

Antes que nada, nos alegra muchísimo que estés participando de este challenge.
¡Felicitaciones y mucha suerte!

Desafío Propuesto:

Contexto para analistas:

Una institución educativa desea analizar el rendimiento de sus estudiantes y desarrollar un modelo predictivo para identificar factores que influyen en las notas finales. El dataset proporcionado incluye datos históricos sobre el cursado de los estudiantes y actividades como entrega, exámenes rendidos y vencimiento de tareas. La variable objetivo es `nota_final_materia`, donde una nota mayor o igual a 5 indica aprobación.

Un aspecto importante del dataset es que incluye distintas "actividades" registradas para el mismo periodo de tiempo. Esto significa que puede haber múltiples registros para una misma partición temporal, dependiendo de las diversas actividades realizadas por el alumno. Les recomendamos agrupar por `user_uuid/course_uuid/particion`
Cuidado! Hay algunas inconsistencias que esperamos tomen decisiones y nos cuenten cuáles y porqué.

Objetivos del Desafío:

1. Análisis Exploratorio de Datos (EDA):

- **Analistas de Datos:** Realizar un análisis descriptivo para identificar tendencias y patrones en el rendimiento académico, y presentar los hallazgos en un tablero interactivo de Power BI.
- **Data Scientists:** Además del EDA básico, aplicar técnicas estadísticas avanzadas para inferir relaciones y significancia.
- **ML Engineers:** Colaborar en la preparación de los datos y asegurar que están listos para el modelado, además de realizar un EDA básico.

2. Ingeniería de Características:

- **Analistas de Datos:** Sugerir posibles características basadas en el análisis descriptivo.
- **Data Scientists:** Crear nuevas características a partir de los datos existentes y evaluar su impacto en modelos predictivos.
- **ML Engineers:** Optimizar la creación y transformación de características para asegurar eficiencia en el entrenamiento de modelos.

3. Modelado Predictivo:

- **Analistas de Datos:** No se espera que desarrollen un modelo predictivo; su enfoque está en la visualización de datos.
- **Data Scientists:** Desarrollar y comparar múltiples modelos predictivos (regresión, árboles de decisión, etc.) y optimizarlos usando técnicas avanzadas.
- **ML Engineers:** Implementar y escalar modelos complejos, optimizar el rendimiento y realizar validación cruzada.

4. Implementación y Escalabilidad:

- **Analistas de Datos:** Crear un tablero de Power BI interactivo que muestre los hallazgos del análisis descriptivo y permita la exploración de los datos.
- **Data Scientists:** Presentar un modelo final con métricas de rendimiento y justificar la elección del modelo.

- **ML Engineers:** Desarrollar una solución escalable para la implementación del modelo en un entorno de producción, asegurando eficiencia y robustez. Uso de buenas practicas

Diccionario:

VARIABLE	DESCRIPCIÓN
user_uuid	Índice de usuario de la plataforma
course_uuid	Índice de curso
legajo	Legajo administrativo del alumno
particion	Partición temporal del 1 a 60 que serían los días de cursado
periodo	Semestre más año de cursado(Ejemplos: 1-2023 equivale al primer Semestre del año y 2-2023 equivale al segundo
nota_final_materia	Última nota del examen final
nota_parcial	Nota del examen parcial
nombre_examen	Tipo de examen rendido
course_name	Nombre del curso
fecha_mesa_epoch	Fecha de mesa de examen en formato epoch
assignment_id	Id de tarea
ass_name	Nombre de la tarea
ass_created_at	Fecha de creación de la tarea
ass_due_at	Fecha de vencimiento de la tarea
ass_unlock_at	Fecha de apertura de la tarea
ass_lock_at	Fecha de cierre de tarea
points_possible	Puntos posibles por tarea
ass_name_sub	Nombre de la entrega
sub_uuid	Id de entrega de tarea
score	Puntaje de la entrega
submission_type	Tipo o formato de tarea entregada
s_submitted_at	Fecha de entrega
s_graded_at	Fecha de corrección
s_created_at	Fecha de creacion de la entrega

Entregables Esperados:

- **Analistas de Datos:**
 - Informe detallado del EDA.
 - Propuesta de características relevantes.
 - Tablero de Power BI interactivo con visualizaciones clave y capacidad de exploración de datos. Modelo de datos.
- **Data Scientists:**
 - Informe del EDA con análisis estadístico avanzado.
 - Nuevas características creadas y justificación.

- Comparación de múltiples modelos predictivos con métricas de rendimiento.
- **ML Engineers:**
 - Código optimizado para la preparación de datos.
 - Implementación del mejor modelo predictivo.
 - Documentación sobre la escalabilidad y eficiencia del modelo.

Evaluación:

- **Análisis de Datos:** Claridad, profundidad y precisión del análisis exploratorio, y efectividad del tablero de Power BI. Proporciona una documentación clara y concisa sobre el proceso de desarrollo y decisiones tomadas.
- **Modelado:** Calidad de los modelos predictivos y justificación de la selección. Proporciona una documentación clara y concisa sobre el proceso de desarrollo y decisiones tomadas.
- **Implementación:** Eficiencia, escalabilidad y robustez del código y la solución final. Uso de buenas practicas. Proporciona una documentación clara y concisa sobre el proceso de desarrollo y decisiones tomadas.

Link al dataset:

<https://drive.google.com/file/d/1SuDzjGnovzUzu9LBktRyx3IgUvVboAiK/view?usp=sharing>

Formas de envío:

Puedes agregar mi mail al repo de github: karina.alem@edmachina.com

Puedes comprimir el proyecto y adjuntarlo en un mail

Puedes subirlo a alguna plataforma de almacenamiento y compartirme el link

Una vez finalizado:

Agendaremos una instancia para la defensa del challenge y continuar el proceso.

Desde ya muchas gracias por tu interés y buena suerte!