MACHINE LEARNING PARCIAL (20%)

Docente: Juan Sebastián Parra Sánchez

Estás participando para ser parte del equipo de Datos y Analítica de una empresa financiera, en donde hay una posición para Científico de Datos Junior.

Se te entrega una base de datos con información del crédito, el usuario de dicho crédito y el pago del mismo(si se hizo a tiempo o cayó en mora). Se quiere evaluar su capacidad para generar insights valiosos para la organización a partir de estos datos y, además, su capacidad para comunicarlos adecuadamente.

Le entregan el DataFrame pero no un diccionario de datos, por lo que debes hacer uso de tu creatividad, búsqueda en internet, etc para entender mejor el negocio y las variables que allí se tienen.

Este entregable debe entregarse en .ipynb y en parejas. Una sola persona entrega ANTES de la fecha límite por el campus virtual.

¿Qué debe tener el archivo?

1. **Carga de datos**
2. **Exploración inicial de datos**
   2.1 Descripción general de los datos. Caracterización de los datos: categóricos, numéricos, ordinales, nominales, dicotómicos, politómicos. Revisión de nulos.
   2.2 Eliminación de variables irrelevantes.
   2.3 Convertir los datos en su tipo correcto (numéricos, categóricos, booleanos, fechas, etc) y corrección de los datos si es necesario, para cada columna tenga un tipo de dato uniforme.
3. **Exploración de datos y descripción (EDA)**

   3.1 Análisis univariable.
   Se recomienda hacer un describe() una vez se hayan ajustado los tipos de datos.
   Se recomienda para las variables numéricas: histograma para distribución, boxplot, tablas pivote.
   Para las variables categóricas: countplot, tablas pivote, value_counts()
   Descripción estadística (dependiendo del tipo de dato): por ejemplo para los numéricos: tendencia Central (media, mediana, moda, max, min, etc), medidas de dispersión (Rango, IQR, cuartiles, varianza, desviación estándar.

   3.2 Análisis bivariable.
   Se recomienda generar gráficos y tablas con respecto a la variable objetivo y comentar.

El análisis debe contener:

* Compresión detallada de las características y el esquema de datos. Muchos comentarios, interpretaciones, análisis de lo que se encuentra.
* Analice los datos para crear las reglas de validación de los datos que serán usadas en otra etapa del proyecto. Por ejemplo, rango de edades.
* Para los proyectos de aprendizaje supervisado, identifique los atributos objetivo (target).

4. Feature Engineering

4.1 Limpieza de datos.

***** Completar los valores faltantes (por ejemplo, con cero, media, mediana ...) o eliminar las filas (o columnas). Indique claramente por qué.

***** Corregir o eliminar valores atípicos. Los valores atípicos pueden separarse del dataset dependiendo del problema del proyecto (por ejemplo, detección de anomalías).

* Descartar los atributos que no proporcionan información útil para el proyecto. O son redundantes de acuerdo con el EDA.

4.3 Transformación de variables:

- Encode variables categóricas, texto y las que sean necesarias para poder ser usadas para el modelamiento (LabelEncoder, get_dummies, replace, etc). Crear todas estas transformaciones usando transformadores y pipelines de scikit-learn

5. Modelado

- Generar un modelo supervisado de Machine Learning (árbol de decisión, k-vecinos, etc) y seleccionar el mejor modelo.
- Comentar las métricas de evaluación del modelo. Escoger una sola métrica especificando por qué.

6. Registro de pruebas

- Para un nuevo usuario inventar un registro y generar la salida que predice el modelo.