1. Procesamiento de datos.

Limpieza de datos.

La limpieza de datos es un paso fundamental en el preprocesamiento de datos en el machine learning. Este proceso implica identificar y corregir o eliminar datos incorrectos, incompletos, duplicados o inconsistentes en un conjunto de datos. Aquí se explica más detalladamente la limpieza de datos:

1. **Identificar datos incorrectos o erróneos:** Durante la recopilación de datos, es posible que se hayan introducido errores o que los datos estén mal formateados. En este paso, se deben identificar y corregir los errores obvios, como valores que están fuera de rango o que no se ajustan al formato esperado. Por ejemplo, si se tiene un conjunto de datos que contiene información sobre el género de los empleados de una empresa, los datos incorrectos serían los que no son ni "masculinos"ni "femeninos".

2. **Tratamiento de datos faltantes:** Los conjuntos de datos a menudo contienen valores faltantes, ya sea porque no se recopilaron o porque se perdieron durante el proceso de almacenamiento o transferencia. Los valores faltantes pueden afectar el rendimiento de los modelos de machine learning, por lo que es importante abordarlos. Esto implica decidir si se deben eliminar las instancias con valores faltantes, imputar los valores faltantes utilizando técnicas de imputación (como el promedio, la mediana o modelos más avanzados), o considerarlos como una categoría separada.

3. **Eliminación de datos duplicados:** En algunos casos, puede haber instancias duplicadas en el conjunto de datos. Estas instancias duplicadas no aportan información adicional y pueden sesgar los resultados. Por lo tanto, es esencial identificar y eliminar las instancias duplicadas para garantizar la integridad y la calidad de los datos.

4. **Resolución de inconsistencias y valores atípicos (outliers):** Los valores atípicos son observaciones que difieren significativamente del resto de los datos. Pueden ser resultado de errores de medición o indicar situaciones inusuales. Es importante evaluar si los valores atípicos deben ser corregidos, eliminados o si contienen información relevante y deben mantenerse.

5. **Normalización y estandarización:** Los datos pueden tener diferentes escalas y rangos, lo que puede afectar el rendimiento de algunos algoritmos de machine learning. En este paso, se pueden aplicar técnicas de normalización o estandarización para asegurar que los datos tengan una distribución más uniforme y comparable

6. **Manejo de datos desbalanceados:** En algunos problemas de clasificación, puede haber una falta de equilibrio entre las clases objetivo, lo que significa que una clase puede tener muchos más ejemplos que las demás.

En estos casos, se deben utilizar técnicas de muestreo o ponderación para abordar el desequilibrio y evitar que el modelo se sesgue hacia la clase mayoritaria

1.1. Transformación y normalización de datos

La transformación y normalización de datos son técnicas utilizadas en el preprocesamiento de datos en machine learning para ajustar los datos a una escala o distribución específica. Estos procesos son importantes para garantizar que los datos sean adecuados para su uso en algoritmos de machine learning y que no se vean afectados por diferencias en las unidades o escalas de las características. A continuación, se explica con más detalle la transformación y normalización de datos: