Análisis de datos

Preprocesamiento

Esta actividad tiene como propósito realizar los procesos de limpieza y transformación sobre los datos disponibles en:

- https://github.com/jpospinalo/MachineLearning/blob/main/Logistic%20Regression/german_credit_data.csv
-

A continuación se presenta una descripción de las columnas disponibles:

Columna	**Descripción**
Age	Edad del solicitante en años.
Sex	Género del solicitante (masculino o femenino).
Job	Tipo de empleo del solicitante (sin empleo, calificado, altamente calificado, etc.).
Housing	Tipo de vivienda (propia, alquilada, gratuita).
Saving accounts	Nivel de ahorros del solicitante (ninguno, bajo, moderado, alto).
Checking account	Estado de la cuenta corriente (ninguno, bajo, moderado, alto).
Credit amount	Monto del crédito solicitado en unidades monetarias.
Duration	Duración del crédito en meses.
Purpose	Motivo del préstamo (ej. automóvil, educación, hogar, etc.).
Risk	Variable objetivo: good = buen pagador, bad = mal pagador.

Su objetivo es realizar el preprocesamiento necesario que permitirá posteriormente utilizar estos datos para entrenar un modelo predictivo. Como parte del proceso usted deberá realizar las siguientes tareas:

A. **Limpieza de datos:**

- Construya un `DataFrame` con la información suministrada
- Verifique el tipo de dato para cada columna y compruebe si es coherente con la información disponible en los campos (por ejemplo, un campo como “edad” debería ser de tipo numérico y no una cadena de caracteres). Para realizar esta verificación puede utilizar el método `info()`. Realice los cambios que considere necesarios.
- Para cada columna, verifique posibles valores duplicados y problemas de formato.
- Para cada columna, verifique si existen valores faltantes. En caso de encontrarlos, determine cuál sería el tratamiento más adecuado considerando el tipo de dato (numérico o categórico). Considere si debido a la cantidad de datos faltantes en una columna, la mejor estrategia podría ser descartarla

B. **Transformaciones:**

En esta etapa se debe realizar cada etapa de transformación de manera manual. El objetivo es reconocer el funcionamiento básico de algunas de las clases disponibles en el módulo de preprocesamiento de sklearn.

- Para las columnas de tipo numérico compruebe la presencia de valores atípicos y decida que tratamiento se les dará. Considere los dos métodos presentados en clase.
- Realice el proceso de imputación para los valores faltantes según la estrategia seleccionada en el punto anterior. Para esto, utilice la clase SimpleImputer disponible en sklearn
- Realice el proceso de escalado de los datos numéricos. Seleccione entre escalado MinMax o el StadarScaler disponibles en sklearn
- Realice el proceso de codificación para datos categóricos según sea necesario. Para esto, utilice las clases OrdinalEncoder y OneHotEncoder disponibles en sklearn

C. **Pipelines de preprocesamiento.**

En esta etapa el objetivo es construir un pipeline que encapsule todo el proceso de transformación. Esto permitirá automatizar este proceso a la hora de realizar predicciones sobre un nuevo conjunto de datos.

- Construya un pipeline que incluya el proceso de transformación para valores numéricos. Incluya todas las transformaciones necesarias. Para esto, utilice la clase Pipeline disponible en Sklearn.
- Construya un pipeline que incluya el proceso de transformación para valores categóricos ordinales. Incluya todas las transformaciones necesarias. Para esto, utilice la clase Pipeline disponible en Sklearn.
- Construya el pipeline final de preprocesamiento combinando los dos pipelines generados anteriormente. Para esto, utilice la clase ColumnTransformer disponible en sklearn
- Imprima las primeras 10 líneas de los datos transformados.

Para tener en cuenta

- Los ejercicios deben entregarse en un notebook.
- El proceso de preprocesamiento debe ser documentado en cada etapa. Esto será tenido en cuenta durante la calificación