LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

MODELO DE REGRESIÓN PARA PREDECIR EL PUNTAJE ESPERADO EN LA SABER 11 PARA LOS ESTUDIANTES DEL COLEGIO CAJASAI

REGRESSION MODEL TO PREDICT THE EXPECTED SCORE IN SABER 11 FOR CAJASAI SCHOOL STUDENTS

John Michael Cardozo Anaya, jmcardozoa@libertadores.edu.co,

José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co

RESUMEN

En la actualidad los estudiantes que aspiran ser profesionales tienen sus esperanzas en los resultados que puedan obtener en las pruebas Saber 11, porque a través de estos es posible aspirar a obtener una beca, un crédito condenable o en su defecto un cupo para ser admitido en alguna institución de carácter superior. De igual manera, las instituciones educativas al lograr que sus estudiantes obtengan puntajes altos, les permite obtener beneficios como acreditación con índices de calidad, estar ubicada en un mejor escalafón y sobre todo el prestigio en el entorno social. Es aquí donde se centra el propósito de construir un modelo de regresión que permita establecer a futuro el puntaje que puedan obtener los estudiantes de la isla de San Andrés pertenecientes al Colegio Cajasai. A partir de los datos suministrados por una empresa educativa (Edúcate) externa al colegio, correspondientes a los resultados de los simulacros realizados durante el año 2022 y de las pruebas Saber 11 del mismo año, se

adiestraron diversos modelos de regresión utilizando aprendizajes automatizados supervisados para predecir el puntaje de los estudiantes de San Andrés. El modelo de regresión con el mayor rendimiento en los datos de testeo fue el Huber con un RSME de 26.2683. Los resultados obtenidos permitirán determinar el puntaje que puede obtener un estudiante al presentar la prueba Saber 11, a partir de los puntajes alcanzados en los simulacros que realiza la empresa asesora.

Palabras clave: Regresión, Correlación, Aprendizaje automatizado, Estudiantes, Saber 11.

ABSTRACT

At present, students who aspire to be professionals have their hopes in the results they can obtain in the Saber 11 tests, because through these it is possible to aspire to obtain a scholarship, a condemnable loan or, failing that, a quota to be admitted in some higher institution. In the same way, educational institutions, by getting their students to obtain high scores, allow them to obtain benefits such as accreditation with quality indices, being located in a better ranking and, above all, prestige in the social environment. It is here where the purpose of building a regression model that allows establishing in the future the score that students from the island of San Andrés belonging to Colegio CAJASAI can be obtained is focused. Based on the data provided by an educational company (Educate) external to the school, corresponding to the results of the drills carried out during the year 2022 and the Saber 11 tests of the same year, various regression models were trained using supervised automated learning to predict the score of the students of San Andrés. The regression model with the highest performance on the test data was the Huber with an RSME of 26.2683. The results obtained will make it possible to determine the score that a student can obtain when taking the Saber 11 test, based on the scores achieved in the drills carried out by the consulting company.

Keywords: Regression, Correlation, Machine Learning, Student, Saber 11.

INTRODUCCIÓN

En la actualidad los estudiantes que aspiran ser profesionales tienen sus esperanzas en los resultados que puedan obtener en las pruebas Saber 11, porque a través de estos es posible aspirar a obtener una beca, un crédito condenable o en su defecto un cupo para ser admitido en alguna institución de carácter superior.

Para centrar el problema, es necesario describir brevemente el contexto institucional del colegio: El colegio CAJASAI, el cual lleva el nombre del acrónimo Caja de Compensación Familiar de la Cámara de Comercio de San Andrés, Providencia y Santa Catalina. Es una institución de carácter privado mixto en el que la población estudiantil pertenece a todos los estratos socioeconómicos y acoge a estudiantes que presenten alguna necesidad cognitiva, visual, auditiva o motriz.

El Proyecto Educativo Institucional se concibe como un proceso permanente y sistemático de reflexión pedagógica, el cual se centra en el constructivismo, que tiene como fin ofrecer una educación de alta calidad en donde el estudiante construya su propio aprendizaje. Las áreas básicas se vincula al proyecto centrando el aprendizaje escolar, en el hacer actividad matemática y consolidando comunidades en las que los estudiantes aprendan a pensar de tal manera que construyan conocimiento relevante y útil para el abordaje y solución de situaciones problema en contextos propios de la disciplina, otras disciplinas y la vida cotidiana (PEI, 2022).

Todas las áreas cuentan con un equipo externo de asesores los cuáles provee diferentes tipos de materiales (videos, talleres, cartillas) para su implementación en el aula enfocados en el modelo de evaluación basado en evidencias. Este modelo de evaluación en la clase de se

implementa en los siguientes momentos: exploración (anclaje), estructuración, práctica o ejecución, transferencia y valoración. Adicional a lo anterior, la empresa asesora por año académico, aplica simulacros tipo Saber 11 desde los grados noveno a undécimo, estos consisten en tres simulacros estilo Saber 11, con el primero verifica el estado inicial del estudiante para implementar estrategias pedagógicas; con el segundo determina el alcance de las estrategias implementadas y con el último valida el estado final del estudiante en cuánto a las competencias establecidas por el Instituto Colombiano para la Evaluación de la Educación (ICFES) para la Saber 11.

Actualmente la Saber 11 evalúa cinco áreas básicas las cuales son (Lectura Crítica, Matemáticas, Ciencias Sociales, Ciencias Naturales e Inglés), cada una de estás enmarcadas en competencias. Los puntajes posibles para obtener por área van de 0-100 y global de 0 – 500 puntos, lo cuál determina que tan competente es el estudiante para la sociedad.

Dentro de los avances de la literatura, se encuentran pocos estudios a nivel nacional que midan el nivel del estudiante antes de presentar la Saber 11, debido a la limitante de datos para evaluar a los estudiantes a lo largo del tiempo. Se encontraron diferentes estudios, incluyendo algunos para la ciudad de Bogotá (Serrano Leon, et al., 2020; Rodriguez Revilla, 2015; Campos, et al., 2017; Casas et. al, 2002). Todos estos se han enfocados en medir el valor agregado de la educación superior, haciendo diferencias entre estos estudios por IES (Institución de Educación Superior), programa o ciudad, haciendo uso de los exámenes del ICFES Saber 11 y Saber Pro a partir de distintas metodologías, donde algunos utilizaron Mínimos Cuadrados Ordinarios y variables instrumentales pero la gran mayoría se centró en el modelo lineal jerárquico.

Es aquí donde se centra el propósito de construir un modelo de regresión que permita establecer a futuro el puntaje que puedan obtener los estudiantes de la isla de San Andrés

pertenecientes al Colegio Cajasai, a partir de los datos suministrados por la empresa educativa (Edúcate) externa al colegio, correspondientes a los resultados de los simulacros realizados durante el año 2022 y de las pruebas Saber 11 del mismo año.

METODOLOGÍA

Datos

El presente estudio se desarrolló usando los datos suministrados por la empresa educativa Edúcate, el cual contiene información de los simulacros de la prueba Saber 11 a los estudiantes de la institución educativa CAJASAI del año 2022. Además se observa el desempeño final de los mismos estudiantes en la prueba Saber 11 respecto al año 2022. Este desempeño será la variable objetivo de este trabajo.

Las fases del presente estudio fueron: (i) limpieza de la base de datos y selección de los registros relevantes, (ii) realización de análisis descriptivos y (iii) identificación de posibles modelos y la respectiva evaluación de su desempeño, usando la librería Pycaret (https://pycaret.org/). En este enlace se comparte el notebook con los desarrollos (https://colab.research.google.com/drive/1axcw9dqSxpQWq2_yI4CzB9uX5sGp4jN?usp=share_link).

Preparación inicial de los datos y las variables

En la primera fase, se eliminaron los registros sin información en cada una de las variables que fueron nueve registros por cada una así: número de estudiante, tarjeta identidad, nombre, apellidos, sexo, puntaje simulacro 1, puntaje simulacro 2, puntaje simulacro 2 y puntaje saber 11, ver tabla 1.

Variable	Descripción
Número de estudiante	Código para identificar al estudiante

Sexo	Sexo del estudiante
P. Simu 1	Puntaje numérico total que obtuvo durante el primer periodo.
P. Simu 2	Puntaje numérico total que obtuvo durante el segundo simulacro.
P. Simu 3	Puntaje numérico total que obtuvo durante el tercer simulacro.
Punt. Saber 11	Puntaje numérico total que obtuvo durante la Saber 11 del 2022.

Tabla 1: Descripción de las variables

En la modelación no se tiene encuentra la identificación del estudiante, además se imputaron los valores nulos por el promedio. Posteriormente, en la segunda fase, se realizaron los respectivos análisis descriptivos de cada variable, con el fin de identificar la distribución de los datos como se evidencia en las figuras 1, 2, 3 y 4.

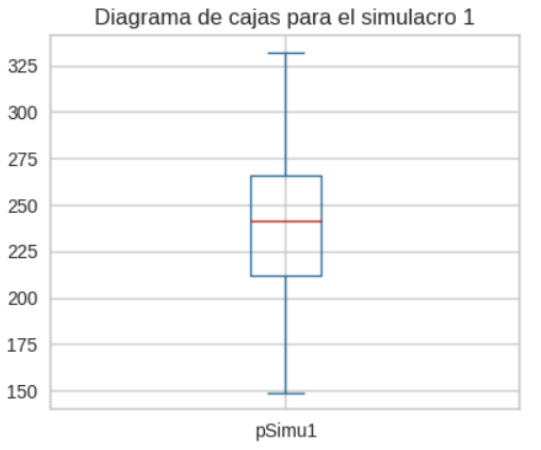

Figura 1: Diagrama de caja para el puntaje del simulacro 1

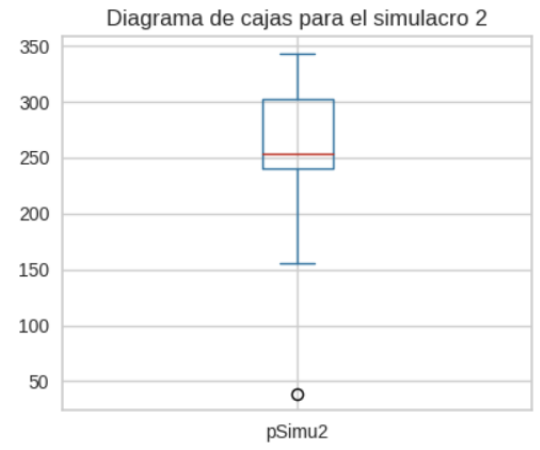

Figura 2: Diagrama de caja para el puntaje del simulacro 2

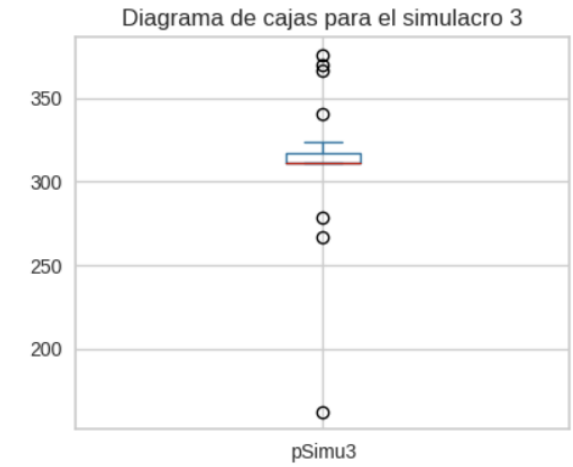

Figura 3:Diagrama de caja para el puntaje del simulacro 3

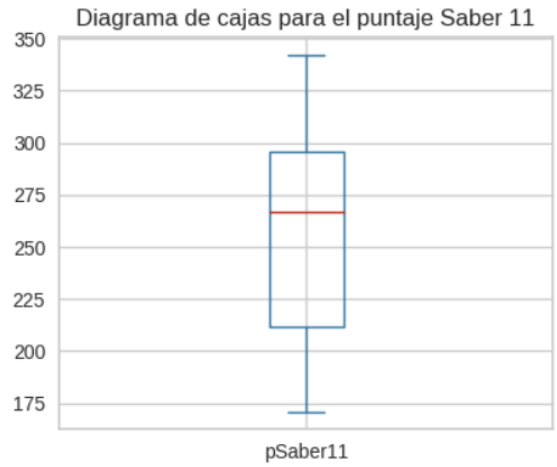

Figura 4:Diagrama de caja para el puntaje Saber 11

Procesamiento y modelación

En la tercera fase del análisis, se desarrollan los diferentes modelos y se selecciona aquel con mejor desempeño en el error cuadrático medio (RMSE). Para el desarrollo del modelo se consideró ajustarlo con 15 registros y se utilizaron 7 para evaluar el desempeño del modelo en datos no observados.

RESULTADOS

Se encontró que la variable *pSimu1* tiene fuerte correlación positiva con *pSaber11* ambas tienen la misma estructura, tiene una correlación fuerte positiva con *pSimu3*, pero es asimétrica y con *pSimu2* una correlación no tan fuerte positiva. Así mismo, la variable *pSimu2* tiene fuerte correlación con positiva con *pSaber11* ambas tienen la misma estructura, tiene una correlación no tan fuerte positiva con *pSimu3*, pero es asimétrica y con *pSimu1* una correlación no tan fuerte positiva. Por último, *pSimu3* se observa que tiene fuerte correlación con positiva con *pSaber11*, pero asimétrica ambas tienen la misma estructura, tiene una correlación no tan fuerte positiva con *pSimu1*, pero es asimétrica y con pSimu2 una correlación no tan fuerte positiva, esto se evidencia en la figura 5.

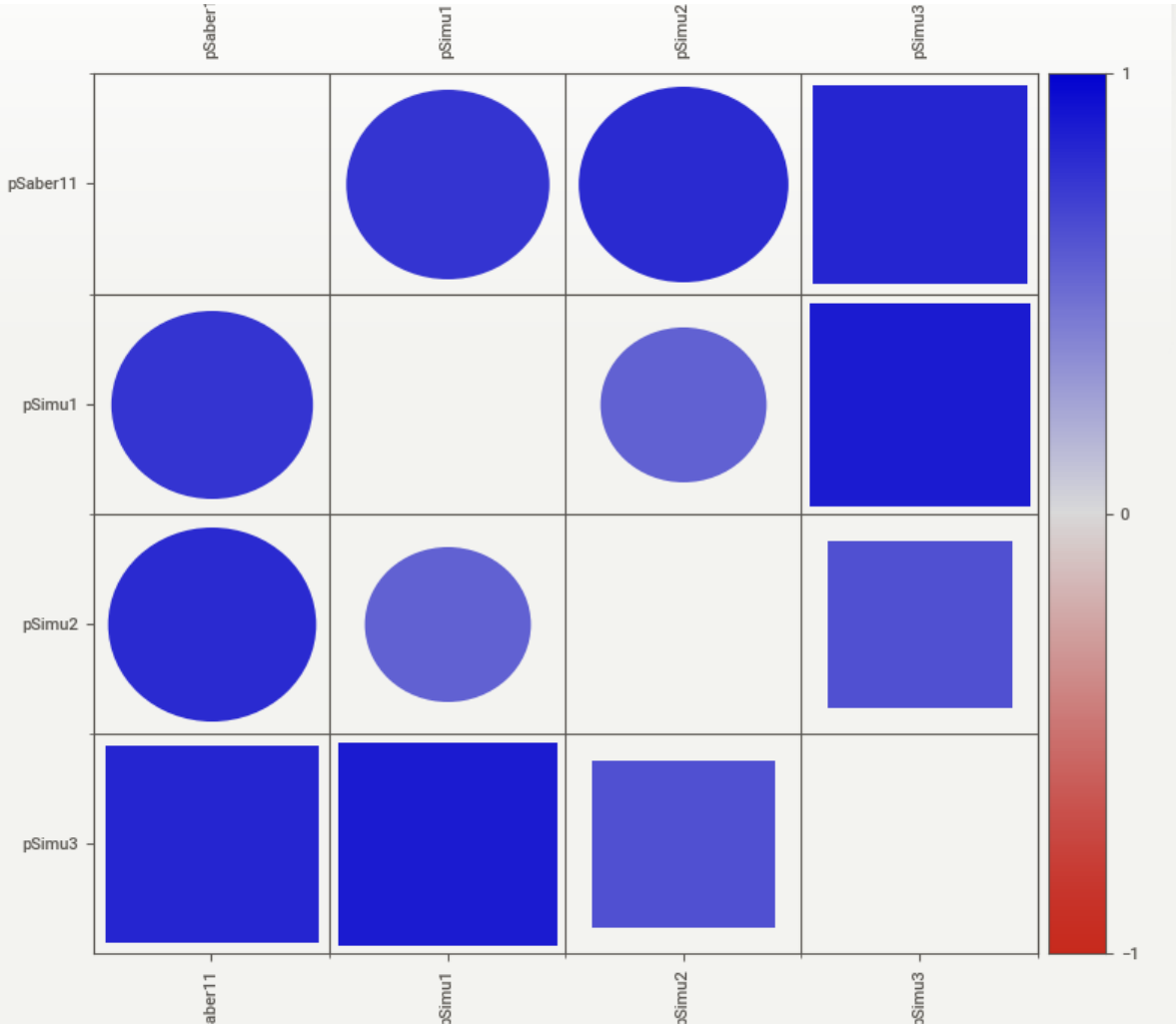

Figura 5: Mapa de correlación entre las variables

Del proceso de comparación de los modelos el mejor modelo escogido fue el Huber con un RMSE de 26.2683 como se evidencia en la tabla 4.

Model	*MAE*	*RMSE*	*MAPE*
Regresión Lineal	24.7294	26.9793	0.1103
Huber	24.0229	26.2683	0.1071
Lasso	24.6441	26.8834	0.1099
Ridge	24.7262	26.9755	0.1103

Tabla 2: Comparación de modelos por el criterio RMSE

Al evaluar el rendimiento del modelo con los datos de testeo y entrenamiento, se encontró que el modelo explica el 66,61% de la varianza lo cual es muy significativo dado el objetivo principal.

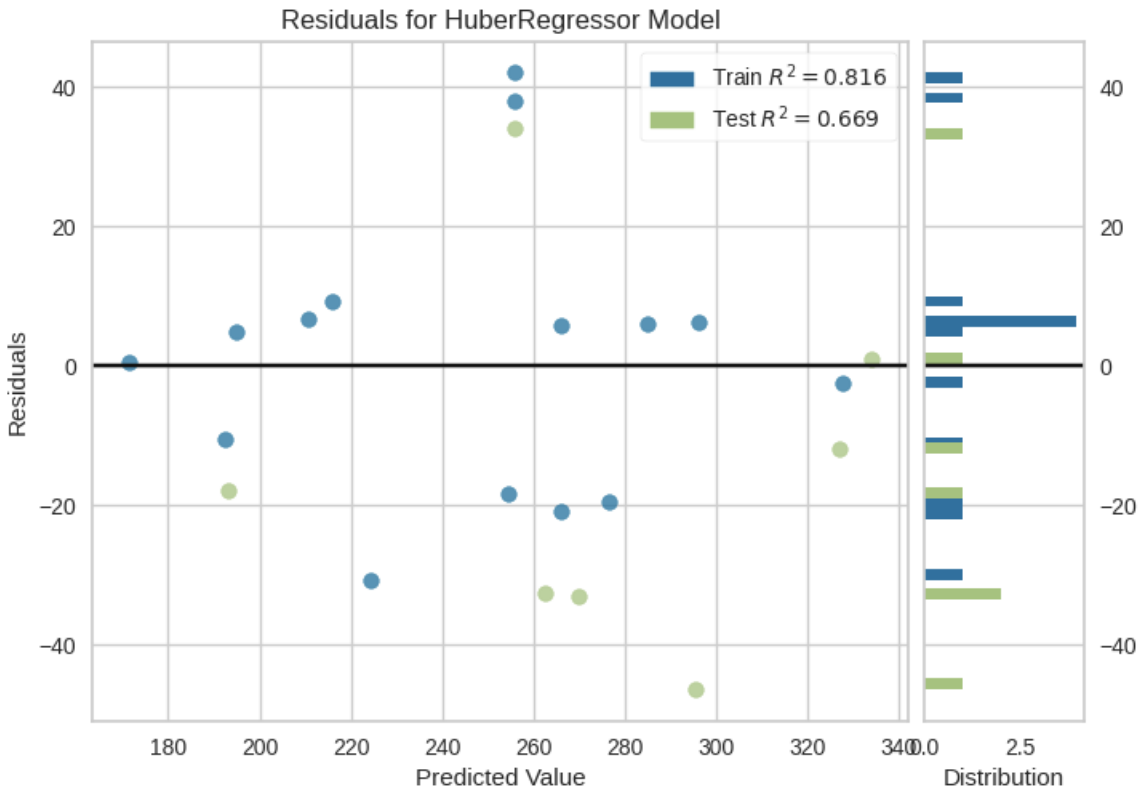

Figura 6: Diagrama de residuos para el modelo Huber

Por último, ser comparó el aporte significativo de las variables predictoras respecto a la variable objetivo, en la figura 7 se muestra esta relación.

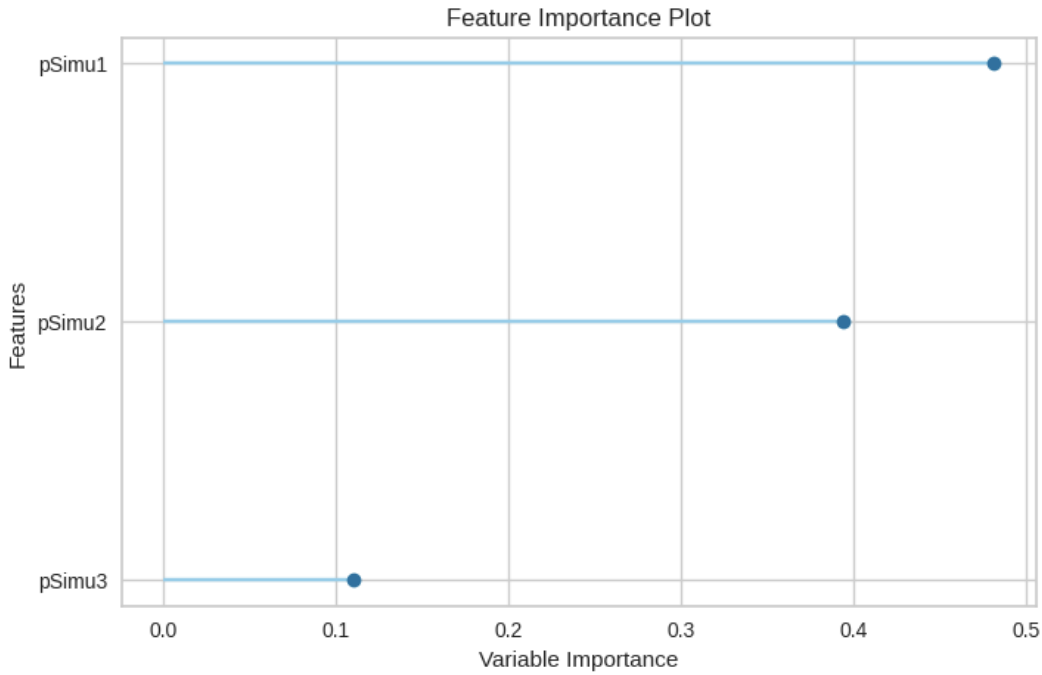

Figura 7: Importancia de las variables predictoras respecto a la variable objetivo

Sin embargo, se encontró que el p-valor para el tercer simulacro es mayor al 5% (ver tabla 5, lo que evidencia que no es significativo para la predicción del puntaje obtenido en la prueba Saber 11. Lo cual puede sugerir una omisión en la aplicación de este o una eliminación del modelo, en cualquier caso, sugiere una revisión al proceso después de la aplicación del segundo simulacro para analizar qué factores están influyendo en este.

DISCUSIÓN DE RESULTADOS

De acuerdo con la metodología planteada y los resultados obtenidos el mejor modelo fue Huber, este explica los datos con un 66.60% (ver figura 8) lo cual es muy significativo para el colegio porque permite determinar el posible puntaje que pueda tener un estudiante antes de presentar la prueba Saber 11 y así implementar estrategias a corto plazo para subir los niveles académicos.

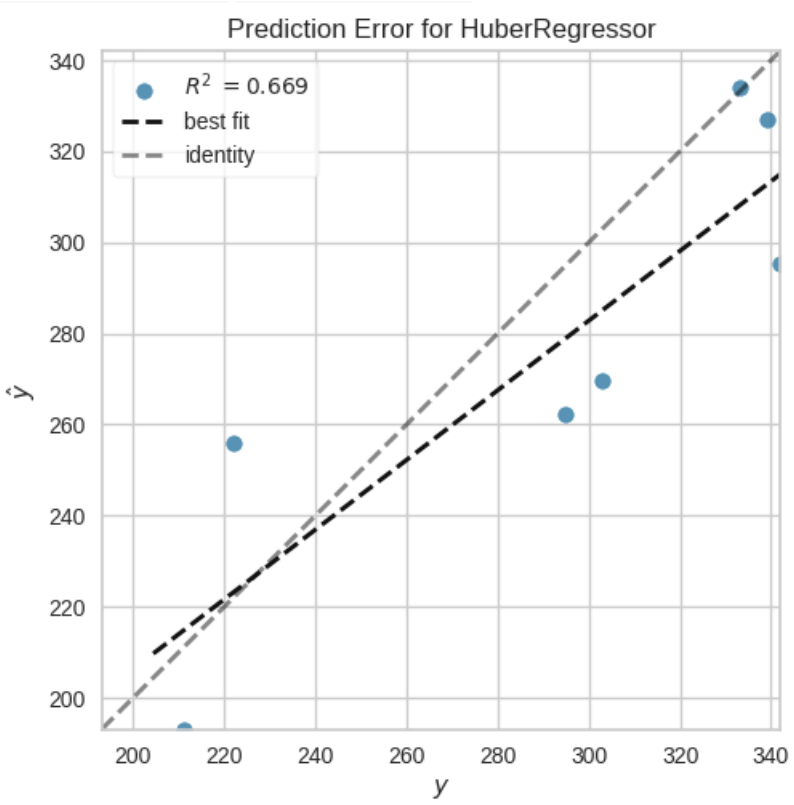

Figura 8: Coeficiente de determinación del modelo Huber

CONCLUSIONES

El tercer simulacro no está aportando a la predicción del resultado final de la prueba saber 11, esto se corrobora con el p-valor obtenido para esta variable, el cuál fue mayor al 5% y el diagrama de caja (figura 3) donde se observan muchos datos atípicos; tal vez este comportamiento sucede por la fecha de aplicación y por la intensidad, la cuál es a finales del mes de agosto, con una duración de 8 horas y media. Una posible alternativa de solución sería aplicar este tercer simulacro en dos sesiones evitando la sobrecarga al estudiante.

El modelo Huber fue el mejor modelo que predice el puntaje en la Saber 11 con un RSME de 26.2683, este modelo nos permite dar una primera mirada a los posibles puntajes a futuro a obtener con los resultados de los simulacros previos.

Con el modelo obtenido se pude prever una solución inicial a la problemática de este trabajo, que está asociada al índice de calidad de la institución, la cual radica en la incertidumbre de los resultados obtenidos por los estudiantes del grado Once en la prueba Saber 11. Dicha solución surgiría de los puntajes obtenidos por el modelo a través de los resultados de los dos simulacros, planteando un panorama futuro sobre las áreas con desempeño bajo, permitiendo emplear estrategias inmediatas sobre estás.

A futuro se enfatizar en realizar un proceso más riguroso al momento de registrar de la trazabilidad de los puntajes obtenidos por los estudiantes no únicamente de la institución CAJASAI, sino de todos los adscrito al programa pre-ICFES ofrecido por la empresa Edúcate, porque permitiría ajustar mejor los simulacros y conseguir un mejor pronóstico.

REFERENCIAS BIBLIOGRÁFICAS

Aguilera-Prado, M., Martinez Cervera, D. E., & Salcedo Parra, O. J. (2021). Forecasting model with machine learning in higher education ICFES exams. International Journal of Electrical and Computer Engineering (IJECE), pp. 5402~5410.

Ahumada, H., Herrera, M., Gabrielli, M., & Sosa, W. (2018). Una nueva econometría. Automatización, big data, econometría espacial y estructural. Bahía Blanca, Argentia: Editorial de la Universidad Nacional del Sur.

Al-Odwan, H. F. (2020). How Machine Learning affects Economics/Econometrics? A critical review of Machine Learning and Econometrics. Oxford Brookes Business School.

Bonaccorso, G. (2018). Machine Learning Algorithms. Birmingham, UK: Packt Publishing Ltd.

Bowen, D., & Ungar, L. (2020). Generalized SHAP: Generating multiple types of explanations in Machine Learnin

Colegio CAJASAI. (2022). Proyecto Institucional Educativo (PEI)

Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach.

ANEXOS

OLS Regression Results

Dep. Variable:	pSaber11	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.788
Method:	Least Squares	F-statistic:	27.00
Date:	Tue, 04 Apr 2023	Prob (F-statistic):	6.99e-07
Time:	04:04:32	Log-Likelihood:	-99.390
No. Observations:	22	AIC:	206.8
Df Residuals:	18	BIC:	211.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-17.5581	41.605	-0.422	0.678	-104.967	69.851
pSimu1	0.4043	0.154	2.622	0.017	0.080	0.728
pSimu2	0.4372	0.090	4.833	0.000	0.247	0.627
pSimu3	0.2216	0.150	1.478	0.157	-0.093	0.537

Omnibus:	1.144	Durbin-Watson:	1.588
Prob(Omnibus):	0.564	Jarque-Bera (JB):	0.980
Skew:	-0.470	Prob(JB):	0.613
Kurtosis:	2.570	Cond. No.	3.77e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.77e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Tabla 4: Resultados de una regresión OLS