REVISTA TIA
- Revista TIA - Tecnología, Investigación y Academia -
Publicación Facultad de Ingeniería y Red de Investigaciones de Tecnología Avanzada - RITA

Predicción rendimiento estudiantes pruebas saber pro en pandemia junto con las características socioeconómicas

Prediction of student performance saber pro test in pandemic together with socioeconomic characteristics

Sebastian Camilo Vanegas-Ayala[1] , Daniel David Leal-Lara[2] y Julio Barón-Velandia[3]

Citar este documento: Vanegas-Ayala, S.C., Leal-Lara, D.D. y Barón-Velandia, J. (2022). Predicción rendimiento estudiantes pruebas saber pro en pandemia junto con las características socioeconómicas. Revista TIA - Tecnología, Investigación y Academia, 9(2), 5-16.

1 Ingeniero de Sistemas, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, scvanegasa@correo.udistrital.edu.co.
2 Ingeniero de Sistemas, Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, ddleall@correo.udistrital.edu.co.
3 PhD en Ingeniería Informática, Decano Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, jbaron@udistrital.edu.co.

Artículo de Investigación - Revista TIA - Tecnología, Investigación y Academia -
Fecha de recepción: 30 de noviembre de 2021 - Fecha de aceptación: 13 de diciembre de 2021
Artículo de investigación. ISSN: 2344-8288 Vol. 9 No. 2. 2021 Bogotá-Colombia

Resumen. El propósito del presente artículo de investigación es exponer los resultados obtenidos al aplicar la minería de datos al ámbito educativo de las pruebas saber pro obtenidas en épocas de pandemia del año 2020 con el objeto de determinar las variables más influyentes en el desempeño de los estudiantes y como se vieron afectadas por la pandemia. Para tal efecto se utilizaron los resultados obtenidos en el año 2019 y 2020, aplicando la metodología CRISP-DM, la cual es un referente en el ámbito de minería de datos. Se realizó un proceso de selección inicial de atributos teniendo en cuenta aquellos que en la literatura los autores encontraron relevantes tales como la información socioeconómica, hábitos de estudio, entre otros. Se procedió a limpiar y transformar en un repositorio de datos con los resultados obtenidos en estas pruebas y se aplicó la técnica de predicción basada redes neuronales profundas y regresión lineal para evidenciar el comportamiento a nivel de predicción de los atributos seleccionados. Se obtuvo que los atributos que influyen de manera importante en los procesos de predicción son los asociados a la educación de los padres y su estado laboral, evidenciando su afectación por la pandemia.

Palabras clave. COVID-19; Desempeño Académico; Minería de Datos Educativa; Pruebas Saber Pro; Redes Neuronales; Regresión Lineal.

Abstract. The purpose of this research is to present the results obtained when applying data mining to the educational field of the saber pro test obtained in times of pandemic (2020) in order to determine the most influential variables in the performance of the students and how they were affected by the pandemic. For this purpose, the results obtained in 2019 and 2020 were used, applying the CRISP-DM methodology, which is a benchmark in the field of data mining. An initial selection process of attributes was carried out taking into account those that the authors found relevant in the literature, such as socioeconomic information, study habits, among others. We proceeded to clean and transform a data repository with the results obtained in these tests and the prediction technique based on deep neural networks and linear regression was applied to show the behavior at the prediction level of the selected attributes. It was found that the attributes that significantly influence the prediction processes are those associated with the parents' education and their employment status, evidencing their impact by the pandemic.

Key Words. COVID-19; Academic Performance; Educational Data Mining; Saber Pro tests; Neural Networks; Linear Regression.

Introducción

La educación es considerada por diversos referentes a nivel mundial como uno de los procesos más importantes a tener en cuenta en el desarrollo de los países dado que mediante la educación se forman personas que adquirirán diferentes roles en el país. De manera la educación es de esencial importancia para los gobiernos tanto en el poder como entrantes(Marly Johana Bahamón & Ruiz, 2014; Rodríguez Rosero et al., 2021). Específicamente en Colombia, para conocer el nivel de calidad de la educación impartida, en sus niveles diversos niveles, y a su vez, establecer el estado de la misma en cada una de las diversas etapas de los ciudadanos, se crea el Instituto Colombiano para el Fomento de la Educación Superior (ICFES), encargado de la aplicación de pruebas estandarizadas que permiten medir la calidad de la educación (Koretz & Langi, 2018). Lo anterior a su vez es fundamentado en la Ley 30 de 1992, en la cual se establece los mecanismos de evaluación de la calidad de los programas académicos de las instituciones de Educación Superior (Marly Johana Bahamón & Ruiz, 2014; Rodríguez Rosero et al., 2021; Timarán Pereira et al., 2016).

Los criterios de evaluación de la calidad de la educación permiten establecer criterios de mejoramiento de la calidad de la misma. Para esto es necesario analizar, comprender y predecir el efecto de diversos factores que varían el desempeño de los estudiantes en las Pruebas Saber Pro, de tal manera que se puedan diseñar planes y estrategias de política social que permitan apoyar la mejora de estos aspectos (Poh & Smythe, 2015; Rodríguez Rosero et al., 2021). Diversos autores han podido establecer una relación entre la variación del rendimiento académico de estas pruebas y diversos factores como el estrato socioeconómico, el medio ambiente, el acceso a herramientas tecnológicas de aprendizaje como computadoras y conexión a internet, el nivel educativo de los padres de familia, el ingreso, entre otros (García-González & Skrita, 2019; Ospina, 2019; Poh & Smythe, 2015; Rodríguez Rosero et al., 2021). Adicionalmente se ha encontrado que los estudiantes que provienen de familias de altos ingresos se le asocian generalmente los mejores resultados. Esto debido que al disponer de recursos como lo son los libros, herramientas de comunicación e información; en el entorno familiar generan un impacto positivo en las experiencias de aprendizaje del estudiante (García-González & Skrita, 2019).

Entre los diversos resultados hallados en la revisión bibliográfica, se encuentran los obtenidos por Garcia-Gonzales y Skrita en (García-González & Skrita, 2019), quienes establecieron que las variables que permiten predecir mejor el rendimiento de los estudiantes son los relacionados con el nivel educativo de la madre, el estrato socioeconómico del hogar, el nivel educativo del padre y el acceso a dispositivos electrónicos. Contrario a esto, los autores Oliveo Carrascal y Jiménez Giraldo en (Oviedo Carrascal & Jiménez Giraldo, 2019) establecieron que el género, la acreditación de la institución educativa y el ingreso de los familiares son los factores más relevantes para predecir desempeño de las pruebas Saber Pro. Por último, otros autores como Rodriguez y Correa en (Rodríguez & Correa, 2018) y de Bahamón y Ruiz en (Marly Johana Bahamón & Ruiz, 2014) determinan que el municipio donde residen los estudiantes, el tipo de institución y el estrato socioeconómico para entender como estos factores repercuten en su rendimiento.

La pandemia causada por COVID-19 ha provocado una crisis sin precedentes. En la educación, esta emergencia ha dado lugar al cierre masivo de las actividades presenciales de instituciones educativas en más de 190 países con el fin de evitar la propagación del virus y mitigar su impacto (Comisión Económica para América Latina y el Caribe (CEPAL); Oficina Regional de Educación para América Latina y el Caribe, 2020). Según datos de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), a mediados de mayo de 2020 más de 1.200 millones de estudiantes de todos los niveles de enseñanza, en todo el mundo, habían dejado de tener clases presenciales en la escuela. De ellos, más de 160 millones eran estudiantes de América Latina y el Caribe. (Comisión Económica para América Latina y el Caribe (CEPAL); Oficina Regional de Educación para América Latina y el Caribe, 2020). Esto a su vez ha aumentado la desigualdad en el acceso a oportunidades educativas por la vía digital dificultando la socialización y la inclusión en general.

La presente investigación pretende proporcionar una visión de las variables más representativas en el momento de una predicción del rendimiento de los estudiantes que presentan las pruebas Saber Pro y cuál ha sido su afectación por la pandemia causada por COVID-19. El presente artículo se divide en 5 secciones, las cuales son introducción, metodología, resultados, discusión y finalmente conclusiones.

Metodología

La metodología del presente proyecto se basa en la denominada CRISP-DM, la cual es una de las más ampliamente utilizadas en el desarrollo de proyectos relacionados a la minería de datos (Timarán-pereira et al., 2020). La adopción de esta metodología para el presente proyecto de investigación permitirá encontrar patrones relacionados al desempeño académico en las pruebas Icfes Saber Pro en la pandemia

causada por el virus COVID-19. La descripción de cada uno de los pasos elaborados bajo la metodologia CRISP-DM son los siguientes:

• *Análisis del problema:* En esta etapa se define claramente el objeto del estudio y los objetivos a alcázar con el mismo.

• *Comprensión de los datos:* Durante esta etapa se busca y recolecta los datos a utilizar en el presente proyecto. Para tal efecto se utilizaron los resultados obtenidos en el año 2019 y 2020 de las pruebas Icfes Saber Pro.

• *Preparación de los datos:* En esta etapa se hace el proceso de transformación de los datos entre lo cual se encuentra la generación de variables adicionales, integración de diferentes orígenes de datos, eliminación de valores anómalos y la reducción de los variables, basado en lo abordado en (Ruiz Escorcia et al., 2018).

• *Modelado:* Para la elaboración de los modelos se establecieron dos técnicas diferentes reconocidas para la predicción las cuales son: redes neuronales y la regresión lineal.

• *Evaluación:* Con el fin de evaluar los resultados de los modelos, se establece esta etapa que permitirá encontrar posibles errores que puedan afectar el comportamiento general del modelo.

• *Despliegue:* Finalmente los hallazgos de las anteriores fases se extraen con el objetivo de que este pueda ser transmitido y socializado con la comunidad científica.

Herramienta tecnológica usada:
Teniendo en cuenta la necesidad de optar por una herramienta software que permita desarrollar a plenitud los pasos establecidos en la metodología de CRISP-DM, se opta por manejar el software de Orange, desarrollado por la Universidad de Ljubljana (Orange, 2021).

Resultados

Conjunto de datos
Para la predicción del rendimiento de los participantes en las pruebas saber pro se utilizan dos conjuntos de datos para definir las características que permiten una predicción, estos dos conjuntos se describen a continuación:

El conjunto de datos principal se compone de 246.436 registros de los resultados para los 5 módulos de competencias genéricas de las pruebas Saber Pro para el año 2020, los cuales contienen 97 atributos de cada registro, sobre la información personal, socioeconómica, de la IES y datos académicos de cada participante.

El conjunto de datos secundario contiene 378 registros de las instituciones de educación superior en Colombia con 24 atributos entre los que se encuentran la información geográfica, de acreditación y financiación.

Preparación de datos
Se inicia con la selección de atributos para cada uno de los conjuntos de datos procurando reducir su dimensionalidad y dejando solo las características relevantes del estudio, para esto se analizaron los resultados alcanzados en otras investigaciones acerca de la relevancia de los diversos factores en el rendimiento de los estudiantes en las pruebas de carácter similar al Saber Pro. Partiendo de la información

obtenida en el análisis de los referentes bibliográficos se seleccionan 26 características para el conjunto de datos principal y 2 para el conjunto de datos secundario, como se describe en la Tabla 1, señalando el nombre de la característica en el conjunto de datos y el tipo de dato según la tipología de Orange.

Tabla 1. Características seleccionadas conjuntos de datos.

Conjunto de datos	Nombre en el archivo de datos	Tipo
Principal	ESTU_GENERO	Categórico
Principal	ESTU_FECHANACIMIENTO	Fecha
Principal	ESTU_VALORMATRICULAUNIVERSIDAD	Categórico
Principal	ESTU_PAGOMATRICULABECA	Categórico
Principal	ESTU_PAGOMATRICULACREDITO	Categórico
Principal	ESTU_PAGOMATRICULAPADRES	Categórico
Principal	ESTU_PAGOMATRICULAPROPIO	Categórico
Principal	ESTU_COMOCAPACITOEXAMENSB11	Categórico
Principal	FAMI_EDUCACIONPADRE	Categórico
Principal	FAMI_EDUCACIONMADRE	Categórico
Principal	FAMI_ESTRATOVIVIENDA	Categórico
Principal	FAMI_TIENEINTERNET	Categórico
Principal	FAMI_TIENECOMPUTADOR	Categórico
Principal	FAMI_TIENELAVADORA	Categórico
Principal	FAMI_TIENEHORNOMICROOGAS	Categórico
Principal	FAMI_TIENESERVICIOTV	Categórico
Principal	FAMI_TIENEAUTOMOVIL	Categórico
Principal	FAMI_TIENEMOTOCICLETA	Categórico
Principal	FAMI_TRABAJOLABORPADRE	Categórico
Principal	FAMI_TRABAJOLABORMADRE	Categórico
Principal	ESTU_HORASSEMANATRABAJA	Categórico
Principal	ESTU_METODO_PRGM	Categórico
Principal	ESTU_INST_MUNICIPIO	Categórico
Principal	INST_CARACTER_ACADEMICO	Categórico
Principal	ESTU_AREARESIDE	Categórico
Principal	ESTU_INSE_INDIVIDUAL	Numérico
Secundario	ACREDITADA_ALTA_CALIDAD	Categórico
Secundario	SECTOR	Categórico

Los datos son enviados a un preprocesamiento donde se realiza la eliminación de registros con valores faltantes que equivalen alrededor del 1.5% de los datos, y a continuación se realiza una discretización de los campos ESTU_FECHANACIMIENTO y ESTU_INSE_INDIVIDUAL usando el método de discretización de igual frecuencia (Equal frequency discretization) y finalmente las clases numéricas son normalizadas entre 0 y 1 tomando los valores máximo y mínimo del atributo.

Después se realiza una búsqueda y eliminación de registros con valores atípicos usando el método de bosques aleatorios (Isolation Forest), para esto se define una contaminación del 5% obteniendo un total de 184.553 registros. Teniendo los registros con valores anómalos y atípicos eliminados se realiza una selección de atributos a partir de las 11 clases objetivo definidas para la predicción del rendimiento, el

método de selección utilizado se muestra en la Tabla 2 donde se relaciona la clase objetivo junto con el método utilizado y el tipo de dato según la tipología de Orange.

Tabla 2. Selección de atributos basado en el atributo de clase.

N	Atributo clase	Método de selección	Tipo
1	MOD_RAZONA_CUANTITAT_PUNT	RReliefF	Numérico
2	MOD_RAZONA_CUANTITAT_DESEM	RReliefF	Numérico
3	MOD_LECTURA_CRITICA_PUNT	RReliefF	Numérico
4	MOD_LECTURA_CRITICA_DESEM	RReliefF	Numérico
5	MOD_COMPETEN_CIUDADA_PUNT	RReliefF	Numérico
6	MOD_COMPETEN_CIUDADA_DESEM	RReliefF	Numérico
7	MOD_INGLES_PUNT	RReliefF	Numérico
8	MOD_INGLES_DESEM	Fast Correlation Based Filter	Categórico
9	MOD_COMUNI_ESCRITA_PUNT	RReliefF	Numérico
10	MOD_COMUNI_ESCRITA_DESEM	RReliefF	Numérico
11	PUNT_GLOBAL	RReliefF	Numérico

Se realiza una selección de atributos RReliefF para los conjuntos de datos donde el atributo de clase es de tipo numérico, por otra parte la selección de atributos para los conjuntos de datos donde el atributo de clase es de tipo categórico se usa el método Fast Correlation Based Filter.

Se establece un total de 15 atributos relevantes seleccionados para cada clase, como se muestra en la Tabla 3, donde se indica si el atributo es usado para la predicción de la clase, numeradas de 1 a 11, según la Tabla 2, indicando el valor de relevancia que se obtuvo al aplicar el método de selección, siendo 1 el más relevante y 15 el menos relevante, este valor se promedia en cada una de las clases y junto con las apariciones se establece un Ranking, definido como la división del número de apariciones sobre el promedio, donde los atributos con mayor Ranking representan los más significativos para la predicción de los atributos de clase.

Tabla 3. Atributos seleccionados por clase.

N°	Atributo seleccionado	1	2	3	4	5	6	7	8	9	10	11	Promedio	Apariciones	Ranking
1	FAMI_TRABAJOLABORPADRE	1	1	1	1	1	1	1		1	1	1	1,0	10	10,0
2	ESTU_INST_MUNICIPIO	5	2	2	6	2	4	5		2	5	3	3,6	10	2,8
3	FAMI_TRABAJOLABORMADRE	4	7	5	3	3	5	2		6	2	2	3,9	10	2,6
4	FAMI_EDUCACIONPADRE	2	3	3	2	5	3	3	12	5	6	5	4,5	11	2,5
5	FAMI_EDUCACIONMADRE	3	4	6	4	4	2	6	10	4	4	6	4,8	11	2,3
6	ESTU_HORASSEMANATRABAJA	7	6	4	5	6	6	4		3	3	4	4,8	10	2,1
7	ESTU_VALORMATRICULAUNIVERSIDAD	6	5	8	7	7	8	9	9	7	7	8	7,4	11	1,5
8	ESTU_FECHANACIMIENTO	8	8	9	8	9	7	8	3	8	8	9	7,7	11	1,4
9	FAMI_ESTRATOVIVIENDA	9	9	7	9	8	9	7	8	9	9	7	8,3	11	1,3
10	ESTU_COMOCAPACITOEXAMENSB11	10	10	10	10	10	10	10		10	10	10	10,0	10	1,0
11	ESTU_GENERO	11	12	11	11	15	11	12	6	12	11	12	11,3	11	1,0
12	ESTU_INSE_INDIVIDUAL	13	13	13	12	14	12	13	1	13		11	11,5	10	0,9
13	FAMI_TIENEMOTOCICLETA	12	11	14	14		13	11	4	11	12		11,3	9	0,8
14	ESTU_PAGOMATRICULACREDITO					11	14	14	7		15	14	12,5	6	0,5
15	ACREDITADA_ALTA_CALIDAD			12					2				7,0	2	0,3

16	FAMI_TIENEHORNOMICROOGAS		15	15	13				14				14,3	4	0,3
17	ESTU_METODO_PRGM								11		13	15	13,0	3	0,2
18	INST_CARACTER_ACADEMICO							15	13		14		14,0	3	0,2
19	FAMI_TIENEAUTOMOVIL		14		15				15				14,7	3	0,2
20	ESTU_AREARESIDE						15		5				10,0	2	0,2
21	ESTU_PAGOMATRICULAPROPIO					12						13	12,5	2	0,2
22	FAMI_TIENESERVICIOTV	14								14			14,0	2	0,1
23	ESTU_PAGOMATRICULAPADRES	15				13							14,0	2	0,1
24	ESTU_PAGOMATRICULABECA									15			15,0	1	0,1
25	FAMI_TIENEINTERNET														
26	FAMI_TIENELAVADORA														
27	FAMI_TIENECOMPUTADOR														
28	SECTOR														

Modelo de predicción

Para el desarrollo del modelo de predicción se utilizaron dos técnicas las redes neuronales y la regresión lineal. Estas técnicas fueron configuradas de la siguiente manera:

La Red Neuronal se configuro con 3 capas ocultas de 32, 16 y 4 neuronas para de esta manera, simular el comportamiento de los datos de entrada para llegar a una única salida. Como algoritmo de optimización se escogió al algoritmo de Adam el cual una extensión del descenso de gradiente estocástico y como función de activación se definió Relu. A partir de esto, y teniendo en cuenta que se tienen dos tipos de resultado en los atributos clase, los cuales son valores entre 0 y 500 para el puntaje, y valores de 1 a 4 para el cuartil; se definen dos subconfiguraciones de las redes neuronales de la siguiente manera: se define un total de 200 iteraciones para aquellos atributos con valores entre 0 y 500 y 100 iteraciones de entrenamiento para los otros atributos.

Para la Regresion Lineal se utilizó la regresión de Lasso (L1), buscando de esta manera bajar la complejidad del modelo buscan do que la solución sea más sencilla de interpretar al tomar como irrelevantes ciertas características. La configuración de alpha se definió en 0.0001.

Se realiza el entrenamiento de los modelos propuestos de predicción con el 70% de los datos, los resultados obtenidos para cada atributo de clase a predecir se muestran en la Tabla 4, donde se relacionan los dos modelos usados para predecir el atributo de clase identificado según la Tabla 2, junto con su error cuadrático medio (MSE), raíz de la desviación cuadrática media (RMSE) y error absoluto medio (MAE) para medir su precisión en la predicción.

Tabla 4. Resultados precisión en entrenamiento de los modelos.

ATRIBUTO DE CLASE	MODELO	MSE	RMSE	MAE
1	Red neuronal	0,009	0,092	0,073
	Regresión lineal	0,009	0,095	0,075
2	Red neuronal	0,066	0,257	0,214
	Regresión lineal	0,069	0,263	0,217
3	Red neuronal	0,008	0,089	0,071
	Regresión lineal	0,008	0,090	0,072
4	Red neuronal	0,063	0,251	0,204
	Regresión lineal	0,065	0,255	0,209

5	Red neuronal	0,009	0,094	0,073
	Regresión lineal	0,009	0,095	0,074
6	Red neuronal	0,067	0,259	0,213
	Regresión lineal	0,066	0,257	0,212
7	Red neuronal	0,008	0,087	0,066
	Regresión lineal	0,008	0,089	0,069
9	Red neuronal	0,016	0,128	0,097
	Regresión lineal	0,017	0,129	0,099
10	Red neuronal	0,073	0,269	0,215
	Regresión lineal	0,074	0,273	0,223
11	Red neuronal	0,007	0,086	0,068
	Regresión lineal	0,008	0,087	0,070

A partir de los resultados obtenidos se comparan el promedio de los valores MSE, RMSE y MAE entre los dos modelos usados para predicción, tanto para los atributos de clase que representan un puntaje o los que muestran el desempeño por cuartil, como se muestra en la Tabla 5, donde la mayor precisión se obtuvo usando redes neuronales en ambos escenarios.

Tabla 5. Promedio resultados precisión en entrenamiento de los modelos.

TIPO ATRIBUTO DE CLASE	MODELO	MSE	RMSE	MAE
PUNT	Red neuronal	0,010	0,096	0,075
	Regresión lineal	0,010	0,098	0,077
DESEM	Red neuronal	0,067	0,259	0,212
	Regresión lineal	0,069	0,262	0,215

Posteriormente al entrenamiento los modelos son evaluados en la fase de prueba con el 30% de los datos restantes, obteniéndose el MSE, RMSE, MAE de la Tabla 6, para los modelos de red neuronal y regresión lineal de los atributos de clase numéricos.

Tabla 6. Resultados precisión en prueba de los modelos.

ATRIBUTO DE CLASE	MODELO	MSE	RMSE	MAE
1	Red neuronal	0,009	0,092	0,073
	Regresión lineal	0,009	0,095	0,075
2	Red neuronal	0,067	0,258	0,214
	Regresión lineal	0,069	0,263	0,217
3	Red neuronal	0,008	0,089	0,071
	Regresión lineal	0,008	0,090	0,072
4	Red neuronal	0,063	0,250	0,203
	Regresión lineal	0,065	0,254	0,209
5	Red neuronal	0,009	0,093	0,073
	Regresión lineal	0,009	0,095	0,074
6	Red neuronal	0,067	0,259	0,213
	Regresión lineal	0,066	0,257	0,212

7	Red neuronal	0,008	0,087	0,067
	Regresión lineal	0,008	0,089	0,069
9	Red neuronal	0,016	0,128	0,097
	Regresión lineal	0,017	0,129	0,099
10	Red neuronal	0,073	0,270	0,215
	Regresión lineal	0,074	0,272	0,222
11	Red neuronal	0,007	0,085	0,068
	Regresión lineal	0,008	0,087	0,069

El promedio de los valores MSE, RMSE y MAE, se compara en la Tabla 7, para los modelos de red neuronal y regresión lineal en los dos escenarios de puntaje evaluados, obteniéndose en ambos casos con mayor precisión al modelo basado en redes neuronales.

Tabla 7. Promedio resultados precisión en prueba de los modelos.

TIPO ATRIBUTO DE CLASE	MODELO	MSE	RMSE	MAE
PUNT	Red neuronal	0,010	0,096	0,075
	Regresión lineal	0,010	0,098	0,076
DESEM	Red neuronal	0,068	0,259	0,211
	Regresión lineal	0,069	0,262	0,215

Por otra parte para la variable de tipo categórico MOD_INGLES_DESEM que clasifica el desempeño del estudiante en un nivel de inglés -A1, A1, A2, B1, B2, la predicción se realiza bajo modelos que permiten valores categóricos como lo es las redes neuronales, en este caso en la Tabla 9, se muestra los valores de exactitud (AUC), puntaje de clasificación de precisión (CA), calidad (Precisión), cantidad de identificación correcta (Recall) y combinación de precisión y recall (F1), obtenidos para las fases de entrenamiento y prueba del modelo.

Tabla 8. Resultados precisión en entrenamiento y prueba de MOD_INGLES_DESEM

VALORES	ENTRENAMIENTO	PRUEBA
AUC	13,842	-0,872
CA	0,402	0,404
F1	0,379	0,381
Precision	0,423	0,434
Recall	0,402	0,404

Discusión

En el proceso de selección de los atributos que determinan en mayor grado los atributos de clase, que miden mediante el puntaje en diferentes competencias el desempeño y calidad de la educación de un evaluado se encontraron que entre las más relevantes se encuentran el trabajo y la educación de los padres, por ello es importante diseñar políticas que permitan garantizar unos mínimos de educación a la población y así mismo un nivel de ingresos. Además, se vio que el municipio de donde es la institución de educación superior se estableció como un indicador de rendimiento en pandemia, por lo que este impacta directamente en la conectividad de los estudiantes.

El aspecto de educación y trabajo está altamente relacionado junto con los ingresos que requiere una familia de un estudiante, es por ello que otros factores que inciden en su rendimiento son el valor de la matricula, el estrato, el índice socioeconómico y el tiempo que requiere el estudiante trabajar para garantizar esos ingresos.

Otro aspecto importante que influye en el desempeño es la edad que se vuelve un atributo importante en el proceso de predicción encontrándose que esta incide en el nivel de ingresos, en el acceso a la educación y en pandemia al manejo de medios tecnológicos de manera más sencilla.

Teniendo lo anterior en cuenta se deben retomar propuestas por parte en primera instancia de las Instituciones de Educación Superior, como sistemas de apoyo de bienestar estudiantil (Marly Johana Bahamón & Ruiz, 2014), en coordinación con programas de riesgo académico, y políticas gubernamentales que se concentren en garantizar unos mínimos básicos en los atributos mencionados, lo cual mejorara el rendimiento de los estudiantes y por ende una mejor competitividad a nivel internacional.

Por otra parte, la pandemia obligo a mejorar las competencias con herramientas tecnológicas o implicando técnicas de estudio más sofisticadas que involucren las TIC, pero también deben acompañarse de políticas sociales que garanticen que el estudiante pueda obtener los mínimos en su condición social, todo esto acompañado de un sistema de evaluación que además de brindar información del aprendizaje determine en qué condiciones se aprende (Timarán Pereira et al., 2016), lo cual viene altamente influido del entorno familiar y social del estudiante.

Desde la perspectiva de los resultados obtenidos se observa que tanto los modelos de predicción basados en redes neuronales y regresión lineal obtienen resultados con una alta precisión, lo cual los hace una solución factible para el problema propuesto, adicionalmente las redes neuronales presentan un mejor rendimiento, el cual puede verse incrementado por configuraciones que impliquen un mayor número de neuronas o una disposición diferente.

Por último, se puede observar que los errores presentados son similares en los 10 atributos numéricos lo cual tiene una relación con la homogeneidad de los atributos seleccionados en la predicción y su nivel de relevancia en cada una de las clases, presentándose comportamientos similares basados en la semejanza de la selección y relevancia de los atributos que predicen la clase.

Conclusiones

Se pueden caracterizar diferentes factores socioeconómicos asociados al rendimiento de los estudiantes de instituciones de educación superior técnica, tecnológica y universitaria en los diferentes componentes de la prueba ICFES Saber Pro los cuales permiten realizar una predicción con un MSE menor que 0,074 para todos los casos. En el marco de la pandemia se observa que características como la tenencia de internet o de computador, se vuelven obligatorias, por lo cual no determinan los resultados del modelo de predicción, ya que se presupone que se tienen estas dos características, además el lugar geográfico donde se ubica la institución de educación superior cobra gran relevancia para la predicción de los valores.

Los resultados obtenidos del entrenamiento y prueba de los diversos modelos obtenidos mediante redes neuronales y regresión lineal evidencian que las redes neuronales en general, obtienen un mejor rendimiento, con menor error sobre la regresión lineal.

Los modelos generados muestran que las características comunes a todos los atributos de clase y que a su vez influyen de manera importante en el proceso de predicción son los asociados con la educación de los padres y su estado laboral, por lo cual se debe promover el acompañamiento hacia los padres en el proceso educativo y de enseñanza de sus hijos, volviéndose vital para el desarrollo de actitudes y capacidades a lo largo de la vida.

Referencias bibliográficas

Comisión Económica para América Latina y el Caribe (CEPAL); Oficina Regional de Educación para América Latina y el Caribe. (2020). Educación en tiempos de pandemia (covid-19). Revista Universidad de La Salle, 1(85), 51–59. https://doi.org/10.19052/ruls.vol1.iss85.4

García-González, J. D., & Skrita, A. (2019). Predicting academic performance based on students' family environment: Evidence for Colombia using classification trees. Psychology, Society and Education, 11(3), 299–311. https://doi.org/10.25115/psye.v11i3.2056

Koretz, D., & Langi, M. (2018). Predicting Freshman Grade-Point Average from Test Scores: Effects of Variation Within and Between High Schools. Educational Measurement: Issues and Practice, 37(2), 9–19. https://doi.org/10.1111/emip.12173

Marly Johana Bahamón, M., & Ruiz, L. R. (2014). Characterization of the intellectual ability, Sociodemographic and academic factors of students with high and low performances in the Saber Pro exam – 2012. Avances En Psicologia Latinoamericana, 32(3), 459–476. https://doi.org/10.12804/apl32.03.2014.01

Orange. (2021). Minería de datos de Orange - Minería de datos.

Ospina, D. R. (2019). Relaciones de Clase en el Sistema Universitario y su Efecto sobre el Rendimiento Académico: El Caso de Bogotá. Multidisciplinary Journal of Educational Research, 9(1), 1–24.

Oviedo Carrascal, A. I., & Jiménez Giraldo, J. (2019). Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO. Revista Politécnica, 15(29), 128–140. https://doi.org/10.33571/rpolitec.v15n29a10

Poh, N., & Smythe, I. (2015). To what extend can we predict students' performance? A case study in colleges in South Africa. IEEE SSCI 2014 - 2014 IEEE Symposium Series on Computational Intelligence - CIDM 2014: 2014 IEEE Symposium on Computational Intelligence and Data Mining, Proceedings, 416–421. https://doi.org/10.1109/CIDM.2014.7008698

Rodríguez, M., & Correa, J. (2018). Impacto del contexto municipal sobre el desempeño académico individual. Lecturas de Economía, 90(90), 159–193. https://doi.org/10.17533/udea.le.n90a06

Rodríguez Rosero, D. D., Ordoñez Ortega, R. E., & Hidalgo-Villota, M. E. (2021). Determinantes del rendimiento académico de la educación media en el departamento de Nariño, Colombia. Lecturas de Economía, 94, 87–126. https://doi.org/10.17533/udea.le.n94a341834

Ruiz Escorcia, R. R., Arévalo Medrano, J. B., Morillo, G. P., & Acosta-Humánez, P. B. (2018). Análisis de componentes principales aplicado a la prueba estatal Colombiana Saber 11. Espacios, 39(10).

Timarán-pereira, R., Hidalgo-troya, A., & Vidal-alegría, F. (2020). Una Mirada al Desempeño Académico en las Pruebas Saber Pro de los Estudiantes de Ingeniería desde la Minería de Datos Educativa. 29–43.

Timarán Pereira, R., Vidal Alegria, F. A., & Solís Flórez, D. (2016). Identificación de Patrones de Rendimiento 2012-2014, en las Competencias Lectura Crítica Académico en las Pruebas Saber Pro entre

y Comunicación Escrita con Técnicas Predictivas de Minería de Datos. Descubrimiento de Patrones de Desempeño Académico Con Árboles de Decisión En Las Competencias Genéricas de La Formación Profesional, 51–64. https://doi.org/10.16925/9789587600490