UNIVERSIDAD MAYOR DE SAN ANDRES

FACULTAD DE CIENCIAS PURAS Y NATURALES

CARRERA DE INFORMATICA

INTEGRANTES

Univ. Joaquín Gabriel Kapa Cruz (INF – 354)

Univ. Vladimir Ariel Lizarro Velásquez (DAT – 245)

Univ. Samuel Alejandro Aruquipa Mamani (DAT – 245)

DOCENTE

Ph. D. Moises Martin Silva Choque

MATERIA

Inteligencia Artificial DAT-245/INF-354.

PROYECTO DE INTELIGENCIA ARTIFICIAL

1. RESUMEN

El presente artículo analiza cómo diversos factores personales, académicos y sociales afectan el rendimiento académico de estudiantes, utilizando un conjunto de datos con 6,607 registros. Se aplicaron modelos de regresión lineal y técnicas como el Análisis de Componentes Principales (PCA) para optimizar el rendimiento del modelo. Se evaluó la consistencia del modelo mediante validación cruzada y métricas como MSE, MAE y R^2.

2. INTRODUCCION

Se busca identificar los elementos que impactan en los resultados académicos, utilizando variables como horas de estudio y asistencia. Este análisis es útil para docentes y responsables de políticas educativas.

3. RESULTADOS

- **Descripción general**

  El dataset cuenta con exactamente 6.607 registros que corresponden a los estudiantes, tenemos 20 clases que influyen factores cuantitativos como ser horas dedicadas al estudio o ejercicio físico, además, están presentes factores cualitativos como ser tipo de escuelas o nivel de motivación, la variable objetivo es 'Exam_score' que representa el puntaje obtenido por cada estudiante en un examen reciente.

- **Factores analizados**
  - Personales: horas de estudio, motivación.
  - Familiares: ingresos económicos, educación parental.
  - Escolares: calidad docente, asistencia a tutorías.
  - Sociales: influencia de compañeros.

- **Preprocesamientos realizados y balanceo de datos**

  En el dataset se presentaron datos nulos y datos categóricos, realizando el 'ReplaceMissingValues, One_hot_Encoding, Normalize' se rellenaron los datos nulos, se volvieron de datos categóricos a datos numéricos y se normalizo los datos ya numéricos.

- **Matriz de correlación**

  Tras realizar el calculo del MAE, MSE y R^2, obtuvimos que el modelo presenta un rendimiento optimo, además que, las predicciones se desvían 0.28 unidades a los datos reales, y que el R^2 resulto 98.94% que significa que nuestro modelo presenta una predicción casi exacta.

- **Splits**

  Tras realizar el análisis de (80/20) y (50/50) para garantizar que nuestro modelo es robusto, obtuvimos que el MSE y MAE de (80/20) es de 0.1059 y 0.2684 respectivamente y el MSE Y MAE de (50/50) es de 0.1069 y 0.2734 respectivamente, dichos resultados son bajos y relativamente similares lo que

significa que nuestro modelo es consistente y preciso, además que dichos valores al ser bajos y similares en ambos Splits, el modelo no está sobreajustado y mantiene un desempeño constante a través de las distintas particiones de datos.

- **PCA**

  Al usar 10 componentes principales se conserva el 57.69% de la varianza mientras que usando 3 componentes se conserva el 17.69% de la varianza, mientras haya una mayor varianza se representan mejor en el espacio reducido.
  Al usar 10 componentes principales produce el menor MSE con 7.3285 lo que significa que esta configuración captura mejor las características relevantes para la predicción, el MAE menor con 2.1563, en lo que respecta al usar menos componentes se pierde información mientras que al usar más componentes se logra conservar un equilibrio y mantener un error bajo.

- **Aprendizaje no supervizado**

  Se llevó a cabo un análisis de aprendizaje no supervisado utilizando un algoritmo de clustering para agrupar estudiantes según sus características relacionadas con el rendimiento académico. El objetivo es identificar patrones sin una variable objetivo definida previamente. Las variables analizadas incluyen horas de estudio, asistencia a clases, involucramiento parental, acceso a recursos educativos, participación en actividades extracurriculares, horas de sueño, calificaciones previas, motivación, acceso a Internet, sesiones de tutoría, ingresos familiares, calidad del profesorado, tipo de escuela, influencia de compañeros, actividad física, discapacidades de aprendizaje, nivel educativo de los padres y distancia a la escuela.
  Para facilitar la visualización de los clusters, se aplicó reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA). Esto permite representar los datos en un formato más comprensible, donde cada punto en el gráfico representa a un estudiante y los colores indican su pertenencia a un cluster. La separación visual sugiere agrupaciones significativas basadas en características académicas y personales.

- **Algoritmo de las n-reinas**

  El algoritmo de recocido simulado resuelve el problema de ubicar N reinas en un tablero de N x N sin que se ataquen entre sí (no pueden compartir fila, columna ni diagonal). Comienza con una configuración inicial aleatoria y calcula su "costo", que corresponde a la cantidad de amenazas entre las reinas.

  Inspirado en el proceso de enfriamiento de un metal, el algoritmo inicia con una temperatura alta, lo que permite

aceptar soluciones peores para evitar quedar atrapado en óptimos locales. A medida que la temperatura disminuye, el algoritmo se vuelve más selectivo y solo acepta soluciones que mejoren la configuración actual.

El proceso finaliza cuando se encuentra una solución sin amenazas (costo = 0) o cuando ya no hay mejoras posibles. Además, el sistema permite visualizar el tablero con las posiciones de las reinas o informar si no se encontró una solución

4. METODOLOGIA

El método empleado fue la indagación acerca de la información que hay que realizar para el análisis de nuestro dataset, además que, se utilizó software para realizar los distintos análisis como ser Python.

5. CONCLUSIONES

Concluimos que el dataset utilizado tiene variables como ser 'Hours_Studied, Attendance, Previus_Score' presentan correlaciones positivas, esto para poder determinar que el modelo puede predecir el 'Exam_Score' que es nuestra variable objetivo, tras realizar el cálculo del MSE Y MAE obteniendo valores bajos lo que indica que el modelo tiene una buena capacidad de predicción, junto con el 98.94% de exactitud lo que indica que el modelo tiene una predicción casi perfecta.

Con toda esta información se puede utilizar para el desarrollo de estrategias pedagógicas personalizadas mejorando la asistencia o el tiempo de estudio, además de crear políticas públicas que coadyuven a mejorar los programas educativos.

6. REFERENCIAS

- Liu, Q., & Chen, G. (2021). Key Influences on Students' Academic Success: Insights from Scholarly Research. *Journal of Educational Research*, 2(1), 9-19.
- Kotsiantis, S. B. (2007). *Supervised Machine Learning: A Review of Classification Techniques*. Informatica. ISBN: 978-1-58603-780-2
- Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, 1(1), 3–17.