

Proyecto entrega 1

(1) Problema Predictivo a Resolver:

El objetivo es predecir el rendimiento de los estudiantes en función de diversas características, como género, nivel de educación de los padres, tiempo de estudio, asistencia, etc. Esto permitirá identificar a los estudiantes que pueden necesitar apoyo adicional y mejorar el proceso de enseñanza.

(2) Dataset Utilizado:

El dataset que vamos a utilizar es el "Student Performance Data" de Kaggle, que contiene información sobre el rendimiento de los estudiantes. Puedes encontrar el dataset en el siguiente enlace: <https://www.kaggle.com/devansodariya/student-performance-data>

Este dataset incluye las siguientes columnas representativas:

Género del estudiante (por ejemplo, "gender").

Nivel de educación de los padres (por ejemplo, "parental level of education").

Tiempo de estudio semanal (por ejemplo, "study time").

Asistencia a clases (por ejemplo, "attendance").

Notas en diferentes materias (por ejemplo, "math score", "reading score", "writing score").

(3) Métricas de Desempeño Requeridas:

Matriz de Confusión: Utilizaremos la matriz de confusión para evaluar el rendimiento de los modelos de clasificación. Esto nos permitirá conocer la precisión, el recall, la especificidad y la precisión de los modelos en la predicción del rendimiento de los estudiantes.

Cross Validation: Utilizaremos la validación cruzada para evaluar la capacidad de generalización de nuestros modelos. Esto nos ayudará a evitar el sobreajuste y a estimar el rendimiento real del modelo en datos no vistos.

(4) Desempeño Deseable en Producción:

El desempeño deseable en producción puede ser definido de la siguiente manera:

Para la matriz de confusión, buscamos un modelo que tenga un alto valor de precisión y recall para predecir el rendimiento estudiantil. Específicamente, deseamos minimizar los falsos negativos (estudiantes que necesitan apoyo pero no se identifican) y los falsos positivos (estudiantes que no necesitan apoyo pero se identifican incorrectamente).

Para la validación cruzada, esperamos que el modelo mantenga un buen rendimiento en datos no vistos y que no haya una brecha significativa entre el rendimiento en los datos de entrenamiento y en los datos de prueba.

En última instancia, el éxito en la producción dependerá de cómo se integre el modelo en la operación de la institución educativa. Si el modelo puede identificar eficazmente a los estudiantes que necesitan apoyo y esto conduce a una mejora en el rendimiento académico, entonces se consideraría exitoso. Sin embargo, el aumento en las ventas no es aplicable en este contexto, ya que estamos tratando con una institución educativa y no un negocio.

Este proyecto se enfoca en evaluar el rendimiento de los modelos de machine learning en la predicción del rendimiento estudiantil, con la esperanza de mejorar el proceso educativo y brindar apoyo a los estudiantes que lo necesitan.