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.

Contenido

PROYECTO DE INTELIGENCIA ARTIFICIAL

1. RESUMEN

El presente artículo presenta un análisis sobre los factores que influyen en el rendimiento académico de los estudiantes, utiliza un dataset con 6,607 registros de estudiantes, incluyendo factores tanto personales como académicos.

Se realizó un preprocesamiento de los datos, incluyendo la gestión de valores nulos, la codificación de variables categóricas y la normalización de los datos. Posteriormente, se aplicaron modelos de regresión lineal para predecir el rendimiento académico, con variables como 'Hours_Studied, Attendance, Previous_Scores' siendo las más influyentes. Se utilizó validación cruzada (80/20 y 50/50) para asegurar la consistencia y robustez del modelo. Además, se aplicaron métricas de confiabilidad como MSE, MAE y R^2 para evaluar el rendimiento del modelo.

Aparte de la regresión lineal, se exploraron técnicas de Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos y optimizar el rendimiento del modelo. También se realizó un análisis de aprendizaje no supervisado mediante clustering para agrupar estudiantes según características comunes.

2. INTRODUCCION

El dataset presentado reúne información detallada sobre diversos factores que podrían influir en el rendimiento académico de los estudiantes, diseñado para analizar cómo variables sociales, económicas, personales y educativas afectan los puntajes en los exámenes, este conjunto de datos es ideal para estudios de aprendizaje y desarrollo académico.

Nuestro objetivo principal es identificar y comprender los elementos que tienen mayor impacto en los resultados académicos, siendo la columna 'Exam_Score' la que proporciona dicha información.

Esto puede ser útil para docentes, instituciones educativas y responsables de políticas para implementar estrategias que mejoren el desempeño estudiantil.

3. RESULTADOS

3.1. 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.

3.2. Factores analizados

- Factores personales siendo horas de estudio, hábitos de sueño, nivel de motivación y participación en actividades extracurriculares.
- Factores familiares como ser el ingreso económico, nivel educativo de los padres y su participación en el proceso educativo.
- Factores escolares, es el tipo de escuela, calidad de los maestros y la asistencia a las tutorías.
- Factores sociales como la influencia de los compañeros, acceso a recursos y la distancia entre la escuela y sus casas.

3.3. Objetivos secundarios

- Mejorar las prácticas pedagógicas desarrollando estrategias personalizadas para mejorar el aprendizaje.
- Establecer políticas públicas para desarrollar programas efectivos reduciendo así las brechas en el rendimiento.
- Identificar y apoyar a aquellos estudiantes que necesiten recursos adicionales o intervención educativa.

3.4. Valores nulos

El dataset por su gran cantidad de datos, presenta datos nulos que podrían afectar a la hora de realizar los preprocesamientos, manejar variables categóricas o normalizar los datos.
Las columnas 'Teacher_Quality', 'Parental_Education' y 'Distace_From_Home' presentan valores nulos, para ello:

- Si el porcentaje de valores nulos es bajo, rellenar con la media o con la moda.
- Si el porcentaje es alto, considerar eliminar las columnas evaluando la relevancia que estas tienen.

Los valores nulos podrían generar sesgos si no se manejan de forma adecuada.

3.5. Codificación de variables categóricas

Las columnas 'Gender', 'Parental_Involment' y 'School_Type' presentan información textual que se puede convertir a datos numéricos realizando:

- One-Hot Encoding para las columnas sin un orden inherente.
- Ordinal Encoding para columnas con datos que presentan algún orden lógico.

Esto realizamos puesto que, a la hora de realizar el análisis, no se puede realizar bajo datos categóricos, necesariamente hay que volverlos a datos numéricos.

3.6. Normalización de los datos

Al realizar el análisis es necesario que los datos una vez ya convertidos a datos numéricos, hay que pasar los datos por el

preprocesamiento para columnas como ser 'Hours_Studied', 'Attendance' y 'Exam_Score' realizando:

- Normalización para aquellos valores que están en distintos rangos, esto para algoritmos sensibles a escalas como ser KNN, regresión logística.
- Estandarización si se busca ajustar los valores a la distribución normal estándar, útil para modelos lineales.

Esto realizamos para mejorar la convergencia de algoritmos y evitar que variables de mayor rango predominen en los resultados.

3.7. Balanceo de los datos

La variable objetivo es 'Exam_Score' que dicha variable es continua, es decir, que no requiere un balanceo directo como tal, sin embargo, es necesario analizar su distribución.

Si los puntajes están sesgados hacia ciertos valores, es decir, que varios estudiantes tienen notas muy altas o muy bajas, siendo el caso así sería necesario realizar un balance.

Además, para realizar un balance sería necesario considerar:

- Submuestreo de la variable objetivo, cuando hay una distribución desigual, dividir los datos en rangos y asegurarse de que las clases estén equilibradas para ciertos análisis.
- Transformar los datos aplicando algoritmos o técnicas de suavizado si hay sesgos extremos en los puntajes.

Dado que el problema principal es de regresión, el balanceo tradicional como 'Oversampling' o 'Undersampling' no es aplicable el balanceo directo, sin embargo, si se transforma en un problema de clasificación, en esos casos sería necesario.

3.8. Exam_Score como variable objetivo

En el ámbito educativo es esencial entender los factores que inciden en el rendimiento académico, siendo nuestro objetivo identificar las variables que puedan predecir las puntuaciones que obtienen en los exámenes, siendo| representada en la columna 'Exam_Score'. Este es un claro ejemplo de regresión supervisada donde buscamos obtener un modelo que puede predecir valores continuos basándose en variables predictoras.

En dicha columna se aprecia de manera directa el éxito académico de un estudiante, siendo la variable más optima a la hora de realizar los análisis.

3.9. Preprocesamientos realizados y balanceo de datos

Para realizar el correcto análisis del dataset, es necesario aplicar algunos preprocesamientos para poder aplicar algoritmos o técnicas de análisis de datos.

3.9.1. ReplaceMissingValues

Este método convierte aquellas celdas nulas en valores bajo las condiciones descritas en punto **3.4**.

3.9.2. One_hot_Encoding

Tras haber realizado el relleno de valores nulos y para aplicar los algoritmos de análisis, es necesario que todos los datos categóricos sean datos numéricos, este método realiza ese preprocesamiento.

3.9.3. Normalize

Una vez ya no están presentes datos nulos, que los datos categóricos estén en datos numéricos, dichos datos numéricos tienen que estar en un rango que se obtiene mediante la siguiente fórmula:

$$\text{rango} = \frac{\text{valor} - \text{minimo\_valor}}{\text{mayor\_valor} - \text{minimo\_valor}}$$

3.10. Clasificador aplicado

El clasificador aplicado es la regresión lineal, esto para poder determinar qué clases podrían predecir el rendimiento académico de cualquier estudiante en base a los datos ya preprocesados de manera no supervisada.

Dicho clasificador lo hemos utilizado puesto que tras realizar y visualizar cómo se relacionan las siguientes clases con nuestra variable objetivo obtuvimos que:

- 'Hours_Studied' y 'Exam_Score' presentan una correlación positiva, siendo que, si aumentan las horas de estudio, el puntaje obtenido es mayor, básicamente son proporcionalmente directa la relación entre ambas variables.
- Con la clase 'Attendance' igualmente se presenta una correlación positiva, mientras mayores asistencias, más puntaje obtenido.
- 'Exam_Score' y 'Previous_Scores' presentan una correlación positiva, sin embargo, a diferencia de las anteriores relaciones, esta presente una variabilidad lo que podría reducir la precisión del modelo.
- Con 'Sleep_Hours' no hay una relación clara, presentándose una dispersión alta lo que implicaría que esta variable no es muy predictora.
- 'Motivation_Level' está representada en los siguientes rangos discretos (0.0, 0.5, 1.0, 2.0), sin embargo, se observa una ligera correlación positiva.
- 'Exam_Score' y 'Physical_Activity' no presentan una correlación.

La regresión lineal como clasificador es viable puesto que las variables 'Hours_Studied, Attendance, Previous_Scores' tienen correlaciones positivas con nuestra variable objetivo, además que, a regresión lineal es fácil de interpretar y eficiente al momento de realizar procesos de análisis.

Sin embargo, como hay variables con relaciones positivas, hay variables que presentan correlaciones que no son claras, lo que

implica regularizar dichas variables mediante 'Lasso Regresion' para mejorar la generalización.
Además, se realizó y visualizo la matriz de correlación entre todas las variables obteniendo:

- Hay variables influyentes siendo 'Hours_Studied, Attendance, Previous_Scores', hay variables con baja relación siendo 'Motivation_Level, Sleep_Hours, Physical_Activity, Internet_Access, Family_Income, Gender, Distance_from_Home' que no tienen una relación fuerte con la variable objetivo, además, las correlaciones entre variables independientes son bajas en su mayoría con valores menor a 0.3, lo que indica una baja multicolinealidad. Esto es importante para un modelo de regresión lineal, ya que minimiza el riesgo de redundancia entre predictores.

Realizando otros cálculos como la confiabilidad, R^2, MSE se puede evaluar si el modelo explica la variabilidad de nuestra variable objetivo.

3.11. Confiabilidad y matriz de confusión

Tras realizar todos los análisis anteriores, es necesario también calcular el error cuadrático medio o MSE, el error absoluto medio o MAE y el coeficiente de determinación R^2:

- El MSE mide la diferencia promedio al cuadrado entre los valores predichos por el modelo y los valores reales, un resultado bajo indica que las predicciones del modelo son más cercanas a los valores reales. Obtuvimos que el MSE es de 0.1154 siendo un resultado relativamente bajo lo que implica que el modelo tiene un rendimiento óptimo en cuanto a la predicción.
- El MAE mide la diferencia promedio entre las predicciones y los valores reales, un resultado más bajo indica que hay menor error promedio en las predicciones del modelo. Obtuvimos un MAE de 0.2848 lo que significa que las predicciones se desvían aproximadamente en 0.28 unidades a los datos reales, además que, el modelo tiene un buen rendimiento.
- El R^2 mide la proporción de la variabilidad de las variables dependientes que en nuestro caso sería 'Exam_Score', si se obtiene un valor de 1, el modelo explica perfectamente todos los datos, sin embargo, un valor nulo indica que no hay relación entre las predicciones y los valores reales. Obtuvimos un resultado de 0.9894 lo que indica que nuestro modelo tiene una precisión casi exacta.

3.12. Splits para validar el modelo

Con la mediana de confiabilidad que son los resultados calculados en el punto **3.11**, es necesario garantizar que nuestro modelo sea robusto garantizando que el modelo funcione en diferentes configuraciones, evitar el sobreajuste y garantizar la

generalización del modelo lo que significa que el modelo pueda predecir con datos distintos a los ya utilizados.

3.12.1. Split 80% para entrenar y 20% para prueba

En esta división es un estándar común en el aprendizaje automático entrenando el modelo con el 80% de los datos y utilizando el 20% restante para evitar que el modelo 'memorice' los datos.

3.12.2. Split 50% para entrenar y 50% para prueba

En esta división se realiza una evaluación más rigurosa puesto que los datos de entrenamiento son menos, esto sirve para probar que el modelo generaliza a la hora de predecir.

Con estos porcentajes para en entrenamiento y la prueba del modelo, utilizamos la validación cruzada con 100 asignaciones esto para evaluar el rendimiento asegurando que el modelo se evalué de manera consistente y se reduzca la variabilidad en las métricas debido a la partición inicial de los datos.

Tras realizar los cálculos de la mediana de ambos errores en ambas métricas obtuvimos que la media del MSE y del MAE bajo el Split (80/20) es de 0.1059 y 0.2684 respectivamente, la media del MSE y del MAE bajo el Split (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.

3.13. Primer código

https://colab.research.google.com/drive/117sVmBAhLI-AlGHQ2KbtM6ZH5nKSXa_9?usp=sharing

3.14. PCA

El Análisis de Componentes Principales (PCA) es una técnica utilizada para simplificar conjuntos de datos complejos reduciendo su número de dimensiones, mientras se mantiene la mayor cantidad posible de información relevante. Esta reducción de dimensionalidad facilita el procesamiento y análisis de los datos, eliminando redundancias y destacando las relaciones más importantes entre las características.

El primer paso en el proceso de PCA es centrar los datos. Este paso consiste en ajustar los datos para que estén centrados en torno al punto de origen, de forma que el promedio de cada característica sea igual a cero, este ajuste evita que las posiciones absolutas de los datos influyan en el análisis y permite estudiar las relaciones entre las características de manera más precisa.

A continuación, se calcula la matriz de covarianza, la cual mide cómo varían las características entre sí, la covarianza nos ayuda a identificar qué tan relacionadas están las variables entre sí.

Una vez obtenida la matriz de covarianza, el siguiente paso es calcular los valores propios y vectores propios, los vectores propios representan las direcciones principales en las que los datos varían, mientras que los valores propios indican la cantidad de variabilidad que cada vector propio captura. Estos autovectores son las direcciones más importantes que describen el comportamiento de los datos. Los autovectores más relevantes corresponden a las direcciones con mayor variabilidad, lo que permite comprender las principales fuentes de información en el conjunto de datos.
Luego, se seleccionan las componentes más importantes, aquellas asociadas con los autovalores más grandes, estos componentes forman la matriz de proyección, que es una representación reducida de los datos. Al seleccionar solo las componentes más significativas, reducimos la cantidad de información sin perder lo esencial. Este paso puede ser comparado con organizar el contenido de una maleta de manera eficiente, eligiendo las direcciones que permitan empaquetar la mayor cantidad de información en el menor espacio.
Finalmente, los datos se proyectan en este nuevo sistema de coordenadas. Esto implica transformar los datos originales en un formato más simple y reducido, pero que sigue conservando la mayor parte de la información relevante. Este paso permite representar los datos de manera más compacta y eficiente.
PCA ofrece varias ventajas clave. Primero, simplifica los datos al eliminar redundancias y facilita la identificación de patrones subyacentes. Además, reduce el ruido al enfocarse en la información más importante, mejora la visualización de los datos, permitiendo su análisis en 2D o 3D, y optimiza la eficiencia computacional al trabajar con menos dimensiones.
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 halla 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.

3.15. Aprendizaje no supervisado sin tomar en cuenta Exam_Score

Se realizó un análisis de aprendizaje no supervisado utilizando el algoritmo de clustering sobre un conjunto de datos relacionados con el rendimiento académico de los estudiantes. Este análisis tiene como objetivo agrupar a los estudiantes en función de sus características, sin la necesidad de tener una variable objetivo previamente definida. Las variables utilizadas incluyen las horas de estudio, la asistencia a clases, el nivel de involucramiento de los padres, el acceso a recursos educativos, la participación en

actividades extracurriculares, las horas de sueño, las calificaciones previas, el nivel de motivación, el acceso a Internet, el número de sesiones de tutoría, los ingresos familiares, la calidad del profesorado, el tipo de escuela, la influencia de los compañeros, la actividad física, las discapacidades de aprendizaje, el nivel educativo de los padres y la distancia desde la escuela.

Luego de realizar el clustering, se obtuvieron los centroides de los clusters, los cuales representan el "centro" de cada grupo en el espacio de características. Estos centroides reflejan el promedio de cada característica dentro de cada grupo. Por ejemplo, si un cluster tiene un promedio alto en las horas de estudio, esto indica que los estudiantes en ese grupo tienden a estudiar más que aquellos en otros grupos. El análisis de estos centroides proporciona información sobre qué características definen a cada grupo y cómo se agrupan los estudiantes según estos factores.

Además, se calculó el promedio de cada característica por cluster. Esto permitió identificar patrones dentro de cada grupo y observar diferencias significativas entre los clusters. Por ejemplo, si un cluster tiene un promedio más alto en motivación, podemos inferir que los estudiantes de ese grupo están más motivados en comparación con los de otros clusters. Estos promedios proporcionan una visión más clara de las características predominantes en cada grupo de estudiantes.

Finalmente, se utilizó la reducción de dimensionalidad a través del Análisis de Componentes Principales (PCA) para facilitar la visualización de los clusters. Este enfoque reduce la complejidad de los datos y permite representar los grupos de estudiantes de manera más comprensible. En el gráfico resultante, cada punto representa a un estudiante, y los colores indican a qué cluster pertenece. La separación entre los colores en el gráfico sugiere que los estudiantes se agrupan de manera significativa según sus características, lo que indica que existen grupos diferenciados en términos de comportamiento y rendimiento académico. Esta visualización ayuda a comprender cómo se estructuran los datos y cómo los estudiantes se agrupan según sus características académicas y personales.

3.16. Algoritmo de las n-reina

El algoritmo que utiliza el enfoque de recocido simulado para resolver el problema de las N reinas comienza con una solución aleatoria. Este problema consiste en ubicar N reinas en un tablero de ajedrez de N x N de manera que ninguna reina se pueda atacar a otra, lo cual implica que no deben compartir fila, columna ni diagonal.

Primeramente, el algoritmo genera una configuración inicial al azar, colocando las reinas en filas aleatorias dentro del tablero. Luego, calcula el "costo" de esta configuración, el cual

corresponde al número de amenazas entre las reinas. Las amenazas se definen cuando dos reinas comparten una misma columna o diagonal. El objetivo del algoritmo es minimizar este costo moviendo las reinas en el tablero y reduciendo el número de amenazas.

De igual forma, el enfoque de recocido simulado se inspira en el proceso físico de enfriamiento de un metal. Comienza con una temperatura alta, lo que permite aceptar soluciones peores con mayor probabilidad, lo que ayuda a evitar quedar atrapado en soluciones subóptimas (también conocidas como óptimos locales). Conforme avanza el proceso, la temperatura disminuye gradualmente, y el algoritmo se vuelve más selectivo, favoreciendo solo aquellas soluciones que realmente mejoren el estado actual.

Finalmente, el algoritmo termina cuando encuentra una solución sin amenazas (un costo de 0) o cuando ya no es posible mejorar la solución. Además, el código proporciona funciones para imprimir el tablero de ajedrez y mostrar las posiciones de las reinas de forma clara y comprensible. Si el algoritmo logra encontrar una solución válida, muestra el tablero con las reinas correctamente ubicadas; de lo contrario, informa que no se ha encontrado solución.

3.17. Segundo código

https://colab.research.google.com/drive/10RGDKqiOs0pyQ0_Ny9A-UjrlO3nr8-Yi?usp=sharing

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.