Introducción

Este análisis de datos tiene como objetivo predecir la variable "Nota\_03" en dos datasets que contienen información de los cursos de Cálculo Diferencial y Álgebra Vectorial de la Facultad de Ingeniería de la Universidad de Antioquia, los cuales contienen información de Características sociodemográficas de los estudiantes y Calificaciones de los tres parciales. El análisis de datos implica un proceso de exploración, limpieza, transformación, normalización y modelado de los datos para extraer su significado y valor, busca patrones, tendencias y correlaciones para comprender el pasado, presente y futuro. Además, provee una serie de beneficios al interior de las compañías como lo son; Apoya la toma de decisiones, identifica patrones y tendencias para anticipar cambios y planificar estrategias, facilita la predicción de resultados futuros, detecta irregularidades en los datos para la conservación y calidad de la información. Entre tanto, este proyecto tiene como objetivo predecir la variable "Nota\_03" para identificar a los estudiantes que podrían reprobar cada curso e Identificar las variables más importantes que influyen en el rendimiento académico e identificar factores que podrían afectar la aprobación de los cursos.

Descripción de los datos (AED)

El dataset "Cálculo Diferencial" contiene 397 registros y 33 variables (17 nominales y 16 numéricas). Se eliminaron dos registros duplicados para asegurar la integridad de los datos. El dataset no contiene valores nulos.

Tratamiento de outliers:

* Se aplicó un método gráfico para identificar outliers en las variables numéricas, contando valores únicos y calculando el Z-score.
* Se eliminaron los outliers con valores absolutos del Z-score mayores a 3.
* Se imputaron dos outliers en la variable "edad" con la mediana.
* Se transformó la variable "faltas" a booleana, creando las categorías "No falta a clase" y "Falta a clase".
* Se recategorizaron las variables "faltas" y "ausencias", imputando los 6 outliers de "ausencias" con la mediana.
* Se decidió no tratar el outlier en la variable "nota 02" (valor 0), pues se considera que brinda información relevante y el Z-score no muestra atípicos.

Análisis univariado:

* La variable objetivo (Nota\_03) no sigue una distribución normal según la prueba Shapiro-Wilk.
* El análisis estadístico de las variables numéricas muestra:
  + Edad promedio de 16-17 años con poca variabilidad.
  + Promedio de ausencias de 5-6, con alta dispersión (hasta 28 ausencias).
  + Promedio de Nota\_01 y Nota\_02 de 10-11, con distribución relativamente simétrica, aunque Nota\_02 tiene mayor dispersión y un mínimo de 0.
* Se consideran las categorías con menos del 5% en las variables categóricas como posibles ruidos, dependiendo del contexto.

Análisis bivariado:

* Se encontró alta correlación entre Nota\_01, Nota\_02 y la variable objetivo (Nota\_03).
* Se realizó una prueba de correlación entre Nota\_01 y Nota\_02 para verificar multicolinealidad.
* Se decidió eliminar Nota\_01, ya que tiene menor correlación con la variable objetivo.

Variables categóricas:

* Se calculó la correlación entre las variables categóricas utilizando la tabla de contingencia Cramér's V y el estadístico Chi-cuadrado.
* No se encontraron problemas de colinealidad o multicolinealidad, ya que las correlaciones entre las variables categóricas tenían un nivel de significancia por encima del 80%.

Transformación de variables:

* Se graficaron los valores únicos de cada variable para identificar variables que no podían ser booleanas.
* Se aplicó el método One-Hot transform para las variables categóricas con más de dos valores únicos, creando columnas binarias que indican la presencia o ausencia de la categoría.

Normalización de variables numéricas:

* Se normalizaron las columnas numéricas dividiendo cada valor por el valor máximo de la columna, lo que resulta en valores entre 0 y 1.
* La variable a predecir "nota\_03" no fue normalizada.

Se aplicó un modelo de regresión lineal al dataset, realizando los siguientes pasos:

1. Análisis de la variable objetivo: Se graficó la variable objetivo "nota\_03" y se observó una asimetría a la derecha. Para corregir esto, se aplicó una transformación logarítmica, lo que ayudó a estabilizar la varianza y a tratar los datos sesgados.
2. Separación de datos: Se dividieron los datos en conjuntos de entrenamiento (train) y prueba (test) para entrenar y evaluar el modelo.
3. Entrenamiento y predicción: Se entrenó el modelo de regresión lineal utilizando los datos de entrenamiento y se realizaron predicciones sobre los datos de prueba.
4. Evaluación del modelo: Se calcularon métricas para evaluar el rendimiento del modelo tanto en los datos de entrenamiento como en las predicciones.

Table 1-Métricas del modelo de regresión lineal

|  |  |  |
| --- | --- | --- |
| METRICA | TEST | TRAIN |
| Accuracy | 0.6810255802594993 | 0.6810255802594993 |
| MSE | 0.23 | 0.16 |
| RMSE | 0.48 | 0.41 |
| MAE | 0.34 | 0.29 |
| MAPE | 495242212174736.12 | 338203402723446.75 |
| R2 | 0.68 | 0.70 |
| R2-adjusted | 0.45 | 0.64 |

El modelo de regresión lineal fue entrenado y evaluado utilizando datos de entrenamiento y prueba. Los resultados muestran que:

* Accuracy: El modelo obtuvo la misma precisión (Accuracy) en ambos conjuntos de datos.
* MSE y RMSE: La media del error cuadrático (MSE) y la raíz del error cuadrático medio (RMSE) disminuyeron en un 7% en los datos de prueba en comparación con los datos de entrenamiento, lo que indica un mejor ajuste del modelo y una mejor predicción de los valores reales.
* MAE: El error absoluto medio (MAE) también disminuyó en un 5%, lo que significa que la diferencia entre los valores predichos y los valores reales es menor.
* R2: El coeficiente de determinación (R2) mejoró del 68% al 70% en los datos de prueba, pero aún se puede mejorar.
* R2 ajustado: El R2 ajustado mejoró significativamente en 19 puntos, alcanzando un 64% de precisión, lo que indica una mejor estimación de la precisión del modelo de regresión lineal para la predicción de la variable objetivo.

El dataset "Álgebra Vectorial" contiene 651 registros y 33 variables (17 nominales y 16 numéricas). Se eliminaron dos registros duplicados y no se encontraron datos nulos.

Análisis de las variables numéricas:

* La edad de los estudiantes oscila entre 17 y 22 años, con poca variabilidad.
* El máximo de faltas a clase es 3.
* El máximo de ausencias es 32.
* Las calificaciones promedio de Nota\_01 y Nota\_02 son 11, con un mínimo de 0 y un máximo de 19.

Tratamiento de outliers:

* Se aplicó un método gráfico y se calculó el Z-score para identificar outliers.
* No se consideraron outliers para la variable "faltas" ya que solo había un registro entre 2 y 3 faltas.
* Se transformó la variable "faltas" a booleana, creando las categorías "No falta a clase" (0 faltas) y "Falta a clase" (1, 2 o 3 faltas).

Análisis de la variable objetivo:

* Se graficó la variable objetivo y se observó una asimetría a la derecha.
* La prueba de Shapiro-Wilk indicó que la variable objetivo no sigue una distribución normal, ya que el valor p (p-value) fue menor a 0.05.

Análisis de variables categóricas:

* Se graficaron las variables categóricas para observar el porcentaje de cada categoría.
* No se consideró necesario reagrupar ninguna variable categórica.

Análisis de variables booleanas:

* Se estableció un umbral de 80-20 para determinar la significancia de las variables booleanas.
* Se eliminarán del modelo las variables booleanas que se distribuyan por encima del 80% o por debajo del 20%, ya que no se consideran significativas.

Análisis de variables numéricas:

* Se convirtió en variables dummy las variables categóricas.
* Se encontró alta correlación entre Nota\_01, Nota\_02 y la variable objetivo.
* Se realizó una prueba de correlación entre Nota\_01 y Nota\_02 para verificar multicolinealidad.
* Se eliminó Nota\_01 por tener menor correlación con la variable objetivo.

Análisis de variables categóricas:

* Se empleó la tabla de contingencia Cramér's V para calcular el estadístico Chi-cuadrado y medir la asociación entre las variables categóricas.
* No se encontraron problemas de colinealidad o multicolinealidad entre las variables categóricas ordinales, ya que las correlaciones tenían un nivel de significancia por encima del 80%.

Variables categóricas:

* Se verificó la correlación entre las variables categóricas booleanas y no se encontraron problemas de colinealidad o multicolinealidad (nivel de significancia > 80%).
* Se aplicó el método One-Hot a las variables categóricas con más de dos valores únicos, creando columnas binarias para indicar la presencia o ausencia de la categoría.
* Se identificó que algunas categorías tenían una baja representación, lo que podría afectar la capacidad predictiva del modelo.
* Se agrupó valores con características similares dentro de las variables categóricas con baja representación, mejorando la proporción de datos en cada categoría y creando clases más equilibradas.
* Se renombraron las nuevas variables agrupadas.

Normalización de variables numéricas:

* Se normalizaron las columnas numéricas dividiendo cada valor por el máximo valor de la columna, lo que resultó en valores entre 0 y 1.
* La variable a predecir "nota\_03" no fue normalizada.

Análisis univariado de nuevas variables booleanas:

* Se estableció un umbral de 80-20 para determinar la significancia de las variables booleanas.
* Se eliminarán del modelo las variables booleanas que se distribuyan por encima del 80% o por debajo del 20%, ya que no se consideran significativas.

Modelo de regresión logistica:

Se implementó un modelo de regresión logística para predecir la variable objetivo. Para evaluar el rendimiento del modelo, se utilizó la matriz de confusión, obteniendo los siguientes resultados:

* Verdaderos positivos (VP): 89. El modelo predijo correctamente 89 casos como 1 cuando realmente eran 1.
* Falsos positivos (FP): 19. El modelo predijo erróneamente 19 casos como 1 cuando realmente eran 0.
* Falsos negativos (FN): 6. El modelo predijo erróneamente 6 casos como 0 cuando realmente eran 1.
* Verdaderos negativos (VN): 81. El modelo predijo correctamente 81 casos como 0 cuando realmente eran 0.

Ahora, evaluamos las siguientes métricas del testeo con resultados así:

Table 2-Métrica del modelo de regresión logística

|  |  |  |
| --- | --- | --- |
| METRICA | TEST | TRAIN |
| Accuracy test | 0.841025641025641 | 0.8876651982378855 |
| Precicion | 0.8076923076923077 | 0.8768656716417911 |
| Recall (Sensibilidad) | 0.8842105263157894 | 0.9288537549407114 |
| F1-score | 0.8442211055276382 | 0.9021113243761996 |
| Especificidad | 0.8 | 0.835820895522388 |

Se evaluó el rendimiento del modelo de regresión logística utilizando diversas métricas. La comparación con los resultados anteriores muestra:

* Accuracy: Disminuyó en 8 puntos porcentuales, pero esto se debe a un cambio en la cantidad de verdaderos positivos y negativos, lo que no necesariamente indica un mal rendimiento.
* Precisión: Aumentó en 7 puntos porcentuales, lo que significa que el modelo identifica mejor los verdaderos positivos.
* Sensibilidad: Aumentó del 88% al 92%, indicando que el modelo es más preciso para detectar los casos positivos.
* F1-score: Aumentó 6 puntos alcanzando un 90%, lo que se considera una buena calificación y muestra que el modelo cumple con su propósito.
* Especificidad: Aumentó 3 puntos, llegando al 83%, indicando una mejor detección de los casos negativos.

Curva ROC: La curva ROC es alta con un área bajo la curva amplia, lo que indica un buen rendimiento del clasificador.

Comparación de los modelos

Ambos modelos muestran buen rendimiento, aunque la regresión logística parece tener mejor rendimiento en términos de precisión, sensibilidad, F1-score y especificidad.

* Accuracy: La regresión logística tiene una precisión significativamente mayor (0.841 vs. 0.681) en el conjunto de prueba.
* MSE y RMSE: La regresión lineal tiene valores más bajos de MSE y RMSE, lo que indica un mejor ajuste del modelo.
* MAE: La regresión lineal también tiene un MAE ligeramente menor, lo que significa que la diferencia entre los valores predichos y reales es menor.
* R2: La regresión lineal tiene un R2 ligeramente mejor en el conjunto de prueba (0.68 vs. 0.70).
* R2 Ajustado: La regresión lineal tiene un R2 ajustado significativamente mejor en el conjunto de prueba (0.45 vs. 0.64).
* Precisión, Sensibilidad, F1-score y Especificidad: La regresión logística supera a la regresión lineal en todas estas métricas de clasificación, indicando un mejor rendimiento en la predicción de la variable objetivo.
* Para este problema, la regresión logística es un mejor modelo que la regresión lineal, ya que tiene un mayor Accuracy, mejores métricas de clasificación y se adapta mejor a la naturaleza de la variable objetivo. La regresión lineal, aunque muestra un buen ajuste según los valores de MSE y RMSE, tiene un Accuracy inferior y no es tan adecuado para este tipo de problema.

Aplicación de otros algoritmos de Machine Learning

Preparación del notebook:

* Se preparó el notebook para realizar el análisis de datos.

Selección univariante:

* Se aplicaron métodos de selección univariante para identificar las variables más relevantes para los modelos de regresión lineal, como f regression que nos entrega formación valiosa sobre qué variables considerar para la construcción de un modelo, mejorando así la calidad de las predicciones y la eficiencia del análisis. Y el método mutual info regression que nos ayuda en la construcción y ajuste de modelos más efectivos y mejora la comprensión del problema que estás abordando.

Comparación de métodos:

* Los resultados del método "f-regression" fueron similares al modelo original.
* El método "mutual info regression" mostró un rendimiento significativamente mejor, con aumentos en el R2, R2 ajustado y el accuracy, así como una disminución de los errores.
* Se seleccionaron las 30 mejores variables utilizando los métodos de selección de variables.

Selección de variables para el dataset "Cálculo":

* Variance Threshold: Selecciona características con varianza superior a un umbral.
* Lasso: Utiliza un método de penalización para reducir la influencia de variables irrelevantes.
* Ridge: Similar a Lasso, pero con una penalización diferente.

Optimización de hiperparámetros:

* Se optimizaron los hiperparámetros Alpha, lambda y score para los métodos Lasso y Ridge.

Se implementaron diferentes métodos de selección de variables para mejorar el rendimiento de los modelos de regresión lineal. Se encontró que el método "mutual info regression" ofreció un mejor rendimiento en comparación con el modelo original. Se optimizaron los hiperparámetros de Lasso y Ridge para obtener mejores resultados.