continuo de una dimensionalidad fija o un conjunto de valores numéricos. La red neuronal aprende de estos vectores del *embedding* durante el entrenamiento, y son optimizadas para capturar la relación entre las categorías y la variable respuesta. Esto es más eficiente que el *one hot encoding*, ya que la dimensionalidad de los vectores del *embedding* son mucho más pequeños que el número de categorías en el atributo, y, a la vez, permite a la red neuronal capturar relaciones no lineales entre las variables categóricas y la variable respuesta. En general, usar una capa de embedding puede ser más poderoso que en el *One-hot encoder* porque permite a la red neuronal aprender más relaciones complejas entre las variables categóricas y la variable respuesta.

Sin embargo, también requiere más datos de entrenamiento para aprender buenos *embeddings*, y puede ser más complejo que implementar *One-hot encoding. One-hot encoding* puede ser una buena opción cuando el número de categorías es pequeño o cuando los atributos categóricos no predicen bien la variable respuesta, lo cual no es el caso para esta investigación.

Para realizar el respetivo *embedding* a las variables categóricas, se realizaron los siguientes pasos como preprocesamiento:

1. Las variables categóricas fueron identificadas usando el método “select.dtypes()” para incluir todas las variables con los parámetros “object”. Estas columnas fueron almacenadas en una variable llamada “cat_cols”.
2. Se creó lo que se domina como “*mappings*” para cada atributo categórico usando un diccionario que asigna cada valor único del atributo categórico a un índice numérico. Estas asignaciones se almacenaron en un diccionario llamado “cat_maps”.
3. Luego, se llamó la función map() en cada columna categórica en los atributos del DataFrame para aplicar el mapeo correspondiente a la columna. Esto reemplaza los valores categóricos en la columna con sus valores de índice numérico correspondiente.
4. Por último, se crearon los *embeddings* para las variables categóricas. Los *embeddings* son una forma de representar las variables categóricas como vectores continuos en un espacio de alta dimensión. Esto se hace usando la función nn.Embedding de la librería de PyTorch.

Los *embeddings* aprenden durante el proceso de entrenamiento de la red neuronal, por lo que no deben predefinirse ni entrenarse previamente. Los embeddings se iniciarán y actualizarán durante el *backpropagation* del proceso de entrenamiento.

Matemáticamente, los *mappings* se pueden representar de esta manera:

Sea X una variable categórica, y sea x_i un valor único de X. Se creó un *mapping* m que mapea x_i en un número entero y_i.

$$\text{M: } x_i \rightarrow y_i$$

Los *embeddings*, se representan matemáticamente de la siguiente forma: Sea X una variable categórica, y sea x_i un valor único de X. Se creó un *embedding* de matriz E que mapea x_i en un vector continuo e_i con longitud K.

$$\text{E: } x_i \rightarrow e_i$$

En general, aplicar *mappings* y e*mbeddings* a las variables categóricas es un paso importante en la red neuronal porque permite representar datos cualitativos de una manera que se puedan usar como atributos de entrada (*inputs*) para el modelo. Al mapear variables categóricas a valores numéricos y crear *embeddings*, se pueden capturar relaciones entre variables categóricas y mejorar el desempeño de la red neuronal.

Normalización de los atributos numéricos.

Por otro lado, es también necesario hacer algún tipo de procesamiento con las variables numéricas que contiene la base de datos. Por lo general, es recomendado usar un Escalador

Estándar (*Standard Scaler*), o alguna forma de normalizar los atributos numéricos para entrenar modelos de aprendizaje de máquina. Esta forma de estandarización ayuda a llevar los atributos a una escala similar, lo que puede mejorar el rendimiento de ciertos algoritmos, en particular aquellos que se basan en medidas de distancia y regularización. Se usó el Escalador Estándar debido a que conserva la distribución original de los datos, es robustos para los datos atípicos, mejora la convergencia de optimización para los algoritmos de *Machine Learning* que usan descenso de gradiente, y funciona bien con datos multivariados.

Los atributos numéricos que fueron escogidos de la base de datos Saber 11-2019, fueron entonces normalizados mediante el método de Z-score. Esto se hace para garantizar que los atributos numéricos tengan una escala similar y no dominen en el entrenamiento del modelo de la red neuronal. Ahora, si se consideran estos atributos como nominales en función de estratos, y no se normalizan mediante el método de Z-score, esto conduce a resultados menos favorables en términos de RMSE (37.8146) Y MAE (30.3681).

Mean() y std() se aplicaron en columnas numéricas para calcular la media y la desviación estándar de cada atributo numérico dentro del DataFrame. Luego, los atributos continuos se normalizaron, restando la media y dividiéndola por la desviación estándar.

Matemáticamente, la estandarización se puede representar de la siguiente manera:

Sea X una característica numérica, y μ y σ la media y la desviación estándar de X, respectivamente. Se puede utilizar X para obtener un nuevo atributo X_{std}, usando la siguiente ecuación:

$$X_{std} = \frac{(X - \mu)}{\sigma}$$

Convertir la Data en PyTorch Tensors.

Es importante transformar las variables categóricas y numéricas en Tensors. Los atributos categóricos se transformaron en PyTorch LongTensor. Un LongTensor es un tensor en PyTorch que contiene valores enteros de 64 bits. Esto fue necesario porque PyTorch requiere que los tensores sean de un tipo específico para trabajar con sus funciones de red neuronal. Al convertir las variables categóricas en LongTensor, se asegura de que puedan usarse en el modelo de PyTorch.

Matemáticamente, se representa un LongTensor T de la siguiente manera:

$$T = [t_1, t_2, \dots, t_n]$$

Donde t_i es un valor entero y n es el número de valores en el tensor.

De manera similar, las variables continuas (X_cont) y la variable objetivo (PUNT_GLOBAL), se transformaron en FloatTensor, que contiene valores de punto flotante de 32 bits. Al convertir las

variables continuas y la variable objetivo en FloatTensor, se asegura de que puedan usarse en el modelo de PyTorch.

Matemáticamente, se representa in FloatTensor T de la siguiente manera:

$$T = [t_1, t_2, \dots, t_n]$$

Donde t_i es un valor de punto flotante y n es el número de valores en el tensor.

División del conjunto de datos.

En este paso, se dividió el conjunto de datos en entrenamiento y validación mediante la función train_test_split() de la biblioteca sklearn. El conjunto de entrenamiento se usa para entrenar el modelo de la red neuronal y el conjunto de validación se usa para monitorear el desempeño del modelo durante el entrenamiento y evitar sobreajuste (*overfitting*). Esto ayuda también a evaluar qué tan bien se generalizará el modelo a datos nuevos y no vistos. Las funciones categóricas se dividen en x_train_cat y x_val_cat, y las funciones numéricas de dividen en x_train_cont y x_val_cont, y la variable objetivo se divide en y_train e y_val.

El parámetro test_size especifica la proporción de los datos que deben usarse para la validación, que es este caso se establece como 20%. El restante 80% se usó como el set de entrenamiento. Finalmente, se utilizó una semilla para garantizar que se incluyan los mismos puntos de datos en ambos conjuntos

La Arquitectura de la Red Neuronal.

Ahora, para implementar este modelo, se tomaron los atributos categóricos y numéricos como entrada (*inputs*), y se usaron dos capas completamente conectadas por 500 y 100 redes neuronales respectivamente para predecir una salida (*output*) continua.

Se usó también un ***forward*** *propagation* para calcular los resultados (*outputs*) del modelo, que para este estudio son los resultados de las pruebas de Estado de los estudiantes de once para el segundo semestre del año 2019. A sí mismo, también se crearon los parámetros para establecer un ***backward*** *propagation* para calcular el error, y de esta manera actualizar los pesos y los sesgos de la función de la red para ver si el modelo está aprendiendo y mejora los resultados y las predicciones del output.

La red neuronal se definió utilizando la clase Pytorch nn.Module y el método _init_ para inicializar las capas de la red neuronal. En este modelo, se usaron los *embeddings* para transformar los atributos categóricos en vectores continuos de baja dimensionalidad, que luego se concatenan con los atributos continuos y pasan a través de las capas completamente conectadas. Luego, aprenden a combinar las representaciones de los atributos categóricos y continuos para producir una salida que predice la variable objetivo. Estas capas son seguidas por la normalización de capas, funciones de activación ReLU y capas *dropout.* Finalmente, se tiene una capa de salida (*output*).

Capa de embeddings para los atributos categóricos.

La capa de self.embeddings se definió mediante la función nn.ModuleList. La función nn.Embedding se usa para crear una capa de embedding para cada atributo categórico. Una capa de embedding asigna a cada valor categórico único a un espacio vectorial continuo que la red neuronal puede procesar más fácilmente. El vector del embedding aprende durante el entrenamiento y se optimiza para capturar las relaciones entre diferentes valores categóricos.

Para cada atributo categórico, se crea una capa de embedding con el número de valores únicos del atributo como dimensión de entrada y una dimensión de embedding fija como la dimensión de salida (*output*). La dimensión del embedding suele ser más pequeña que la cantidad de valores únicos del atributo y, en este caso, se estableció en un mínimo de 100 y la mitad de la cantidad de valores únicos. Esto se hizo para reducir la dimensionalidad de los vectores y evitar que haya sobreajuste.

Las capas de embedding luego se almacenan en una lista llamada *embeddings,* que se usa más adelante para recuperar los embeddings para cada atributo categórico.

Capas completamente conectadas.

Después de crear las capas de embedding, los atributos categóricos y continuos se combinaron mediante capas completamente conectadas para generar una sola salida que predice la variable objetivo (output). El modelo tiene dos capas ocultas completamente conectadas: fc1 y fc2. La capa fc1 toma la entrada concatenada de todos los embeddings y las variables continuas. El tamaño de salida de esta capa es de 500 neuronas. La capa fc2 toma como entrada la salida de dropout1, que tiene tamaño de 500. El tamaño de salida de fc2 es 100. Ambas capas totalmente conectadas realizan una multiplicación de matrices entre el tensor de entrada y una matriz de peso, seguida de la suma de un sesgo. Luego, la salida pasa a través de una función de activación no lineal. La ecuación para una capa completamente conectada se puede escribir de la siguiente manera:

$$y = activation(Wx + b)$$

Donde W es la matriz de peso, “x” es el tensor de entrada, “b” es el término que representa el sesgo, *activation* es la función de activación, e “*y”* es el tensor de salida.

En capas totalmente conectadas, también se insertó el módulo nn.Linear, que representa una transformación lineal del tensor de entrada. En la clase “Model”, nn.Linear es usado en las capas totalmente conectadas (fc1, fc2, output_layer) para realizar una multiplicación de matriz entre el tensor de entrada, y una matriz de peso, seguida de la suma de un sesgo. La salida de la capa nn.Linear luego pasa a través de otras capas de la red.

nn.Linear es una parte importante del modelo ya que proporciona una manera de aprender un mapeo no lineal entre los datos de entrada y salida. Usando varias capas nn.Linear con funciones de activación no lineales entre ellas, el modelo puede aprender asignaciones complejas entre los datos de entrada y salida que serían difíciles de aprender con una única transformación lineal.

Función de activación.

Se necesita una función de activación para determinar si una neurona debería ser activada o no. Esto implica que esta función de activación va a utilizar alguna operación matemática simple para determinar si la variable de entrada de la neurona (*input*), va a ser relevante o no en el proceso de predicción. La habilidad de introducir no-linealidad a una red neuronal artificial y generar un resultado (*output*) a partir de una colección de variables de entrada (*input*), alimentadas a una capa, es el propósito de la función de activación.

La habilidad de introducir no-linealidad a una red neuronal artificial y generar unos resultados (*outputs*) con respecto a una colección de valores de entrada (*input*) alimentadas a una capa es el propósito de esta función de activación. De las diferentes funciones de activación que existen, se usó una función no lineal llamada **ReLU**. Dos funciones de activación ReLU fueron usadas en este modelo. La función de activación ReLU es una función de activación no lineal que aplica la función max(0,x), elemento a elemento al tensor de entrada. Esta función de activación ayuda a introducir no linealidad a la red neuronal, y evita que haya problemas como el de **gradientes de fuga**. La ecuación ReLU es:

$$y = \max(0, x)$$

Donde “x” es en tensor de entrada e “y” el tensor de salida

Normalización de las capas.

Se utilizaron dos capas nn.LayerNorm para normalizar las activaciones de las capas completamente conectadas. La normalización de capas estandariza la salida de una capa restando la media y dividiéndola por la desviación estándar de las activaciones a lo largo de una dimensión especifica. A continuación, la salida normalizada se escala y cambia utilizando los parámetros aprendidos. La ecuación para la normalización de capas es:

$$y = \Upsilon \ \frac{x - \mu}{\sigma + \beta}$$

Donde “x” es el tensor de entrada, “μ” y “σ” son la media y la desviación estándar de las activaciones a lo largo de una dimensión en específico. “Ɣ” y “β” son parámetros aprendibles que escalan y cambian los valores normalizados, e “y” es el tensor de salida normalizado.

En general, la normalización de capas ayuda a que el modelo de optimización sea más estable al reducir el cambio de covariable interna, proporciona puntos de referencia consistente, maneja estadísticas de entrada variables y mitiga problemas de inicialización ya que reduce la sensibilidad de la red a pesos iniciales. Todo lo anterior contribuye a un entrenamiento más eficiente y un mejor rendimiento de las redes neuronales.

Dropout.

Dropout es una técnica de regularización comúnmente usada en redes neuronales para evitar que haya un sobreajuste. Sobreajuste ocurre cuando el modelo aprende muy bien los datos de entrenamiento, pero no generaliza muy bien los datos nuevos de validación. Dropout ayuda con este problema al reducir la dependencia del modelo en neuronas especificas durante el entrenamiento, y esto hace que la red aprenda representaciones más robustas y generalizadas

Dos capas nn.Dropout fueron usadas en el modelo para evitar que haya un sobreajuste. Dropout establece aleatoriamente una fracción de las unidades en una capa a cero durante el entrenamiento. La ecuación es la siguiente:

$$\boldsymbol{y = x\ \frac{mask}{1 - p}}$$

Donde “x” es el tensor de entrada, “*mask*” es una máscara binaria con la misma forma de “x” donde cada elemento se establece en 0 o 1 con probabilidad 1-p, e “y” es el tensor de salida. Durante la prueba, la capa Dropout se apaga y el tensor de salida se multiplica por (1-p) para conservar el valor esperado de salida.

Capa de salida (*output layer*).

La capa final de modelo es una capa completamente conectada que toma como entrada la salida de *dropout2*. El tamaño de salida de la capa es 1, que es el valor de salida final que se quiere predecir. La salida (*output*) de la capa final se calcula como:

$$y = Wx + b$$

Donde “W” es la matriz de peso, “x” es el tensor de entrada, “b” es termino de sesgo, e “y” es el tensor de salida final que se quiere predecir.

Forward.

El “método *Forward*” define el cálculo que se realiza en los datos de entrada para generar la salida del modelo. Se necesitan dos entradas:

- X_cat: un tensor que representan los atributos categóricos de los datos de entrada.
- X_cont: un tensor que representan los atributos continuos de los datos de entrada.

En el “método *Forward*”, los datos de entrada pasan primero a través de las capas de *embedding*, y los *embeddings* resultantes se concatenan a lo largo de la dimensión del atributo. Los atributos continuos también se concatenan con los *embeddings* a lo largo de la dimensión del atributo. Este tensor conectado luego pasa a través de las capas completamente conectadas (fc1, fc2), las capas normalizadas (layer_norm1, layer_norm2), las funciones de activación ReLU (relu1, relu2), y las capas de *Dropout* (*dropout1* y *dropout2*) en orden.

El output de la capa final de *Dropout* pasa luego a través de la capa de salida (*output_layer*) para generar la salida de la variable final que se quiere predecir.

Hiperparámetros e inicialización del Modelo.

Los hiperparámetros usados para el entrenamiento de la red neuronal son los siguientes:

***Learning rate*:** Tasa de aprendizaje. Es el tamaño de los pasos que se utilizan para el descenso del gradiente durante la optimización. Un *Learning Rate* alto puede hacer que el optimizador supere la solución óptima, mientras que un *Learning Rate* bajo puede hacer que el optimizador converja muy lentamente. La ecuación para actualizar los parámetros del modelo utilizando el descenso de gradiente estocástico con una tasa de aprendizaje α y un gradiente g es:

$$w_1 = w_0 - \alpha * g$$

W es el peso de los parámetros de la red neuronal. En este caso, se encontró que el *learning rate* que más minimizaba la función de pérdida era de **0.0003**

***Weight Decay*:** es la fuerza de regularización utilizada para evitar el sobreajuste al penalizar los grandes pesos (w) del modelo. La ecuación de regularización L2, que es el “*Weight Decay*” en nuestro caso es:

$$L2 = \lambda * ||w||^2$$

Donde λ es la fuerza de regulación, y “w” es el vector de peso. Para el modelo, se encontró el mejor *Weight decay* como **0.003**

***Dropout rate*:** es la probabilidad de abandonar unidades en la red para evitar el sobreajuste. Se utilizó un *dropout rate* de **0.1.**

Num_epoch: Número de épocas. Es la cantidad de veces que se usará todo el conjunto de datos para entrenar el modelo. Para este estudio, se usó un número de épocas de **1000**.

Batch_ size: Tamaño de lote. Es el número de muestras utilizadas para calcular el gradiente en cada iteración del algoritmo de optimización. Entrenar una red neuronal en todo el conjunto

de datos puede resultar una tarea “costosa” desde el punto de vista computacional. Para evitar esto, el conjunto de datos se divide en lotes. En este caso, se usó un *batch size* de **128**.

Entrenamiento del Modelo.

La red neuronal del modelo fue entrenado usando el 80% de los datos y fue optimizado mediante el **optimizador Adam (torch.optim.Adam())**, que es un algoritmo de optimización de descenso de gradiente estocástico que utiliza tasas de aprendizaje adaptables por cada parámetro. El optimizador Adam calcula una tasa de aprendizaje adaptable para cada parámetro en función del primer y segundo momento de la gradiente. Por otro lado, la **función de pérdida del error cuadrático medio (nn.MSELoss())** se usó para entrenar el modelo, ya que se usa comúnmente para problemas de regresión y mide la distancia entre los valores calculados y los valores verdaderos.

Recorte de Gradiente y Detención Anticipada.

Para evitar que los gradientes exploten durante el entrenamiento, se aplicó *Gradiante Clipping* o “recorte de gradiente” al optimizador. Este recorte de gradiente es una técnica que limita la magnitud de los gradientes a un valor máximo, lo que evita que se vuelvan demasiado grandes y provoquen que el modelo diverja durante el entrenamiento. Por otro lado, para evitar el sobreajuste, el *Early Stopping* o la “Detención Anticipada” se implementó mediante la clase ReduceLROnPlateau de Pytorch. La “detención anticipada” es una técnica que detiene el entrenamiento del modelo antes de tiempo si la pérdida del conjunto de validación no mejora durante cierto número de épocas, lo que se conoce como parámetro paciente. En el código utilizado, la tasa de aprendizaje se reduce en un factor de 0.1 si la pérdida del conjunto de validación no mejora durante 5 épocas consecutivas.

Para procesar esta gran cantidad de data, se utilizaron **TensorDataset** y **DataLoader** que son de ayuda en Pytorch para manejar y procesar gran cantidad de datos durante el entrenamiento. Proporcionan funciones como el preprocesamiento por lotes (*batches*), combinación aleatoria, y la carga de datos en paralelo que pueden hacer que el entrenamiento sea más eficiente y reducir el uso de la memoria.

Loop de entrenamiento.

El ciclo de entrenamiento itera sobre un número fijo de épocas, y cada época consiste en un pase hace adelante (*foward pass*) a través del modelo, una propagación hacia atrás (*backpropagation*) para calcular los gradientes, y un paso en el optimizador para actualizar los pesos del modelo.

Durante cada época, los datos de entrenamiento se pasaron a través del modelo en lotes utilizando la clase de *Dataloader* de Pytorch, la cual carga los datos en lotes y mezcla el orden de los lotes para evitar que el modelo memorice el orden de los datos. En este código, el tamaño del lote fue de 128. La pérdida de entrenamiento (***Training Loss***), se calculó

promediando la pérdida de cada lote y actualizando los pesos del modelo, en función de los otros gradientes calculados durante la retropropagación (*backpropagation*). La pérdida en el conjunto de validación (***Validation Loss***) se calculó utilizando el mismo proceso, pero con los datos de validación en lugar de los datos de entrenamiento. El propósito de la pérdida de validación es monitorear el rendimiento del modelo en datos nuevos, no vistos, y evitar el sobreajuste.

El ciclo de entrenamiento también incluye un "*Early Stopping*", que se implementa mediante el seguimiento de la pérdida de validación y el número de épocas desde que mejoró la pérdida de validación. Si la pérdida de validación no mejora durante 10 épocas, la tasa de aprendizaje se reduce a un factor de 0.1, y el entrenamiento continúa. Si la pérdida de validación no mejora durante toras 10 épocas, se detiene el entrenamiento y se devuelve el modelo con la mejor pérdida de evaluación.

Evaluación del modelo.

Finalmente, el modelo calculó varias métricas de evaluación para los conjuntos de entrenamiento y validación, incluido el raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE). Las métricas se calculan utilizando los valores calculados (train_preds, val_preds) y los valores verdaderos (train_targets, val_targets) obtenidos durante los ciclos de entrenamiento y validación. Las pérdidas tanto del conjunto de validación y entrenamiento se muestran en términos de la raíz del error cuadrático medio (RMSE) entre los valores calculados y los verdaderos valores objetivo, para evaluar el desempeño del modelo. Cuanto más pequeño es el MSE, mejor será el modelo para predecir la variable objetivo en nuevos datos, no vistos.

La raíz del error cuadrático medio se calculó mediante la siguiente formula:

$$RMSE = \sqrt{\frac{\sum(actual\text{ - }prediction)^2}{Number\ of\ observations}}$$

MAE (el error absoluto medio) se calculó mediante la siguiente formula:

$$MAE = \frac{\sum|actual\text{ - }prediction|}{number\ of\ observations}$$

PERMUTATION FEATURE IMPORTANCE (PFI)

Con el modelo predictivo de red neuronal construido, es necesario observar los atributos socioeconómicos que tienen impacto en la variable respuesta por medio de un análisis de importancia de atributos. *Permutation Feature Importance* (PFI) en Python es una técnica independiente del modelo que evalúa la importancia de cada característica al observar cómo cambia el rendimiento del modelo cuando los valores de los atributos en particular se permutan aleatoriamente. Valores negativos sugieren que la permutación de la función tuvo un impacto negativo en el rendimiento del modelo, lo que significa que el atributo es importante para la predicción del modelo. La magnitud de los valores proporciona una medida de relativa importancia, donde los valores más grandes indican una mayor importancia. Específicamente, esa puntuación de la importancia de la variable para un atributo determinado se calcula como la diferencia métrica del rendimiento calculada al conjunto de datos original y la métrica del conjunto de datos permutado, que se promedia en múltiples permutaciones. Para este estudio, la métrica usada fue el *Mean Absolute Error* (MAE).

Se usó esta técnica por varias razones: proporciona una interpretación intuitiva al medir el impacto de la permutación de los valores de los atributos en el rendimiento del modelo. PFI es adecuado para varios tipos de modelos ya que captura relaciones no lineales y es robusta a la correlación de los atributos. Es particularmente útil para grandes conjuntos de datos con numerosas características y puede ayudar a comprender qué variables son más influyentes en la predicción del modelo.

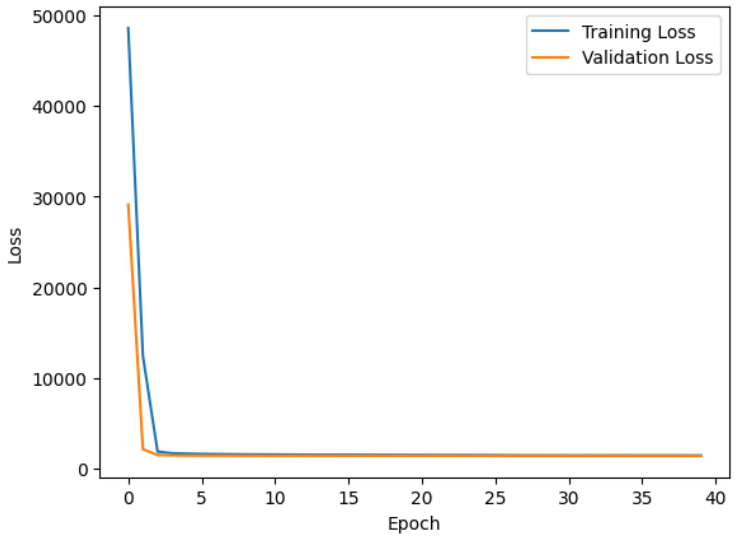
RESULTADOS

La red neuronal tiene la capacidad de aprender atributos complejos categóricos concatenado con las características numéricas de la base de datos. Para analizar los resultados, además de las funciones de pérdida tanto para el conjunto de entrenamiento y validación, se usaron el RMSE (*Root Mean Squared error*) y MAE (*Mean Absolute Error*). Las pruebas se corrieron en un computador ASUS X415M-BV462TS INTEL CELERON N4020 4GB 128GB SSD.

En la **Gráfica 9** y en la **tabla 5** del anexo, se observa que el modelo aprende, hasta cierto punto. La red neuronal es entrenada por 1000 épocas, pero se detiene anticipadamente en la época 40 ya que la función de pérdida del conjunto de validación no mejora en 10 épocas, lo que indica que el modelo ya no está aprendiendo de forma efectiva. A lo largo de las iteraciones, tanto la pérdida de entrenamiento como la de validación disminuyen significativamente en cada época, **lo que sugiere que el modelo está aprendiendo y mejorando sus predicciones**.

Visualización de las predicciones

Gráfica 9. Loss Curve: Validation loss and training loss



Como se evidencia en la Gráfica 9, la función de pérdida de los datos de entrenamiento decreció de 48586.9120 en la primera época a 1461.7518 en la época 40, indicando que el modelo mejora su habilidad de predecir la data de entrenamiento en cuanto aprende. La función de validación sigue un patrón similar, cayendo de 29128.5664 a 1429.2660, lo que significa que el desempeño predictivo del modelo también mejora con datos no vistos.

El RMSE de los datos de entrenamiento empezó en 220.4421 y bajó a 38.2283 al finalizar la época 40. Similarmente los datos de validación, el RMSE disminuyó de 170,6719 a 37.8051, mostrando la capacidad de generalizar del modelo. El MAE, por otro lado, muestra tendencia similar. En el conjunto de entrenamiento, el MAE cayó de 213.0501 a 30.6102, y en el conjunto de validación bajó del 162.9504 a 30.3248. La tasa de aprendizaje se ajustó dos veces: en la época 26, la tasa fue ajustada en $3.00e^{-05}$, y en la época 35 fue de nuevo ajustada al $3.00e^{-06}$. Esta técnica es utilizada para ayudar al modelo a converger más eficientemente cuando las pérdidas empiezan a estancarse. Lo anterior, se puede ver a continuación en la Tabla 4:

Tabla 4

Resumen Comportamiento Aprendizaje del Modelo

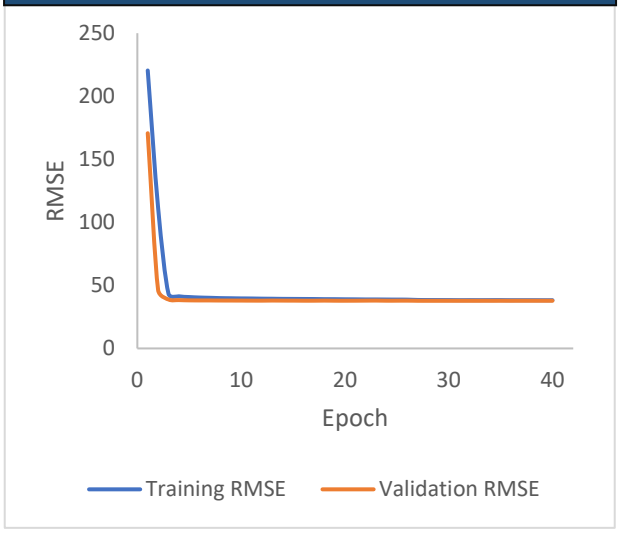
	Training Loss	Validation Loss	Training RMSE	Validation RMSE	Training MAE	Validation MAE
Epoch 26	1497.9692	1436.6665	38.706	37.903	31.0305	30.3697
Epoch 26	reducing learning rate of group 0 to 3.0000e-05.					
Epoch 27	1477.4839	1427.5693	38.4403	37.7827	30.7931	30.32
Epoch 28	1473.8431	1427.5252	38.3908	37.7821	30.7486	30.2869
Epoch 29	1471.3037	1427.4005	38.355	37.7805	30.7363	30.2942

Epoch 30	1470.3845	1428.7813	38.3434	37.7988	30.7147	30.3109
Epoch 31	1464.7885	1427.7291	38.2716	37.7848	30.6536	30.3261
Epoch 32	1466.4733	1427.7636	38.2944	37.7853	30.6845	30.3194
Epoch 33	1465.4453	1428.8078	38.2834	37.7991	30.6663	30.3123
Epoch 34	1466.8986	1428.6248	38.3054	37.7966	30.6848	30.3185
Epoch 35	1467.7764	1428.922	38.3095	37.8005	30.6853	30.3318
Epoch 35	reducing learning rate of group 0 to 3.0000e-06.					
Epoch 36	1463.5555	1428.8784	38.255	37.7999	30.6416	30.3231
Epoch 37	1462.7737	1429.108	38.2495	37.803	30.6324	30.3251
Epoch 38	1463.2555	1429.3757	38.2517	37.8065	30.6485	30.3177
Epoch 39	1461.0104	1429.2151	38.2267	37.8044	30.6204	30.3245
Epoch 40	1461.7518	1429.266	38.2283	37.8051	30.6102	30.3248
Validation loss hasn't improved in 10 epochs. Stopping early.						

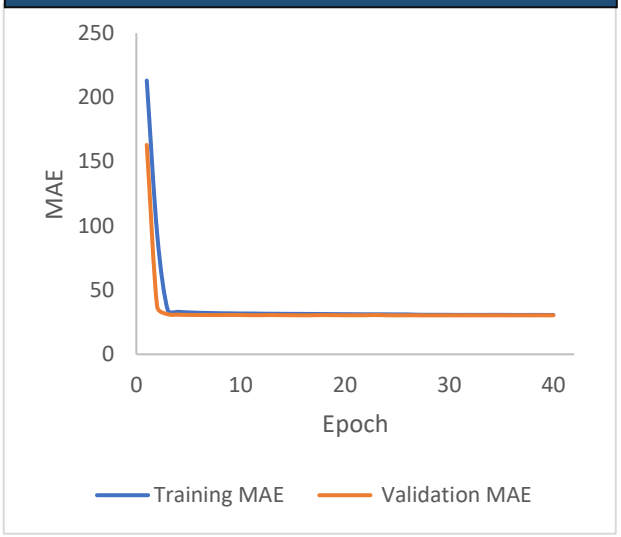
Note. ***Epoch=1000***

En consonancia con lo anterior, el optimizador Adam, que ajusta la tasa de aprendizaje para cada parámetro durante el entrenamiento en un factor de 0.1, logró obtener los siguientes resultados después de la iteración número 40, ajustando la tasa de aprendizaje a un $3.00e^{-06}$: una función de pérdida de 1429.266 y 1461.7518 para el conjunto de validación y entrenamiento respectivamente, con un RMSE 38.2283 y un MAE de 30.3248 para los datos de validación. En la **Gráfica 10** y en la **Gráfica 11**, se muestra el comportamiento tanto de RMSE y MAE tanto para el conjunto de entrenamiento y validación, a lo largo del aprendizaje del modelo:

Gráfica 10: RMSE, Training vs Validation



Gráfica 11: MAE, Training vs Validation



Después de esto, no hubo mejoras en el modelo, indicando que el modelo llegó a su límite dado la arquitectura del modelo y los datos. En general, el modelo parecer estar aprendiendo efectivamente de los datos entrenamiento como lo indican las métricas de pérdida y de error decrecientes. El modelo parece tener un desempeño estable en el conjunto de validación y

entrenamiento, mostrando una buena capacidad de generalización, al menos hasta la época 40. Al finalizar el entrenamiento, los errores del entrenamiento y validación son similares (38 y 40), revelando que hay un buen ajuste sin un aparente sobreajuste (*overfitting*) o desajuste (underfitting).

Finalmente, se aplicó PFI, donde se observa qué tanto impacta cada variable socioeconómica dentro del modelo. Los resultados de todas las variables socioeconómicas que más inciden en el desempeño de los estudiantes de undécimo en las pruebas Saber 11, 2019 se pueden ver en **Tabla 10** del anexo. El valor negativo indica que al permutar la variable conduce a una diminución en el rendimiento del modelo. Los atributos socioeconómicos que más impactan el modelo son:

1. **La jornada del colegio** (-2.0848): la jornada de la institución educativa (completa, mañana, noche, sabatina, tarde, única)**,** es el atributo más importante evaluado en el modelo predictivo. Permutando los valores de esta característica reduce significativamente el rendimiento del modelo, lo que destaca la importancia del horario escolar para predecir el Puntaje Global en las Pruebas Saber 11.

2. **Departamento donde está ubicada la sede educativa** (-1.8912): El departamento donde se ubica el colegio del estudiante es la segunda característica más importante del modelo. La permutación de este atributo provoca una diminución sustancial en el rendimiento del modelo, lo que indica que el departamento tiene un fuerte impacto en los resultados de los estudiantes.

3. **ESTU_NSE_INDIVIDUAL o nivel socioeconómico del evaluado** (-0.8611): “El NSE se usa para caracterizar la población que presenta las pruebas Saber 11. Esta caracterización permite tener en cuenta no solo el nivel de ingresos, sino también posesión de bienes, acceso a servicios y educación del núcleo familiar, lo cual brinda una visión completa del estudiante” (Instituto Colombiano para la Evaluación de la Educación (ICFES) y Ministerio de Educación Nacional, 2019, pág. 5).

4. **Sexo del estudiante** (-0.8495): el sexo del estudiante es la cuarta característica más influye dentro de la tarea predictiva de la red neuronal. Cuando se permutan los valores de esta función, el rendimiento del modelo se ve afectado considerablemente, lo que muestra que el sexo juega un papel crucial en las predicciones.

5. **Tiempo de dedicación a la lectura diaria** (-0.6721): el tiempo que un estudiante dedica a la lectura al día es la quinta variable más importante. Cuando se permuta esta característica, el rendimiento del modelo cae, destacando la importancia del tiempo de lectura diaria en las predicciones.

Lo anterior, se puede ver reflejo en la **Tabla 5** a continuación:

Tabla 5

Resumen Variables Socioeconómicas que impactan el modelo

	Variables Socioeconómica	Impacto en el modelo
1	La jornada del colegio	-2.0848
2	Departamento donde está ubicada la sede educativa	-1.8912
3	Nivel socioeconómico del evaluado	-0.8611
4	Sexo del estudiante	-0.8495
5	Tiempo de dedicación a la lectura diaria	-0.6721

Estos resultados contrastan con los otros trabajos realizados sobre este tema: autores como Blanco (2015), Ávila et al. (2021) y García-González y Skrita (2019) encontraron que la educación de los padres es de las variables socioeconómicas que más influyen en los resultados académicos de los estudiantes. Los resultados obtenidos en este trabajo están más en línea con los resultados encontrados por Timarán-Pereira et al. (2019) . Sin embargo, ninguno de estos autores encuentra el nivel socioeconómico del evaluado (diferente al estrato socioeconómico), como uno de los atributos más influyentes en el modelo.

AJUSTES AL MODELO

Se ajustaron los siguientes hiperparámetros del modelo para tratar de alcanzar el mejor resultado de optimización posible:

1. *Learning rate*: se valoró el modelo usando diferentes tasas de aprendizaje tales como 0.1, 0.001, 0.0001, 0.3, 0.03, 0.003, 0.0003.
2. *Weight Decay*: se evaluó el modelo usando diferentes valores para el *weight decay* tales como 0.1, 0.001, 0.0001, 0.3, 0.03, 0.003, 0.0003.
3. *Dropout rate*: se usaron valores de 0.1 hasta 0.8 para correr las pruebas de optimización.

Adicionalmente, se realizaron modificaciones en la arquitectura de la red neuronal ajustando el número de capas ocultas (1, 2 o 3), el número de neuronas de entrada y de salida, y el número de lote (64 y 128). La función de activación ReLU y la función optimizadora Adam, se usaron en todos los escenarios, ajustando manualmente. Ninguna de estas modificaciones lograron una función de pérdida menor que 1400, tanto para el set de validación como para el de entrenamiento. Es necesario que las funciones de pérdidas para ambos sets tiendan a ser similar, para observar convergencia y saber que el modelo no vaya a presentar sobreajuste. En La **Tabla 6** a continuación, se muestra un resumen de algunos de los diferentes hiperparámetros que se utilizaron con sus respectivos resultados:

Tabla 6

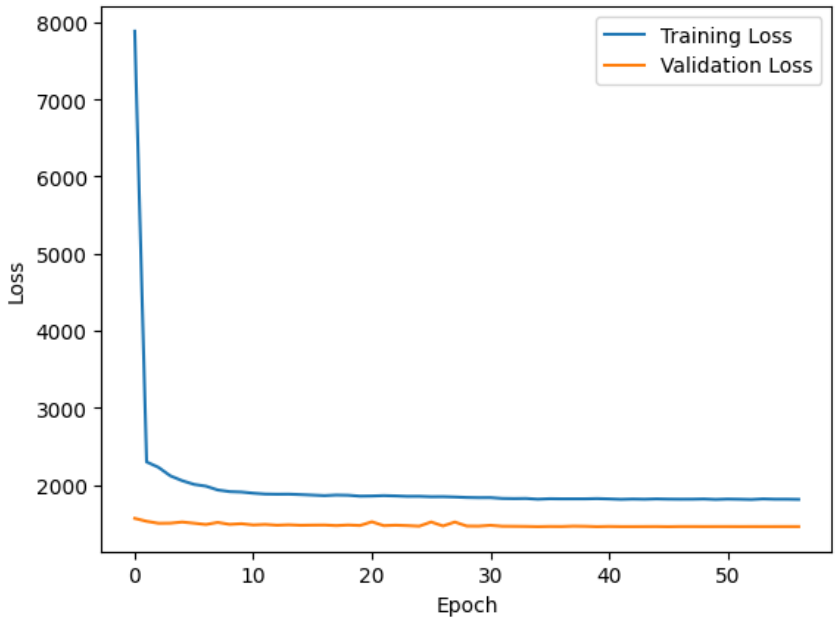
Ajustes Manuales Red Neuronal

Modelo1	Número de Capas Ocultas: 3 Número de neuronas: 220 *Learning rate: 0.003* *Weight Decay: 0.00003* *Dropout rate: 0.4*	*Training Loss: 1869* *Validation Loss:1518*
Modelo2	Número de Capas Ocultas: 3 Número de neuronas: 128 *Learning rate: 0.003* *Weight Decay: 0.0003* *Dropout rate: 0.1*	*Training Loss: 1573* *Validation Loss: 1437*
Modelo3	Número de Capas Ocultas:4 Número de neuronas: 128 *Learning rate: 0.003* *Weight Decay: 0.0003* *Dropout rate:0.1*	*Training Loss: 1642* *Validation Loss: 1434*
Modelo4	Número de Capas Ocultas:2 Número de neuronas: 128 *Learning rate:0.03* *Weight Decay:0.003* *Dropout rate:0.1*	*Training Loss: 1777* *Validation Loss: 1516*
Modelo5	Número de Capas Ocultas: 2 Número de neuronas: 128 *Learning rate:0.003* *Weight Decay: 0.0003* *Dropout rate: 0.5*	*Training Loss: 3801* *Validation Loss: 1747*

En la **Gráfica 12,** se muestra el comportamiento del modelo con las tres capas ocultas, 128 neuronas, una tasa de aprendizaje de 0.0003, un *weight decay* de 0.0003, *dropout rate* de 0.1 y un a *batch size* de 128. Aquí se observa que, a diferencia de la Gráfica 9, este modelo no es el mejor por varias razones. En primer lugar, el conjunto de validación parece no disminuir en ningún momento, manteniéndose constante. Esto no sirve porque da indicios de que no hay un aprendizaje por parte del modelo, especialmente en el set de validación. En segundo lugar, el conjunto de entrenamiento y validación parecen no converger en ningún momento, lo cual es muy importante ya que esto dice, gráficamente, que no hay sobreajuste. Y en tercer lugar, este modelo tiene métricas más altas, alcanzando una función de pérdida 1811 y 1459, para el

conjunto de entrenamiento y validación respectivamente. En resumen, la Gráfica 12 muestra que el modelo no realizó bien su tarea predictiva.

Gráfica 12: 3 Capas, 128 neuronas, Comportamiento



Por otro lado, se utilizaron otras técnicas dentro de Pytorch, incluyendo Optuna, para encontrar la mejor combinación de parámetros posible para bajar tanto la función de error en la red neuronal como el RMSE y el MAE. Con Optuna, se puede definir un espacio de búsqueda, configurar la función objetivo y dejar correr el proceso de optimización, mejorando la última instancia el rendimiento y la eficiencia de los modelos. Optuna se demoró corriendo aproximadamente 12 horas, sin ningún resultado prometedor que pudiera mejorar el rendimiento del modelo, tanto para **set de validación como para el de entrenamiento**. En la **tabla 7** a continuación, se puede ver un resumen de los 19 resultados que arrojó Optuna. Los resultados sólo muestran las funciones de pérdida. En la **Tabla 11**, del anexo, se puede encontrar en detalle los hiperparámetros utilizados para llegar a cada uno de estos resultados.

Tabla 7

Resultados de Optuna

Modelo	Iteración	Training Loss	Validation Loss
1	Epoch 17	2364.8763	1565.2167

2	Epoch 48	2552.2559	1519.9951
3	Epoch 76	1977.842	1447.9375
4	Epoch 99	1821.7439	1438.0112
5	Epoch 55	1852.127	1439.4335
6	Epoch 63	2238.1553	1483.9757
7	Epoch 46	1464.846	1404.9254
8	Epoch 45	1500.5408	1402.3572
9	Epoch 36	1639.1509	1426.3957
10	Epoch 41	1704.8643	1433.8892
11	Epoch 17	1786.6241	1481.6397
12	Epoch 60	1518.083	1405.9135
13	Epoch 68	1520.1644	1403.5144
14	Epoch 56	2162.985	1418.5194
15	Epoch 30	1682.6241	1417.8785
16	Epoch 18	1595.6192	1427.5553
17	Epoch 49	1741.0114	1432.8201
18	Epoch 14	1771.4622	1502.7038
19	Epoch 45	1869.7065	1503.5389

Note. ***Epoch=100***

Para obtener los resultados de la Tabla 7, Optuna buscó los mejores parámetros e hiperparámetros dentro del siguiente espacio de búsqueda:

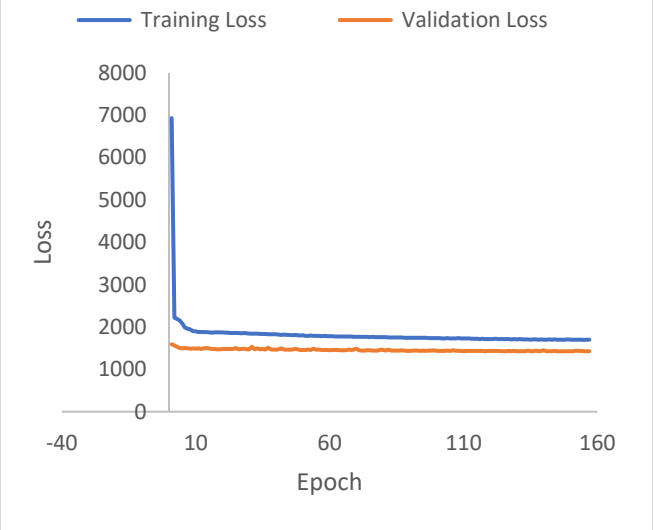
- Número de capas: 1, 3
- Función de Activación 1: ReLU, LeakyReLU, ELU
- Función de Activación 2: ReLU, LeakyReLU, ELU
- *Hidden Layer 1:* 32, 512
- *Hidden Layer 2:* 32, 256
- *Dropout rate: 0.1, 0.5*
- *Weight Decay: :* $1e^{-6}$, $1e^{-2}$
- *Batch size:* 32, 64, 128
- *Learning rate:* $1e^{-5}$, $1e^{-2}$,

El mejor resultado, el modelo 7, el cual cuenta con funciones de perdidas 1464 y 1404 para los conjuntos de entrenamiento y validación respectivamente, no tiene el comportamiento deseado, ya que el set de validación permanece constante y no muestra indicios de que aprende. Optuna, además, combina funciones de activación, algo que no se hizo a la hora de ajustar el modelo manualmente. Este comportamiento se puede ver en la **Gráfica 13**:

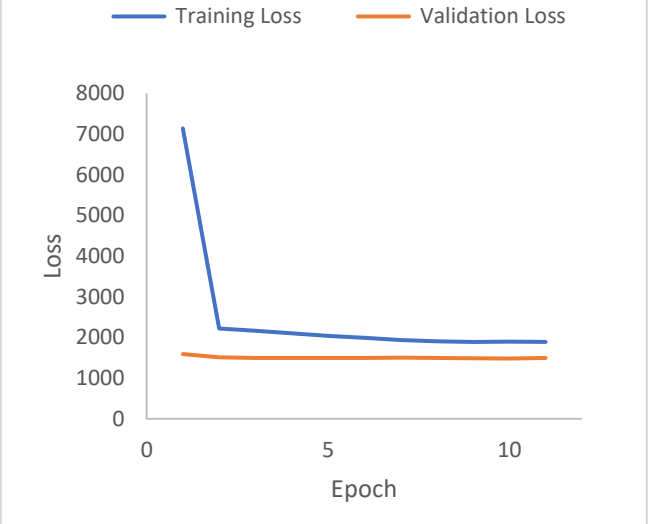
Gráfica 13: Optuna, Modelo 7

También se usó el método de **ensemble learning** para mejorar el rendimiento en la tarea predictiva del modelo. Con ensemble learning, se entrenaron 3 modelos y se promediaron sus predicciones. Con los mismos parámetros e hiperparámetros usados en el modelo original, se obtuvo un promedio RMSE de 38.1807 y promedio MAE de 30.6762. No hubo mejoras. En la En los **Gráficos 14, 15 y 16**, a continuación, se ve un resumen de las iteraciones que se hicieron con los 3 modelos, y el resultado final, en términos de función de pérdida:

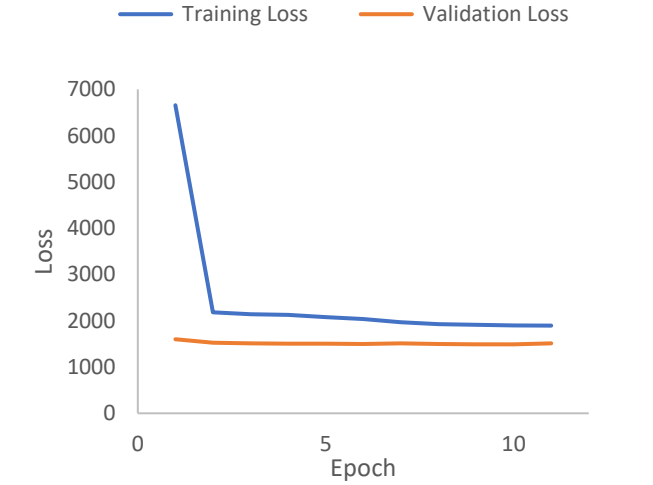
Gráfica 14: Modelo1, Ensemble Learning



Gráfica 15: Modelo2, Ensemble Learning



Gráfica 16: Modelo3, Ensemble Learning



El primero modelo entrenó y validó el conjunto de datos en 157 épocas, el segundo en 11, y el tercero 11, antes de ser detenidas por el *Early Stopping*.

Se utilizó, finalmente, normalización de capas. El módulo nn.LayerNorm en PyTorch realiza una normalización de capas: normaliza las entradas en la dimensión especifica, que es la última dimensión por defecto. Al usar la normalización de capas en la red neuronal, ayuda estabilizar el proceso de entrenamiento y mejorar el rendimiento en general. Este último logró mejorar algo el MAE pasando de 30.7039 a 30.3248 para el set de validación.

LIMITACIONES

Se probaron varios métodos y procedimientos para bajar la función de pérdida del modelo, incluyendo ajustes manuales y un marco de optimización de hiperparámetros de código abierto como Optuna. Sin embargo, el mínimo MAE (mean absolute error) al que el modelo pudo llegar fue de 30.3248. Como se ha observado, mientras que el modelo de red neuronal creada en este estudio tiene resultados prometedores en las predicciones de la variable objetivo, tiene algunas limitaciones en su tarea predictiva para tener en cuenta. Una limitación es el Error Absoluto Medio (MAE) relativamente alto de 30.3248, lo que sugiere que el modelo puede no ser lo suficientemente preciso para algunas aplicaciones. Se quiere que el MAE sea el más bajo posible, lo cual indica que la etiqueta de salida predicha por el modelo se aproxima con la etiqueta de salida real de los datos.

Además, el rendimiento del modelo puede depender en gran medida del calidad y cantidad de los datos de entrenamiento, lo que puede limitar su generalización a conjuntos de datos nuevos o diferentes. Por lo tanto, es importante evaluar cuidadosamente las limitaciones del modelo y seguir considerando otros enfoques para poder mejorar su capacidad predicativa.

Sin embargo, con todos de ajustes y métodos utilizados para aumentar el rendimiento del modelo, **los resultados se mantuvieron robustos en casi todos los modelos considerados en este trabajo**, conservando los resultados del PFI (Permutation Feature Importance) relativamente constantes. Aunque indicadores como el MAE o las funciones de perdida pueden seguir mejorando, los resultados no son descartables y pueden considerarse para observar los factores socioeconómicos que inciden en el desempeño de los estudiantes en la Pruebas Saber 11.

CONSIDERACIONES FINALES

La red neuronal de este trabajo sirvió para consolidar un modelo con una función de pérdida de 1429.2660, RMSE de 37.8051 y un MAE de 30.3248 para el conjunto de validación en la iteración 40 dentro del loop de entrenamiento. **Se escogió este modelo ya que tiene las funciones de pérdida más bajas tanto para el conjunto de entrenamiento como el de validación.** Usando un *Permutation Feature Importance (PFI)*, se pudieron encontrar los atributos que más inciden en el desempeño de los estudiantes de undécimo. La característica con más impacto en el rendimiento del modelo es **la jornada del colegio** (-2.0848), seguida **del departamento donde está ubicada la sede educativa** (-1.8912), el **nivel socioeconómico del evaluado** (-0.8611), **el sexo del estudiante** (-0.8495), y el **tiempo de dedicación a la lectura diaria** (-0.6721).

Además, la creación de esta red neuronal usando herramientas de Deep Learning en Pytorch, contrasta con otras metodologías propuestas por otros trabajos citados en este estudio, tales como el método de análisis de componentes principales de Ortega et al. (2021), o el modelo de clasificación basado en arboles de decisiones usado por Timarán-Pereira et al. (2019) y Ávila et al. (2021).

CONCLUSIONES

Durante esta investigación, se observó cómo se usaron técnicas avanzadas de Aprendizaje de Máquinas, como Deep Learning, para estructurar y entrenar un modelo con 35 variables socioeconómicas a nuestra disposición, para identificar los atributos socioeconómicas que más influyen en los resultados de las Pruebas Saber 11 de más 300,000 estudiantes en el segundo semestre del 2019.

Para lograr este propósito, se creó en Pytorch una red neuronal con 2 capas ocultas normalizadas, utilizando una función de activación ReLU, con 500 neuronas y una tasa de *dropout* de 0,1. Para asegurar un aprendizaje del modelo se estableció una tasa de aprendizaje de 0,0003 y un *weight decay* de 0,0003, usando una función de Optimización Adam y una función de pérdida de Error Cuadrático Medio. El modelo se entrenó para 1000 épocas utilizando un tamaño de lote de 128. Se evaluó la red neuronal llegando a una función de pérdida de 1429, RMSE de 37.80, y un MAE de 30.32 para el conjunto de validación, y una función de pérdida de 1461.75, RMSE de 38,22, y un MAE de 30.61 para el conjunto de entrenamiento.

Finalmente, por medio de un PFI (*Permutation Feature Importance*), se llegó a que las variables socioeconómicas que más inciden en el desempeño de los estudiantes de undécimo en las pruebas de Estado Saber 2019 usando Redes Neuronales son la jornada escolar, el departamento donde está ubicada la sede educativa, el nivel socioeconómico del evaluado, el sexo del estudiante, y el tiempo de dedicación a la lectura diaria.

El modelo construido en esta tesis puede ser una de las varias otras herramientas para la toma de decisiones en la educación media, y así estimular que haya una movilidad social para las nuevas generaciones. Mientras que los exámenes estandarizados son solo un componente entre muchos para evaluar el perfil de un estudiante (además que en los últimos años se ha estado desarrollando tener una mirada holística en la formación del estudiante), la realidad es que siguen siendo un factor importante para que el estudiante tenga un acceso a la educación superior, ya sea universitaria, técnica o tecnológica. También es un indicador de qué tanto tiene el estudiante desarrollado ciertas competencias y aptitudes fundamentales para resolver problemas bajo un contexto especifico a lo largo de la vida.

Mas específicamente, en términos aplicados, este trabajo le puede servir a los directivos, padres de familia y docentes de las instituciones educativas para incrementar el desempeño de los estudiantes de undécimo en las pruebas Saber 11. Se pueden contemplar políticas públicas para aumentar la jornada escolar de los estudiantes, tales como: horas escolares extendidas para brindar tiempo adicional de apoyo a los estudiantes, reducción de ausentismo (seguimiento de asistencia), horarios de aprendizaje flexible, intervenciones dirigidas a los estudiantes con alto riesgo de rezagarse académicamente y brindar apoyo como tutorías individualizadas o mentorías, inversión en infraestructura y recursos para apoyar en la enseñanza efectiva a los estudiantes durante las horas extendidas, capacitación y apoyo a docentes, entre otros.

Asimismo, ajustar el currículo de los docentes para incentivar más la lectura dentro y fuera del aula puede ayudar significativamente a cerrar esas brechas socioeconómicas, de género y regionales, según se vio en este estudio. Esto también le puede servir a ciertos agentes del sector público y privado para de focalizar intervenciones en el sistema educativo en los diferentes territorios de Colombia (programas de lectoescritura, donación de libros, entre otros).

Para futuros trabajos, es necesario seguir ajustando la red neuronal para obtener mejores resultados en cuanto a función de perdida, RMSE y MAE, y así compararlo con otros modelos de aprendizaje de máquinas como los regresores de *Random Forest o Decision Tree*. Por otro lado, se podrían aplicar otras técnicas para hacer un análisis de importancia de atributos como SHAP, y observar si los resultados se mantienen frente a PFI. Sería interesante crear una red neuronal para cada Prueba (matemáticas, inglés, competencias ciudadanes, etc), y observar si las variables socioeconómicas que se usaron en este trabajo inciden de la misma manera en la predicción de los resultados. También se podría ver el impacto postpandemia, y cómo el aprendizaje en esta etapa influenció sus resultados.

Por último, una posibilidad para mejorar el modelo en su tarea predictiva es aumentar los datos de entrada (*inputs*) y así observar si la Red Neuronal puede generar mejores relaciones entre los atributos para calcular la variable objetivo. Para esto, se puede tomar más años de aplicación de la Prueba de Estado, con diferentes estudiantes alrededor del país.

BIBLIOGRAFÍA

Álvarez-López, G., & Matarranz, M. (2020). Quality and evaluation as global educational political trends: Comparative study of national evaluation agencies in compulsory education in Europe. *Revista Complutense de Educacion*, *31*(1), 85–95. https://doi.org/10.5209/rced.61865

Ávila, L. K., Ospino Gutierrez, E., & Paez Reales, J. A. (2021). *Análisis de resultados de las pruebas saber 11 implementando técnicas de minería de datos*.

Blanco, V. P. (2015). *Análisis del Desempeño Académico del Examen de Estado para el Ingreso a la Educación Superior Aplicando Minería de Datos*.

Demarchi, G. D. (2020). *La evaluación desde las pruebas estandarizadas en la educación en Latinoamérica*.

Dirección de Gestión y Evaluación de la Calidad (DGEC). (2019). *Evaluaciones Internacionales LLECE - ERCE 2019*. https://dgec.mep.go.cr/evaluaciones-internacionales-llece

Flotts, M. P., Manzi, J., Jiménez, D., Abarzúa, A., Cayuman, C., & García, M. J. (2016). *Informe de resultdos TERCE: logros de aprendizaje*. www.acentoenlace.cl

García-González, J. D., & Skrita, A. (2019). Predicting Academic Performance Based on Students' Family Environment: Evidence for Colombia Using Classification Trees

Academic Performance Based on Students´FamilyStudents´Family Environment. *Psychology*, *11*(3), 299–311. https://doi.org/10.25115/psye.v10i1.2056

González, L. M. (2019). *Patrones de desempeño en pruebas estandarizadas y de calidad en institucionesde educación superior: evidencia basada en datos a partir de resultadosindividuales en Pruebas Saber*.

Instituto Colombiano para la Evaluación de la Educación (ICFES). (2021). *Guía de orientación Pruebas saber 11° 2022-1*.

Instituto Colombiano para la Evaluación de la Educación (ICFES), & Ministerio de Educación Nacional. (2019). *SABER AL DETALLE*.

Ministerio de Educación Nacional. (2016). *Pruebas Saber*. https://www.mineducacion.gov.co/1759/w3-printer-244735.html

Ministerio de Educación Nacional, & Instituto Colombiano para la Evaluación de la Educación (ICFES). (2022). *DICCIONARIO DE DATOS SABER 11 (APLICACIONES DESDE 2019-1 A 2022-1)*.

Miranda, L., & Schleicher, A. (2009). *La educación peruana en el contexto de PISA*.

Naciones Unidas. (2015a). *EDUCACIÓN DE CALIDAD: POR QUÉ ES IMPORTANTE*. http://www.un.org/

Naciones Unidas. (2015b). *Objetivo 4: Garantizar una educación inclusiva, equitativa y de calidad y promover oportunidades de aprendizaje durante toda la vida para todos*. https://www.un.org/sustainabledevelopment/es/education/#:~:text=En%20abril%20de%202020%2C%20cerca,otras%20fuentes%20de%20nutrici%C3%B3n%20diaria.

Oecd. (2014). *How Was Life? Global Well-being Since 1820*.

Ortega, J. E., Valencia Jiménez, N. N., & Álvarez Cogollo, J. L. (2021). Familia y éxito escolar según las pruebas de calidad del Estado en una región del Caribe colombiano. *Publicaciones de la Facultad de Educacion y Humanidades del Campus de Melilla*, *51*(2), 427–434. https://doi.org/10.30827/publicaciones.v51i2.18150

Portal Educativo Red Académica. (2022, agosto 11). *¿Cuáles son las diferencias entre jornada única y jornada completa?* YouTube. https://www.youtube.com/watch?v=2B74W0aoB08

Rivas, A., & Scasso, M. (2017). *¿Qué países mejoraron la calidad educativa? América Latina en las evaluaciones de aprendizajes*.

Stevens, E., Antiga, L., & Viehmann, T. (2020). *Deep Learning with PyTorch*.

The World Bank. (2022). *Metadata Glossary, Literacy rate, adult total (% of people ages 15 and above)*. https://databank.worldbank.org/metadataglossary/environment-social-and-governance-(esg)-data/series/SE.ADT.LITR.ZS

Timarán, R., Caicedo Zambrano, J., & Hidalgo Troya, A. (2019). Identification of factors associated with academic performance in mathematics in the saber 11th tests applying educational data mining. *Proceedings of the LACCEI international Multi-conference for*

Engineering, Education and Technology, *2019-July*. https://doi.org/10.18687/LACCEI2019.1.1.297

Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. *Revista de Investigación, Desarrollo e Innovación*, *9*(2), 363–378. https://doi.org/10.19053/20278306.v9.n2.2019.9184

undp. (2015, junio 23). *Movilidad social* . undp. https://www.undp.org/es/mexico/publications/movilidad-social

Valora Analitik. (2021, octubre 27). *Colombia: país con más desigualdad de la Ocde y el segundo en América Latina*. Valora Analitik. https://www.valoraanalitik.com/2021/10/27/colombia-pais-mas-desigualdad-ocde-segundo-america-latina/

Zhang, H., Cheng, X., & Cui, L. (2021). Progress or stagnation: Academic assessments for sustainable education in rural China. *Sustainability (Switzerland)*, *13*(6). https://doi.org/10.3390/su13063248

ANEXO

TABLAS

Tabla 8

Datos usados para el análisis

Variable	Descripción	Valores
ESTU_GENERO	Género del estudiante	F - Femenino M – Masculino
ESTU_TIENEETNIA	¿Pertenece el estudiante a un grupo étnico minoritario?	No Si
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de la vivienda según recibo de energía eléctrica	Estrato 1 Estrato 2 Estrato 3 Estrato 4 Estrato 5 Estrato 6 Sin estrato
FAMI_PERSONASHOGAR	¿Cuántas personas confirman el hogar donde vive actualmente, incluido el estudiante?	1 a 2 3 a 4 5 a 6 7 a 8 9 o más
FAMI_CUARTOSHOGAR	En total, ¿en cuantos cuartos duermen las personas de su hogar?	Uno Dos Tres

		Cuatro Cinco Seis o más
FAMI_EDUCACIONPADRE	Nivel educativo más alto alcanzado por el padre	- Ninguno - Primaria incompleta - Primaria completa - Secundaria (Bachillerato) incompleta - Secundaria (Bachillerato) completa -Técnica o tecnológica incompleta -Técnica o tecnológica completa -Educación profesional incompleta - Educación profesional completa - Postgrado -No Aplica -No sabe
FAMI_EDUCACIONMADRE	Nivel educativo más alto alcanzado por la madre	- Ninguno - Primaria incompleta - Primaria completa - Secundaria (Bachillerato) incompleta - Secundaria (Bachillerato) completa -Técnica o tecnológica incompleta -Técnica o tecnológica completa -Educación profesional incompleta - Educación profesional completa - Postgrado -No Aplica -No sabe
FAMI_TRABAJOLABORPADRE	Señale aquella labor que sea más similar al trabajo que realizó su padre durante la mayor parte del último año	- Es agricultor, pesquero o jornalero. - Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial. - Es dueño de un negocio pequeño (tiene pocos empleados o no tiene,por ejemplo, tienda, papelería, etc. - Es operario de máquinas o conduce vehículos (taxista, chofer). - Es vendedor o trabaja en atención al público. - Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente). - Trabaja como personal de limpieza, mantenimiento, seguridad o construcción. - Trabaja como profesional (por ejemplo, médico, abogado, ingeniero). - Trabaja en el hogar, no trabaja oestudia. - Trabaja por cuenta propia (por ejemplo, plomero, electricista). - Pensionado. - No sabe. - No aplica
FAMI_TRABAJOLABORMADRE	Señale aquella labor que sea más similar al trabajo que realizó su madre durante la mayor parte del último año	- Es agricultor, pesquero o jornalero. - Es dueño de un negocio grande, tiene un cargo de nivel directivo o gerencial. - Es dueño de un negocio pequeño (tiene

		pocos empleados o no tiene, por ejemplo, tienda, papelería, etc. - Es operario de máquinas o conduce vehículos (taxista, chofer). - Es vendedor o trabaja en atención al público. - Tiene un trabajo de tipo auxiliar administrativo (por ejemplo, secretario o asistente). - Trabaja como personal de limpieza, mantenimiento, seguridad o construcción. - Trabaja como profesional (por ejemplo, médico, abogado, ingeniero). - Trabaja en el hogar, no trabaja oestudia. - Trabaja por cuenta propia (por ejemplo, plomero, electricista). - Pensionado. - No sabe. - No aplica
FAMI_TIENEINTERNET	¿El hogar del estudiante cuenta con servicio o conexión a internet?	No Si
FAMI_TIENESERVICIOTV	¿El hogar del estudiante cuenta con servicio cerrado de televisión?	No Si
FAMI_TIENECOMPUTADOR	¿Cuáles de los siguientes bienes posee su hogar?: Computador	No Si
FAMI_TIENELAVADORA	¿Cuáles de los siguientes bienes posee su hogar?: Lavadora de ropa	No Si
FAMI_TIENEHORNOMICROOGAS	¿Cuáles de los siguientes bienes posee su hogar?: Horno Microondas u Horno eléctrico oa gas	No Si
FAMI_TIENEAUTOMOVIL	¿Cuáles de los siguientes bienes posee su hogar?: Automóvil particular	No Si
FAMI_TIENEMOTOCICLETA	¿Cuáles de los siguientes bienes posee su hogar?: Motocicleta	No Si
FAMI_TIENECONSOLAVIDEOJUEGOS	¿Cuáles de los siguientes bienes posee su hogar?: Consola para juegos electrónicos (PlayStation, Xbox, Nintendo, etc.)	No Si
FAMI_NUMLIBROS	¿Cuántos libros físicos o electrónicos hay en su hogar excluyendo periódicos, revistas, directorios telefónicos y libros del colegio?	0 a 10 libros 11 a 25 libros 26 a 100 libros Más de 100 libros
FAMI_COMELECHEDERIVADOS	¿Cuántas veces por semana se comen los siguientes alimentosen su hogar? Leche o derivados (queso, yogurt, etc.)	1 o 2 veces por semana 3 a 5 veces por semana Nunca o rara vez comemos eso Todos o casi todos los días
FAMI_COMECARNEPESCADOHUEVO	¿Cuántas veces por semana se comen los siguientes alimentos en su hogar? Carne (pollo, pavo, res, cordero, cerdo, conejo, etc.), pescados o huevos	1 a 2 veces por semana 3 a 5 veces por semana Nunca o rara vez comemos eso Todos o casi todos los días
FAMI_COMECEREALFRUTOSLEGUMBRE	¿Cuántas veces por semana se comen los	1 a 2 veces por semana

	siguientes alimentosen su hogar? Cereales (avena, granola), frutos secos (almendras, maní) o legumbres(frijoles, garbanzos, lentejas)	3 a 5 veces por semana Nunca o rara vez comemos eso Todos o casi todos los días
FAMI_SITUACIONECONOMICA	Con respecto al año inmediatamente anterior, la situación económica de su hogares:	Igual Mejor Peor
ESTU_DEDICACIONLECTURADIARIA	Usualmente, ¿cuánto tiempo aldía dedica a leer por entretenimiento?	No leo por entretenimiento 30 minutos o menos Entre 30 y 60 minutos Entre 1 y 2 horas Más de 2 horas
ESTU_DEDICACIONINTERNET	Usualmente, ¿cuánto tiempo aldía dedica a navegar en internet? Excluya actividades académicas	No Navega Internet 30 minutos o menos Entre 30 y 60 minutos Entre 1 y 3 horas Más de 3 horas
ESTU_HORASSEMANATRABAJA	¿Cuántas horas trabajó usteddurante la semana pasada?	0 Menos de 10 horas Entre 11 y 20 horas Entre 21 y 30 horas Más de 30 horas
COLE_GENERO	Indica el género de la poblaciónde Establecimiento	FEMENINO MASCULINO MIXTO
COLE_NATURALEZA	Indica la naturaleza del Establecimiento	NO OFICIAL OFICIAL
COLE_BILINGUE	Indica si el Establecimiento esbilingüe o no	N – No S – Sí
COLE_CARACTER	Indica el carácter delEstablecimiento	ACADÉMICO TÉCNICO TÉCNICO/ACADÉMICO NO APLICA
COLE_SEDE_PRINCIPAL	¿Esta es la sede principal del Establecimiento Educativo?	N – No S – Sí
COLE_AREA_UBICACION	Área de ubicación de la Sede	RURAL URBANO
COLE_JORNADA	Jornada de la Sede	COMPLETA MAÑANA NOCHE SABATINA TARDE UNICA
COLE_DEPTO_UBICACION	Nombre del departamento donde está ubicada la Sede	Texto
ESTU_NSE_INDIVIDUAL	Nivel Socioeconómico del evaluado	NSE1 (numérico) NSE2 (numérico) NSE3 (numérico) NSE4 (numérico)

ESTU_NSE_ESTABLECIMIENTO	Nivel Socioeconómico del Establecimiento	Numérica
PUNT_GLOBAL	Puntaje total obtenido	Numérico

Referencia: Ministerio de Educación Nacional e Instituto Colombiano para la Evaluación de la Educación (ICFES), 2022.

Red Neuronal

Tabla 9

Comportamiento de La Red Neuronal

	Training Loss	Validation Loss	Training RMSE	Validation RMSE	Training MAE	Validation MAE
Epoch 1	48586.912	29128.5664	220.4421	170.6719	213.0501	162.9504
Epoch 2	12423.9158	2149.4334	111.479	46.3619	93.3333	36.975
Epoch 3	1884.4593	1508.8428	43.4057	38.8435	34.7104	31.2626
Epoch 4	1709.5269	1470.7219	41.3466	38.3496	33.1167	30.8244
Epoch 5	1661.3629	1454.1827	40.7523	38.1334	32.6273	30.6481
Epoch 6	1629.823	1449.6292	40.3692	38.0736	32.3265	30.5912
Epoch 7	1613.1171	1449.0694	40.1626	38.0665	32.1447	30.5593
Epoch 8	1592.5281	1443.351	39.9095	37.9912	31.9595	30.5454
Epoch 9	1583.4855	1441.8461	39.795	37.9713	31.8742	30.614
Epoch 10	1573.6182	1440.466	39.6712	37.9529	31.7675	30.5591
Epoch 11	1570.0112	1437.0077	39.6144	37.9076	31.7366	30.4494
Epoch 12	1558.7319	1436.6818	39.4781	37.9032	31.6297	30.4348
Epoch 13	1553.2696	1441.9753	39.4088	37.973	31.5697	30.5673
Epoch 14	1544.8429	1438.6907	39.3072	37.9296	31.4956	30.4232
Epoch 15	1542.6161	1436.9279	39.2782	37.9065	31.4632	30.3945
Epoch 16	1538.0391	1433.0076	39.2197	37.8547	31.4253	30.3567
Epoch 17	1534.7583	1432.187	39.1784	37.8437	31.3913	30.3981
Epoch 18	1528.1011	1439.6435	39.0897	37.942	31.3252	30.5294
Epoch 19	1525.141	1432.9179	39.0511	37.8533	31.2777	30.4284
Epoch 20	1518.997	1431.866	38.9758	37.8395	31.217	30.3939
Epoch 21	1515.4932	1433.5715	38.9306	37.8624	31.1723	30.3718
Epoch 22	1509.531	1437.5573	38.854	37.915	31.1249	30.4114
Epoch 23	1509.052	1440.4146	38.844	37.9523	31.1368	30.551
Epoch 24	1504.0927	1432.9041	38.7863	37.8533	31.0937	30.3622
Epoch 25	1500.0493	1433.9601	38.7307	37.8672	31.0363	30.3115
Epoch 26	1497.9692	1436.6665	38.706	37.903	31.0305	30.3697
Epoch 26	reducing learning rate of group 0 to 3.0000e-05.					
Epoch 27	1477.4839	1427.5693	38.4403	37.7827	30.7931	30.32
Epoch 28	1473.8431	1427.5252	38.3908	37.7821	30.7486	30.2869
Epoch 29	1471.3037	1427.4005	38.355	37.7805	30.7363	30.2942
Epoch 30	1470.3845	1428.7813	38.3434	37.7988	30.7147	30.3109
Epoch 31	1464.7885	1427.7291	38.2716	37.7848	30.6536	30.3261
Epoch 32	1466.4733	1427.7636	38.2944	37.7853	30.6845	30.3194
Epoch 33	1465.4453	1428.8078	38.2834	37.7991	30.6663	30.3123
Epoch 34	1466.8986	1428.6248	38.3054	37.7966	30.6848	30.3185
Epoch 35	1467.7764	1428.922	38.3095	37.8005	30.6853	30.3318
Epoch 35	reducing learning rate of group 0 to 3.0000e-06.					
Epoch 36	1463.5555	1428.8784	38.255	37.7999	30.6416	30.3231
Epoch 37	1462.7737	1429.108	38.2495	37.803	30.6324	30.3251
Epoch 38	1463.2555	1429.3757	38.2517	37.8065	30.6485	30.3177

Epoch 39	1461.0104	1429.2151	38.2267	37.8044	30.6204	30.3245
Epoch 40	1461.7518	1429.266	38.2283	37.8051	30.6102	30.3248
Validation loss hasn't improved in 10 epochs. Stopping early.						

Note. ***Epoch=1000***

Análisis de Importancia de Atributos

Permutation Feature Importance (PFI)

Tabla 10

Resultados PFI

Variable	Impacto en el modelo
COLE_JORNADA	-2.0848
COLE_DEPTO_UBICACION	-1.8912
ESTU_NSE_INDIVIDUAL	-0.8611
ESTU_GENERO	-0.8495
ESTU_DEDICACIONLECTURADIARIA	-0.6721
FAMI_NUMLIBROS	-0.6294
FAMI_SITUACIONECONOMICA	-0.5458
FAMI_EDUCACIONMADRE	-0.5442
FAMI_EDUCACIONPADRE	-0.4554
FAMI_ESTRATOVIVIENDA	-0.4552
ESTU_HORASSEMANATRABAJA	-0.4195
FAMI_COMELECHEDERIVADOS	-0.408
FAMI_CUARTOSHOGAR	-0.3306
COLE_NATURALEZA	-0.3119
ESTU_DEDICACIONINTERNET	-0.3044
FAMI_COMECARNEPESCADOHUEVO	-0.2715
FAMI_TRABAJOLABORMADRE	-0.2453
FAMI_TRABAJOLABORPADRE	-0.2304
ESTU_NSE_ESTABLECIMIENTO	-0.2166
FAMI_PERSONASHOGAR	-0.2074
ESTU_TIENEETNIA	-0.1955
FAMI_TIENEINTERNET	-0.1317
FAMI_COMECEREALFRUTOSLEGUMBRE	-0.1306
FAMI_TIENECOMPUTADOR	-0.1242
FAMI_TIENEMOTOCICLETA	-0.0962
COLE_CARACTER	-0.0954
COLE_AREA_UBICACION	-0.0795
FAMI_TIENESERVICIOTV	-0.0596
COLE_GENERO	-0.0518
FAMI_TIENEHORNOMICROOGAS	-0.0399
FAMI_TIENECONSOLAVIDEOJUEGOS	-0.0343
COLE_SEDE_PRINCIPAL	-0.0282
FAMI_TIENEAUTOMOVIL	-0.0245
COLE_BILINGUE	-0.0116
FAMI_TIENELAVADORA	-0.0109

Optuna

Tabla 11

Resultados Optuna

Modelo	Número iteraciones	Hiperparámetros Utilizados:	Training Loss	Validation Loss
0	Epoch 17/100	Número de capas: 1 Función de Activación 1: LeakyReLU Función de Activación 2: LeakyReLU *Hidden Layer 1: 265* *Hidden Layer 2: 81* *Dropout rate: 0.334* *Weight Decay:* 0.00021 *Batch size:* 32 *Learning rate:* 0.0032	2364.8763	1565.2167
1	Epoch 48/100	Número de capas: 2 Función de Activación 1: LeakyReLU Función de Activación 2: LeakyReLU *Hidden Layer 1:* 393 *Hidden Layer 2:* 138 *Dropout rate: 0.415* *Weight Decay: 0.000213* *Batch size:* 64 *Learning rate:* 0.0000204	2552.2559	1519.9951
2	Epoch 76/100	Número de capas: 1 Función de Activación 1: ReLU Función de Activación 2: ReLU *Hidden Layer 1: 405* *Hidden Layer 2:* 226 *Dropout rate: 0.208* *Weight Decay: 3.13137385347927E-06* *Batch size:* 64 *Learning rate:* 0.0000365	1977.842	1447.9375
3	Epoch 99/100	Número de capas: 1 Función de Activación 1: ELU Función de Activación 2: ELU *Hidden Layer 1:* 217 *Hidden Layer 2:* 174 *Dropout rate: 0.257* *Weight Decay: 0.000138* *Batch size:* 128 *Learning rate:* 0.000048	1821.7439	1438.0112
4	Epoch 55/100	Número de capas: 1 Función de Activación 1: ELU Función de Activación 2: ReLU *Hidden Layer 1: 78* *Hidden Layer 2:* 225 *Dropout rate: 0.178* *Weight Decay: 1.45646821187267E-06* *Batch size:* 128	1852.127	1439.4335

		Learning rate: 0.000103		
5	Epoch 63/100	Número de capas: 2 Función de Activación 1: ELU Función de Activación 2: ReLU *Hidden Layer 1:* 273 *Hidden Layer 2:* 253 *Dropout rate: 0.348* *Weight Decay: 5.16942414170745E-06* *Batch size:* 32, 64, 128 *Learning rate:* 0.0000182	2238.1553	1483.9757
6	Epoch 46/100	Número de capas: 1 Función de Activación 1: LeakyReLU Función de Activación 2: ReLU *Hidden Layer 1:* 90 *Hidden Layer 2:* 87 *Dropout rate: 0.1330* *Weight Decay: 3.55578800610674E-06* *Batch size:* 64 *Learning rate: 0.001114*	1464.846	1404.9254
7	Epoch 45/100	Número de capas: 3 Función de Activación 1: LeakyReLU' Función de Activación 2: ReLU *Hidden Layer 1:* 50 *Hidden Layer 2:* 52 *Dropout rate: 0.105* *Weight Decay: 0.0000339* *Batch size:* 128 *Learning rate: 0.000646*	1500.5408	1402.3572
8	Epoch 36/100	Número de capas: 2 Función de Activación 1: LeakyReLU Función de Activación 2: LeakyReLU *Hidden Layer 1:* 493 *Hidden Layer 2:* 174 *Dropout rate: 0.170* *Weight Decay: 3.96118942162264E-06* *Batch size:* 128 *Learning rate:* 0.000925	1639.1509	1426.3957
9	Epoch 41/100	Número de capas: 3 Función de Activación 1: ELU Función de Activación 2: ELU *Hidden Layer 1:* 371 *Hidden Layer 2:* 137 *Dropout rate: 0.183* *Weight Decay: 0.000410* *Batch size:* 128 *Learning rate:* 0.000869	1704.8643	1433.8892
10	Epoch 17/100	Número de capas: 1 Función de Activación 1: ReLU Función de Activación 2: ReLU *Hidden Layer 1: 136* *Hidden Layer 2: 90* *Dropout rate: 0.102* *Weight Decay: 0.0000101* *Batch size:* 32, 64 *Learning rate:* 0.009213	1786.6241	1481.6397

11	Epoch 60/100	Número de capas: 3 Función de Activación 1: LeakyReLU Función de Activación 2: ELU *Hidden Layer 1: 33* *Hidden Layer 2: 50* *Dropout rate: 0.1127* *Weight Decay: 0.0000337* *Batch size:* 32, 64 *Learning rate:* 0.000308	1518.083	1405.9135
12	Epoch 68/100	Número de capas: 3 Función de Activación 1: LeakyReLU Función de Activación 2: ELU *Hidden Layer 1: 136* *Hidden Layer 2: 42* *Dropout rate: 0.1036* *Weight Decay: 0.00003371* *Batch size:* 32, 64 *Learning rate:* 0.00021	1520.1644	1403.5144
13	Epoch 56/100	Número de capas: 3 Función de Activación 1: LeakyReLU Función de Activación 2: ELU *Hidden Layer 1: 42* *Hidden Layer 2: 94* *Dropout rate: 0.498* *Weight Decay: 0.00001859* *Batch size:* 32, 64 *Learning rate:* 0.0002433	2162.985	1418.5194
14	Epoch 30/100	Número de capas: 2 Función de Activación 1: LeakyReLU Función de Activación 2: ELU *Hidden Layer 1:* 133 *Hidden Layer 2: 64* *Dropout rate: 0.2475* *Weight Decay: 0.00006174* *Batch size:* 32, 64 *Learning rate:* 0.001798	1682.6241	1417.8785
15	Epoch 18/100	Número de capas: 2 Función de Activación 1: LeakyReLU Función de Activación 2: ELU *Hidden Layer 1:* 215 *Hidden Layer 2:* 109 *Dropout rate: 0.1399* *Weight Decay: 1.00327543041208E-06* *Batch size:* 32, 64 *Learning rate:* 0.000403	1595.6192	1427.5553
16	Epoch 49/100	Número de capas: 3 Función de Activación 1: ReLU Función de Activación 2: ELU *Hidden Layer 1:* 103 *Hidden Layer 2:* 117 *Dropout rate: 0.226* *Weight Decay:* 9.68133383479953E-06 *Batch size: 64* *Learning rate:* 0.000108	1741.0114	1432.8201

17	Epoch 14/100	Número de capas: 2 Función de Activación 1: LeakyReLU Función de Activación 2: ReLU *Hidden Layer 1:* 190 *Hidden Layer 2:* 37 *Dropout rate: 0.134* *Weight Decay: 0.000986* *Batch size:* 32 *Learning rate:* 0.00175	1771.4622	1502.7038