Pasos que realizar para el preprocesamiento:

Limpieza del dataset

La mayor parte de los algoritmos de aprendizaje automático tiene problemas o no funcionan bien cuando tenemos columnas con datos nulos. Es decir si algunas filas tiene datos y otras no.

Si fuera así deberíamos de seguir los siguientes pasos:

- Eliminar los datos con los valores nulos, reduciendo así el número de datos.
- Eliminar por completo, la columna que contiene valores nulos, reduciendo el número de columnas del dataset.
- Asignar un valor al campo de la columna con valores nulos, por ejemplo el valor medio (mediana).

Preprocesando datos con texto en el dataset

En general, los algoritmos de machine learning no trabajan bien con datos con valores que no sean numéricos. Es por ello, que aquellos datos con valores que contengan un texto o sean una categoría, debemos transformarlos a valores numéricos.

Ejemplo de columnas no numéricas: «Color» que nos indica el color en inglés (Red, Blue, etc) y «Spectral_Class» que es otro dato categórico (M, O, A, etc).

- En el primer ejemplo del color, se realiza una transformación directa de cada valor del atributo de texto a un número correlativo. Es decir, si tenemos los valores «A, B, C, A, …» pasar los mismos valores a valor numérico «0, 1, 2, 0, …».

- A veces cuando no existe una relación de cercanía entre los valores de texto de un dato, no es buena idea pasar directamente a valores numéricos consecutivos. En estos casos, lo que hacemos es crear un dato con valores binarios para cada valor de texto. Por ejemplo, en la columna «COLOR» tenemos el valor «RED». Así, creamos un dato binario de nombre «RED» siendo 1 para las instancias cuyo color sea «RED» y 0 en otro caso.

Escalado de los datos

Este será una de las principales tareas que realicemos en el preprocesamiento de datasets. Los principales algoritmos de machine learning que existen no funcionan muy bien cuando existen una gran diferencia entre los valores de una columna. Si tenemos una columna «L» que contiene valores muy distante, por ejemplo, tiene un valor mínimo de 0 y máximo de 849820. Esto es algo que debemos evitar, ya que los algoritmos de aprendizaje no funcionan bien en estos casos.

Existen dos formas claras de solucionar estos problemas de escalas: la normalización de valores y la estandarización.

- **Normalización**

El escalado **min-max** o normalización, es una técnica común a la hora de solucionar el problema de tener diferentes escalas en los valores de una columna. El objetivo que se consigue con esta técnica es que todos los valores de una columna estén comprendido en el intervalo [0-1]. De forma matemática, lo que estamos haciendo es a cada valor le restamos el mínimo y lo dividimos entre el valor máximo.

Otras maneras de utilizar la normalización, además de min-max son:

- Normalización Z-score
- Normalizado por escala decimal

- **Estandarización**

La otra forma de realizar el escalado de valores que vamos a ver se denomina estandarización. Matemáticamente hablando, lo que estamos haciendo en este proceso es restar la media de los valores y dividir por la desviación estándar de los mismos. De esta forma los valores obtenidos tendrán una media de cero y una varianza de uno.

Existen algunas diferencias notables con respecto a la normalización. En primer lugar los valores obtenidos por la estandarización no están acotados en ningún rango ([0-1] por ejemplo).

En segundo lugar, este método consigue solucionar el problema de valores atípicos u outliers que presenta la normalización. Por ejemplo, supongamos que un atributo de temperatura contiene valores entre 0 y 100 por regla general. Por un error de medición, tenemos un valor de 10000 que se consideraría un outlier. Si aplicamos la normalización, la mayor parte de valores estarían en el rango 0-0.1. Sin embargo, esto no se da con la estandarización.

Así pues podemos encontrarnos con distintas etapas del preprocesamiento:

- **Data cleaning:** la limpieza de datos elimina ruido y resuelve las inconsistencias en los datos.
- **Data integration:** con la Integración de datos se migran datos de varias fuentes a una fuente coherente como un Data Warehouse.
- **Data transformation:** la transformación de datos sirve para normalizar datos de cualquier tipo.
- **Data reduction:** la reducción de datos reduce el tamaño de los datos agregandolos.

ETL - Extract, Transform, Load

Las herramientas ETL, ya poseen la mayoría de las técnicas de procesamiento de datos mencionadas anteriormente como la migración de datos y la transformación de datos, esto hace que el seguimiento de estas prácticas de limpieza de datos resulte mucho más conveniente. Además, tales herramientas ETL permiten a los usuarios especificar los tipos de transformaciones que desean realizar con sus datos.

LIBRERÍA SCIKIT-LEARN (sklearn)

Para realizar este paso podemos ir a la librería sklearn para ver los distintos tipos disponibles: https://scikit-learn.org/stable/modules/preprocessing.html