

RITE

VOLUMEN 3 N° 1 (2025)

Revista Iberoamericana en Tecnologías
y Educación

Innovación tecnológica en la Educación

Inteligencia Artificial Educativa

7

Artículos



Título: Modelo predictivo basado en datos históricos y actuales para identificar tempranamente dificultades académicas en estudiantes de Ingeniería Industrial

**María del Socorro Barrientos
Camarena**

Centro Universitario Mar de Cortés
Tel 2225507330
México

Área temática: Competencias digitales en el proceso de Aprendizaje

Tipo de aporte: Innovación con la I.A

Resumen

El estudio desarrolla un modelo de regresión lineal múltiple para predecir el desempeño académico de estudiantes de Ingeniería Industrial y Rentabilidad de Negocios en la Universidad Madero. La investigación surge de la necesidad de identificar factores que influyen en el rendimiento estudiantil, especialmente después de la pandemia. El modelo permite predecir calificaciones con base en variables como el resultado del examen de admisión. Se observó que alumnos con exámenes de admisión deficientes tienen mayor probabilidad de reprobado materias. El modelo mostró potencial para identificar estudiantes en riesgo académico tempranamente. Representa un paso importante hacia sistemas de aprendizaje más personalizados y efectivos. El estudio destaca el potencial de la inteligencia artificial para adaptar y mejorar los procesos educativos, ofreciendo herramientas que permitan intervenciones tempranas y personalizadas.

Palabras clave: *Modelo predictivo, Desempeño académico, Regresión lineal múltiple, Analítica educativa, Inteligencia artificial en educación, Predicción de rendimiento estudiantil*

Introducción

Tradicionalmente, cuando pensamos en la impartición de un curso, visualizamos a un docente frente a un grupo utilizando un pizarrón. Con el tiempo, se han incorporado tecnologías para proyectar contenidos, junto con prácticas, juegos e investigaciones. Por otra parte, los métodos de evaluación incluyen exámenes, proyectos, tareas y participaciones, planeados con anticipación. Sin embargo, analizar estos datos para mejorar el aprendizaje es complicado y consume mucho tiempo, impidiendo acciones efectivas en tiempo real. Aquí es donde la inteligencia artificial puede adaptar y personalizar el aprendizaje, ofreciendo herramientas ajustables a las necesidades de los alumnos.

Metodología

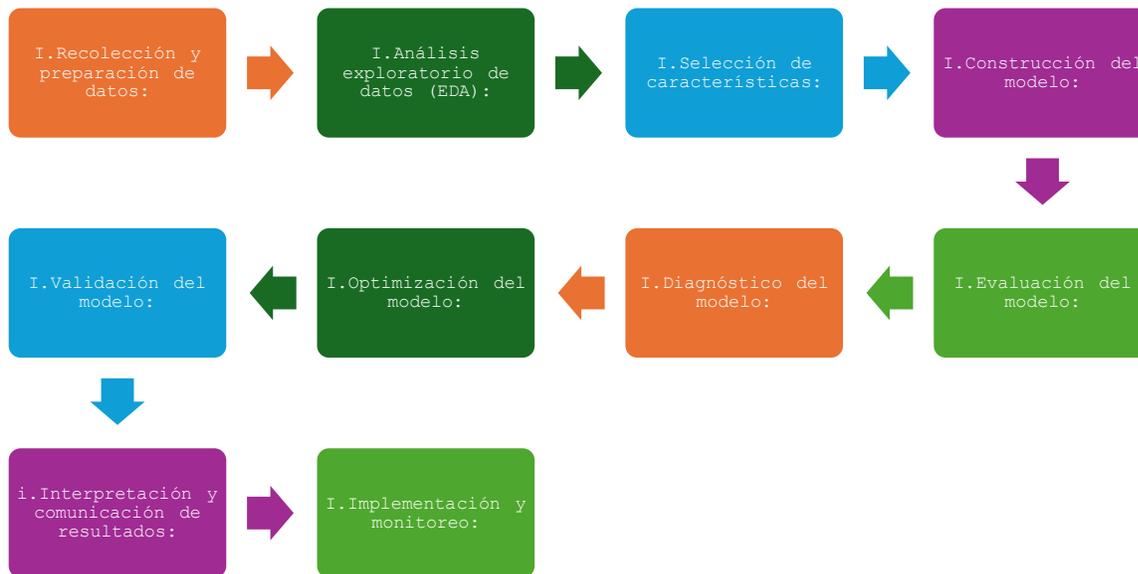
Metodología para un modelo predictivo de Machine Learning, ML, usando regresión lineal múltiple:

- I. Recolección y preparación de datos:
 - i. Recopilar datos relevantes para tu problema.
 - ii. Limpiar los datos (manejar valores faltantes, outliers, etc.).
 - iii. Dividir los datos en conjuntos de entrenamiento y prueba.
- II. Análisis exploratorio de datos (EDA):
 - i. Visualizar la distribución de variables.
 - ii. Examinar correlaciones entre variables.
 - iii. Identificar posibles relaciones no lineales.
- III. Selección de características:
 - i. Elegir variables independientes relevantes.
 - ii. Considerar la multicolinealidad entre variables.
- IV. Construcción del modelo:

- i. Implementar la regresión lineal múltiple.
 - ii. Ajustar el modelo a los datos de entrenamiento.
- V. Evaluación del modelo:
 - i. Calcular métricas de rendimiento (F de Fisher, R^2 , Error Cuadrático Medio, etc.).
 - ii. Analizar los coeficientes y su significancia estadística.
- VI. Diagnóstico del modelo:
 - i. Verificar supuestos de la regresión lineal (linealidad, homocedasticidad, normalidad de residuos, independencia).
 - ii. Identificar observaciones influyentes.
- VII. Optimización del modelo:
 - i. Realizar selección de variables (forward, backward, stepwise).
 - ii. Considerar transformaciones de variables si es necesario.
- VIII. Validación del modelo:
 - i. Evaluar el rendimiento en el conjunto de prueba.
 - ii. Realizar validación cruzada si es apropiado.
 - iii. Interpretación y comunicación de resultados:
 - iv. Explicar el significado de los coeficientes.
 - v. Comunicar las limitaciones y alcances del modelo.
- IX. Implementación y monitoreo:
 - i. Desplegar el modelo en un entorno de producción.
 - ii. Establecer un sistema de monitoreo y actualización del modelo.

Figura 1 Metodología para modelo predictivo basado en

<https://www.codetodevs.com/regresion-lineal-multiple-con-python/>



Resultados y discusión

La Universidad Madero tiene 30 años de experiencia formando ingenieros, por lo general los estudiantes tienen problemas de desempeño académico, sin embargo, después de Pandemia los comportamientos fueron cambiando y es por ello por lo que se requiere identificar los factores que determinan el desempeño académico, para el presente ejercicio se eligieron a los alumnos activos de Ingeniería Industrial y Rentabilidad de Negocios.

Recolección y preparación de datos:

La fuente de datos es el SISE (SISTEMA INTEGRAL DE SERVICIOS EDUCATIVOS) de él se extraen las siguientes variables:

1. Clave de materia
2. Materia
3. Matrícula de estudiante

4. Nombre del alumno
5. Calificación
6. Periodo
7. Materias Reprobadas
8. Examen de admisión

Tratamiento de los datos: por cuestiones de confiabilidad se omiten los datos, los alumnos tienen opción para darse de baja por lo que se sustituye por una calificación de 0 cero, los resultados de examen de admisión se encuentran en forma cualitativa por lo que se transfieren en forma numérica: DEFICIENTE – 1, TERMINO MEDIO NEGATIVO – 2, TERMINO MEDIO – 3, SUPERIOR AL TERMINO MEDIO – 4, SUPERIOR – 5.

Variable de respuesta: se elige la variable “Calificación” como respuesta o “Y”.

Análisis exploratorio de datos (EDA) y selección de características:

Una vez recolectados los datos como se muestra en la Tabla 1, que en total son 583 registros:

Tabla 1 Extracto de base de datos

No.	Matrícula	Sexo	Periodo nuevo	Examen de admisión	Tipo de materia	Calificación
1	20101234	2	1	3	4	9
2	20101234	2	1	3	3	7
3	20101234	2	1	3	8	10
4	20101234	2	1	3	8	10
5	20101234	2	1	3	8	10
6	20101234	2	1	3	8	8
7	20101234	2	1	3	1	8
8	20101234	2	2	3	1	5
9	20101234	2	2	3	2	10
10	20101234	2	2	3	3	10
296	20079166	2	9	3	3	8.5
297	20079166	2	9	3	2	5.8
298	20079166	2	9	3	3	7.3
299	20079166	2	9	3	3	6.9
300	20079166	2	9	3	8	7.6
301	20080479	2	6	3	4	9
302	20080479	2	6	3	3	9
303	20080479	2	6	3	2	10
304	20080479	2	6	3	3	10
305	20080479	2	6	3	8	10
306	20080479	2	6	3	8	5
307	20080479	2	6	3	1	7
308	20080479	2	7	3	1	5
309	20080479	2	7	3	2	5

Se realiza un análisis de correlación:

Tabla 2 Análisis de correlación

	Matrícula	Sexo	Periodo nuevo	Examen de admisión	Tipo de materia	Calificación
Matrícula	1					
Sexo	0.32635657	1				
Periodo nuevo	-0.05276373	0.10371793	1			
Examen de admisión	-0.07709244	0.26843965	-0.00087048	1		
Tipo de materia	-0.05091509	-0.00987391	-0.19799039	-0.01150392	1	
Calificación	-0.17879018	-0.18379599	-0.25287203	0.08540129	0.15252334	1

n = 585
 $r_{\alpha} = 0.164$ para n = 100

Se observan las correlaciones, se utilizó la matrícula en lugar del nombre de alumno para conservar el anonimato se pudiera interpretar que los números más grandes son los alumnos de menores semestres, se agregó la variable sexo para saber si tiene alguna influencia.

Sobre el tipo de materia se había clasificado con las relacionadas con:

Tabla 3 Codificación

Clasificación	Número
Matemáticas	1
Ciencias	2
Ingeniería	3
Programación	4
Diseño	5
Costos	6
Tronco	7

Se esperaba que existiera correlación entre la calificación y el tipo de materia, pero con el factor de correlación no se muestra, sin embargo, al hacer una corrida de regresión múltiple se observa lo siguiente:

	Coefficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%	Inferior 95.0%	Superior 95.0%
Intercepción	10.898894	0.6415025	16.9896362	1.2434E-52	9.63888262	12.1589054	9.63888262	12.1589054
Matrícula	-7.5781E-08	2.5121E-08	-3.01663948	0.00267054	-1.2512E-07	-2.6439E-08	-1.2512E-07	-2.6439E-08
Sexo	-0.58790276	0.17352545	-3.38799155	0.00075292	-0.92873393	-0.24707159	-0.92873393	-0.24707159
Periodo nuevo Examen de admisión	-0.19047419	0.03493924	-5.45158433	7.4592E-08	-0.25910033	-0.12184806	-0.25910033	-0.12184806
Tipo de materia	0.31016072	0.10912714	2.8421961	0.00464168	0.09581793	0.52450352	0.09581793	0.52450352
	0.07514102	0.0288369	2.60572414	0.00940886	0.01850082	0.13178122	0.01850082	0.13178122

Como se puede observar todas las variables tendrían validez para ejecutar un modelo ya que sus probabilidades son menores a $\alpha = 0.05$

Construcción del modelo

Se realiza la corrida de regresión lineal múltiple, obteniendo el siguiente modelo:

$$Y = 10.8989 + -7.57806E-08 X_1 + -0.5879 X_2 + -0.1905 X_3 + 0.3102 X_4 + 0.0751 X_5$$

Donde:

X_1	Matrícula
X_2	Sexo
X_3	Periodo nuevo
X_4	Examen de admisión
X_5	Tipo de materia

Evaluación del modelo

Si suponemos: un alumno de nuevo ingreso que tuvo un examen de admisión deficiente para una materia de matemáticas estaríamos esperando una calificación reprobatoria ya que en la UMAD la calificación mínima es 7

X_1	Matrícula	20140885
X_2	Sexo	2
X_3	Periodo nuevo	9
X_4	Examen de admisión	1
X_5	Tipo de materia	1
	$Y =$	6.867835006

Ahora si mantenemos los mismos datos, pero estimamos que tuvo un examen de admisión con término medio

(3) la calificación es aprobatoria

X_1	Matrícula	20140885
X_2	Sexo	2
X_3	Periodo nuevo	9

X_4	Examen de admisión	3
X_5	Tipo de materia	1
	$Y =$	7.488156451

Optimización y validación del modelo

Parece alentador el modelo, sin embargo, es necesario continuar con la limpieza y tratamiento de variables ya que presenta anomalías en sus residuos estándar y estos están relacionados con alumnos con materias de matemáticas reprobadas, se muestra un fragmento de los resultados:

Tabla 4 Extracto de análisis de residuales

<i>Observación</i>	<i>Pronóstico Calificación</i>	<i>Residuos</i>	<i>Residuos estándares</i>
1	9.24037782	-0.24037782	-0.13664659
2	9.16523681	-2.16523681	-1.23086325
3	9.5409419	0.4590581	0.26095887
4	9.5409419	0.4590581	0.26095887
5	9.5409419	0.4590581	0.26095887
6	9.5409419	-1.5409419	-0.87597289
7	9.01495477	-1.01495477	-0.57696716
8	8.82448058	-3.82448058	-2.17408672
9	8.89962159	1.10037841	0.62552758
10	8.97476261	1.02523739	0.58281248
11	9.3504677	-1.3504677	-0.76769481
12	9.08485249	0.91514751	0.52023013
13	8.63400638	-3.63400638	-2.06580864
14	9.08485249	-1.08485249	-0.61670163
15	8.93457046	1.06542954	0.60566035
16	9.15999351	-2.15999351	-1.22788262
17	9.08485249	-0.08485249	-0.04823575
18	8.44353219	-1.44353219	-0.8205988
19	8.59381423	-0.59381423	-0.33756313
20	8.96951932	-3.96951932	-2.2565363
21	8.8943783	-0.8943783	-0.50842355
22	8.8943783	1.1056217	0.62850821
23	8.77904513	-0.77904513	-0.44286057

Uso de Python

El código se obtiene a partir de una consulta en Claude IA mediante el prompt: ¿Me puedes dar la metodología para hacer un modelo predictivo de Machine Learning utilizando una regresión lineal múltiple?

Para la recolección de datos se realizan consultas para crear una base de datos y se reunió todas las características siguientes: todas las columnas tienen título, todas las variables están

expresadas en forma numérica, no existen espacios vacíos, estos datos se capturaron en un archivo Excel y se pasó a uno csv.

Recolección y preparación de datos

Análisis exploratorio de datos (EDA), en este punto se examinan correlaciones entre variables, e identificación de relaciones no lineales.

Selección de características: en este punto se eligen las variables independientes

Construcción del modelo: aquí se implemente la regresión lineal múltiple y se ajustar el modelo a los datos de entrenamiento.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns

# Cargar datos (asumiendo que tienes un archivo CSV)
data = pd.read_csv('Regresion multiple desempenio alumnos 024.csv')

# Seleccionar variables predictoras y objetivo
X = data[['Tipo de materia', 'Examen de admision', 'Matricula', 'Sexo', 'Periodo nuevo']]
y = data['Calificacion']

# Dividir datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Crear y entrenar el modelo
model = LinearRegression()
model.fit(X_train, y_train)
```

Evaluación del modelo: Calcular métricas de rendimiento (R^2 , Error Cuadrático Medio, etc.) y analizar los coeficientes y su significancia estadística.

```
# Hacer predicciones
y_pred = model.predict(X_test)
```

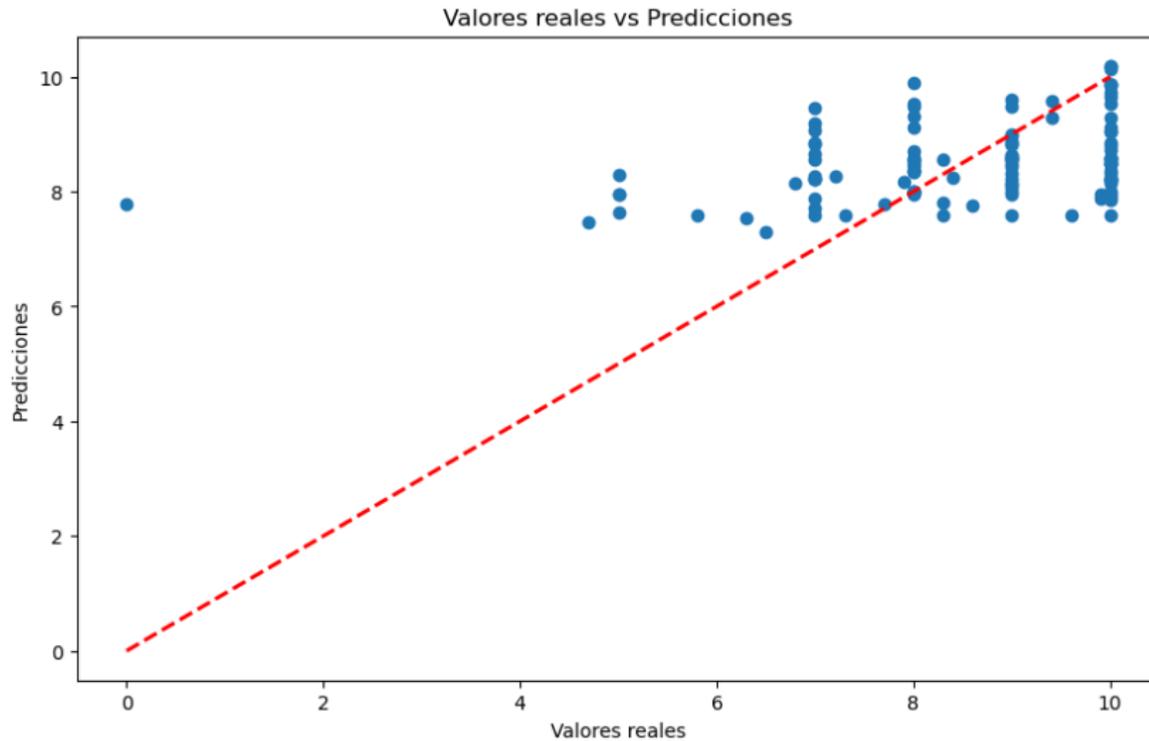
```
# Evaluar el modelo
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

```
print(f'Error cuadrático medio: {mse}')
print(f'R-cuadrado: {r2}')
```

```
Error cuadrático medio: 2.1955312749642473
R-cuadrado: 0.11699660476549856
```

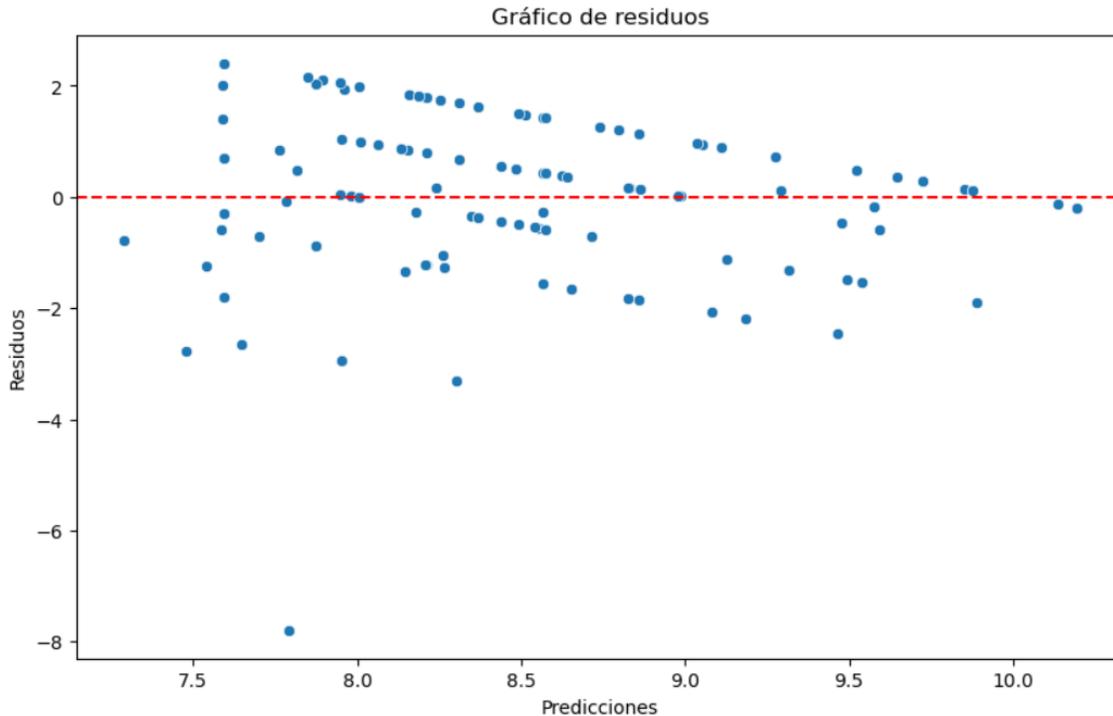
```
# Visualizar resultados
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--', lw=2)
plt.xlabel('Valores reales')
plt.ylabel('Predicciones')
plt.title('Valores reales vs Predicciones')
plt.show()
```

Se observa en el gráfico que será necesario hacer más tratamientos para ajustar el modelo, como se comentó en el Alcance de la Investigación, al ser una fase de pilotaje se tendrá que hacer ajustes posteriores, pero lo importante es que ya se visualizan estas situaciones.



Es importante revisar los residuos

```
# Diagnóstico de residuos
residuos = y_test - y_pred
plt.figure(figsize=(10, 6))
sns.scatterplot(x=y_pred, y=residuos)
plt.axhline(y=0, color='r', linestyle='--')
plt.xlabel('Predicciones')
plt.ylabel('Residuos')
plt.title('Gráfico de residuos')
plt.show()
```



Al analizar los datos de origen se observa una relación fuerte con las calificaciones con 0, que son aquellos alumnos que se dieron de baja.

El análisis de los coeficientes e intercepción se obtiene:

```
# Analizar coeficientes e intercepción
coef_df = pd.DataFrame({'Variable': X.columns, 'Coeficiente': model.coef_})
intercept_df = pd.DataFrame({'Variable': ['Intercepción'], 'Coeficiente': [model.intercept_]})
coef_df = pd.concat([intercept_df, coef_df], ignore_index=True)
print("\nCoeficientes e Intercepción del modelo:")
print(coef_df)
```

Coeficientes e Intercepción del modelo:

	Variable	Coeficiente
0	Intercepción	1.117371e+01
1	Tipo de materia	5.651297e-02
2	Examen de admision	3.038854e-01
3	Matricula	-7.321271e-08
4	Sexo	-6.687606e-01
5	Periodo nuevo	-2.057633e-01

```
# Crear una ecuación del modelo
equation = f"y = {model.intercept_:.4f}"
for i, col in enumerate(X.columns):
    equation += f" + {model.coef_[i]:.4f} * {col}"
print("\nEcuación del modelo:")
print(equation)
```

La ecuación del modelo es:

```
Ecuación del modelo:
y = 11.1737 + 0.0565 * Tipo de materia + 0.3039 * Examen de admision + -0.0000 * Matricula
+ -0.6688 * Sexo + -0.2058 * Periodo nuevo
```

Para poner a prueba el modelo se solicitó el código con el siguiente prompt: Del código "Implementación de Regresión Lineal Múltiple en Python", ¿cómo podría escribir un dataframe de prueba y que me imprima el resultado calculado por el modelo?

Se realiza la prueba con un alumno de primer semestre, hombre, para una materia relacionada con matemáticas y se hace la prueba con 4 resultados de examen de admisión, con 1 como deficiente, 2 inferior al término medio, 3 término medio y 4 superior al término medio, cabe mencionar que en la base de datos no aparece algún alumno con superior que sería 5.

```
# Crear un DataFrame de prueba
df_prueba = pd.DataFrame({
    'Tipo de materia': [1, 1, 1, 1],
    'Examen de admision': [1, 2, 3, 4],
    'Matricula': [20140885, 20140885, 20140885, 20140885],
    'Sexo': [2, 2, 2, 2],
    'Periodo nuevo': [9, 9, 9, 9]
})
```

```
# Hacer predicciones con el DataFrame de prueba
predicciones = model.predict(df_prueba)
```

```
# Añadir las predicciones al DataFrame de prueba
df_prueba['Prediccion'] = predicciones
```

```
# Imprimir el DataFrame de prueba con las predicciones
print("\nDataFrame de prueba con predicciones:")
print(df_prueba)
```

Los resultados son:

```
DataFrame de prueba con predicciones:
   Tipo de materia  Examen de admision  Matricula  Sexo  Periodo nuevo  \
0                1                    1  20140885    2             9
1                1                    2  20140885    2             9
2                1                    3  20140885    2             9
3                1                    4  20140885    2             9

   Prediccion
0    6.870147
1    7.174032
2    7.477918
3    7.781803
```

Como se puede observar en esta prueba es que cuando un alumno de nuevo ingreso resulta deficiente en su examen de admisión reprobaría alguna materia de matemáticas

(la UMAD considera reprobatoria una calificación inferior a 7), que es coherente con la realidad.

EVALUACIÓN DEL IMPACTO EDUCATIVO:

Los modelos predictivos pueden tener un impacto significativo en la mejora educativa al proporcionar información valiosa para la toma de decisiones y la personalización del aprendizaje.

Impacto de los modelos predictivos en la mejora educativa:

Identificación temprana de estudiantes en riesgo: Los modelos predictivos pueden identificar a los estudiantes que están en riesgo de fracaso académico o abandono escolar antes de que ocurran problemas graves (Márquez-Vera et al., 2016).

Personalización del aprendizaje: Estos modelos pueden ayudar a adaptar el contenido y el ritmo de aprendizaje a las necesidades individuales de cada estudiante (Baker & Inventado, 2014).

Optimización de recursos: Permiten una asignación más eficiente de recursos educativos, dirigiéndolos hacia donde son más necesarios (Pistilli & Arnold, 2010).

Mejora en la retención de estudiantes: Al identificar factores de riesgo, las instituciones pueden implementar estrategias para mejorar la retención estudiantil (Thammasiri et al., 2014).

Evaluación continua: Los modelos predictivos facilitan la evaluación continua del progreso del estudiante, permitiendo intervenciones oportunas (Siemens & Long, 2011).

Métricas para evaluar la efectividad de los modelos predictivos en educación:

1. Precisión del modelo:
 - Exactitud (Accuracy): Porcentaje de predicciones correctas.

- Área bajo la curva ROC (AUC-ROC): Mide la capacidad del modelo para distinguir entre clases.
2. Métricas educativas:
 - Tasa de retención estudiantil: Porcentaje de estudiantes que continúan sus estudios.
 - Tasa de graduación: Porcentaje de estudiantes que completan sus estudios en el tiempo previsto.
 - Promedio de calificaciones (GPA): Mejora en las calificaciones generales de los estudiantes.
 3. Métricas de intervención:
 - Tasa de éxito de intervenciones: Porcentaje de estudiantes que mejoran después de las intervenciones basadas en predicciones.
 - Tiempo hasta la intervención: Reducción en el tiempo para identificar y abordar problemas académicos.
 4. Satisfacción de las partes interesadas:
 - Encuestas de satisfacción a estudiantes, profesores y administradores sobre la utilidad del sistema predictivo.
 5. Eficiencia de recursos:
 - Relación costo-beneficio: Comparación entre los recursos invertidos en el sistema predictivo y los beneficios obtenidos.
 6. Equidad educativa:
 - Brechas de rendimiento: Reducción en las disparidades de rendimiento entre diferentes grupos de estudiantes.
 - Identificación de alumnos con diferentes capacidades como es el caso de estudiantes con síndrome de Asperger.

RESULTADOS Y ANÁLISIS:

La reflexión sobre la elección y manejo de variables sería interesante agregar una variable más que indique si el alumno reprobó o se dio de baja ya que al haber puesto un cero a las bajas incremento las anomalías en la prueba de residuos.

Una vez obtenido el modelo a pesar de que tiene validez es necesario incorporar más variables como son: si tienen beca o no, si tienen otras actividades extracurriculares, como incluir variables de retención o deserción, como se indicó en el apartado 5 es necesario investigar el por qué las anomalías son de materias relacionadas con las matemáticas.

CONCLUSIÓN Y RECOMENDACIONES.

Es posible decir que el modelo propuesto es prometedor, sin embargo, presenta ciertas limitaciones que deben abordarse en futuras investigaciones. La generalización de los resultados a diferentes contextos educativos, la forma en la interpretación de los modelos y la ética en el uso de datos son aspectos clave a considerar. A pesar de estos desafíos, el estudio abre nuevas vías para la investigación en analítica educativa y sienta las bases para el desarrollo de sistemas de aprendizaje más personalizados y efectivos.

Este ejercicio da pie a identificar el impacto de un modelo predictivo en la mejora educativa, ayudando a la ubicar en forma temprana a alumnos que tendrían problemas en su desempeño que pueden ir desde baja de calificaciones, pérdida de beca y hasta deserción, por lo que la implementación de modelos predictivos impactaría en forma positiva al dar alertas que ayuden a tomar acciones en forma temprana.

1. REFERENCIAS:

- Mera Castillo, D. E. (2023). La influencia de la inteligencia artificial en la personalización del aprendizaje: *Perspectivas y desafíos en la educación*. *Revista Ingenio Global*, 2(2), 28–39. <https://doi.org/10.62943/rig.v2n2.2023.64>
- Morillo Lozano, M. C. (2016). *Aprendizaje Adaptativo: Adaptive Learning* (Tesis de Máster). Universidad de Valladolid.
- Salgado Reyes, N., & Beltrán Morales, J. (2018). Modelo para predecir el rendimiento académico basado en redes neuronales y análisis de aprendizaje
- Avendaño Prieto, B. L., Avendaño Prieto, G., Cruz, W., & Cárdenas-Avendaño, A. (2014). Guía de referencia para investigadores no expertos en el uso de estadística multivariada. *Diversitas* (17949998), 10(1), 13–27. Retrieved from <https://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,url,custuid&custid=s4231244&db=fua&AN=102024754&site=ehost-live>
- Badii, M. H., Guillen, A., Cerna, E., Valenzuela, J., & Landeros, J. (2012). Análisis de Regresión Lineal Simple para Predicción. *Revista Daena (International Journal of Good Conscience)*, 8(1), 67–81. Recuperado de <https://search.ebscohost.com/login.aspx?direct=true&AuthType=cookie,ip,url,custuid&custid=s4231244&db=fua&AN=93609071&site=ehost-live>
- Carollo Limeres, Carmen. (2011). Regresión Lineal Simple. Universidad de Santiago de Compostela. Recuperado de http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf
- CodeToDevs. (s.f.). Modelos de Regresión Lineal Múltiple con Python. Recuperado de <https://www.codetodevs.com/regresion-lineal-multiple-con-python/>
- Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61-75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4
- Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: A case study with high school students. *Expert Systems*, 33(1), 107-124. <https://doi.org/10.1111/exsy.12135>
- Pistilli, M. D., & Arnold, K. E. (2010). In practice: Purdue Signals: Mining real-time academic data to enhance student success. *About Campus*, 15(3), 22-24. <https://doi.org/10.1002/abc.20025>

- Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, 46(5), 30-32.
- Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2), 321-330. <https://doi.org/10.1016/j.eswa.2013.07.046>