INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Ingenieria en Inteligencia Artificial, Machine Learning

Complejidad de datos

Jiménez Hernández Vicente David.

jimenez.hernandez.vicente.david@gmail.com

1.Introduccion

En esta práctica describimos el análisis de un dataset relacionado con el cáncer de mama, aplicado para la identificación de patrones y la preparación de los datos para tareas de machine learning. El dataset presenta retos comunes en ciencia de datos, como valores faltantes y un desbalance de clases, los cuales pueden afectar significativamente el desempeño de los modelos predictivos. Por lo tanto, se han implementado técnicas como la imputación de valores faltantes y el balanceo de clases mediante el método SMOTE (Synthetic Minority Oversampling Technique). Este reporte detalla las transformaciones realizadas, así como los resultados obtenidos tras el preprocesamiento del dataset.

2.Metodologia

Para este análisis y preprocesamiento, se siguieron los siguientes pasos:

1. **Carga y Exploración de los Datos**:
    - Se utilizó el dataset "breast-cancer-wisconsin.csv", que contiene características relacionadas con células cancerígenas.
    - Se realizó un análisis exploratorio para entender la estructura de los datos, identificar valores faltantes y evaluar la distribución de las clases.
2. **Tratamiento de Valores Faltantes**:
    - Se imputaron los valores faltantes utilizando la técnica de imputación con la media, asegurando así que no se pierda información importante.
3. **Análisis de Desbalance de Clases**:
    - Se evaluó la distribución inicial de la variable objetivo class, observándose un desbalance significativo entre las clases.
4. **Aplicación de SMOTE**:
    - Para balancear las clases, se utilizó el algoritmo SMOTE, que genera datos sintéticos para la clase minoritaria, logrando una distribución más equitativa.
5. **Validación y Almacenamiento**:
    - Se verificaron las transformaciones aplicadas al dataset y se guardó el nuevo conjunto de datos balanceado en un archivo CSV para su uso posterior.

2.2 Resultados

 Análisis **Inicial**:

- El dataset original contenía un total de 683 registros con 11 columnas. La columna objetivo class presentaba un desbalance significativo entre las clases.
- Se identificaron valores faltantes en algunas columnas, los cuales fueron tratados mediante imputación.

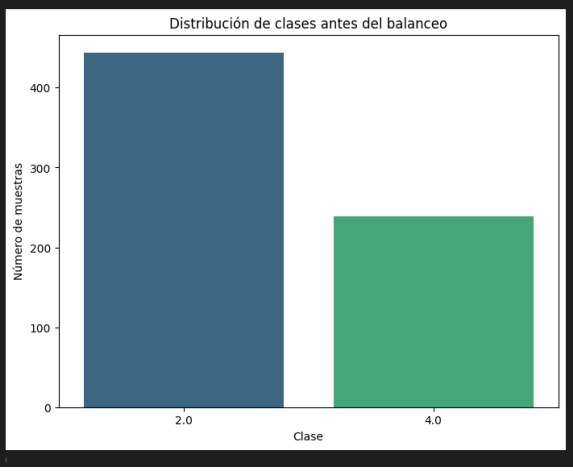
```
Información del dataset antes de la transformación:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 683 entries, 0 to 682
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   code                         683 non-null    int64
 1   clump_thickness              683 non-null    int64
 2   uniformity_of_cell_size      683 non-null    int64
 3   uniformity_of_cell_shape     683 non-null    int64
 4   marginal_adhesion            683 non-null    int64
 5   single_epithelial_cell_size  683 non-null    int64
 6   bare_nuclei                  683 non-null    int64
 7   bland_chromatin              683 non-null    int64
 8   normal_nucleoli              683 non-null    int64
 9   mitoses                      683 non-null    int64
 10  class                        683 non-null    int64
dtypes: int64(11)
memory usage: 58.8 KB

Resumen estadístico inicial:
               code  clump_thickness  uniformity_of_cell_size  \
count  6.830000e+02       683.000000               683.000000
mean   1.076720e+06         4.442167                 3.150805
...
25%          1.000000    1.000000    2.000000
50%          1.000000    1.000000    2.000000
75%          4.000000    1.000000    4.000000
max         10.000000   10.000000    4.000000
```

```
Valores faltantes por columna:
code                           0
clump_thickness                0
uniformity_of_cell_size        0
uniformity_of_cell_shape       0
marginal_adhesion              0
single_epithelial_cell_size    0
bare_nuclei                    0
bland_chromatin                0
normal_nucleoli                0
mitoses                        0
class                          0
dtype: int64

Valores faltantes después de la imputación:
code                           0
clump_thickness                0
uniformity_of_cell_size        0
uniformity_of_cell_shape       0
marginal_adhesion              0
single_epithelial_cell_size    0
bare_nuclei                    0
bland_chromatin                0
normal_nucleoli                0
mitoses                        0
class                          0
dtype: int64
```

Distribución Inicial de Clases:

- Clase 0: 444 muestras.
- Clase 1: 239 muestras.

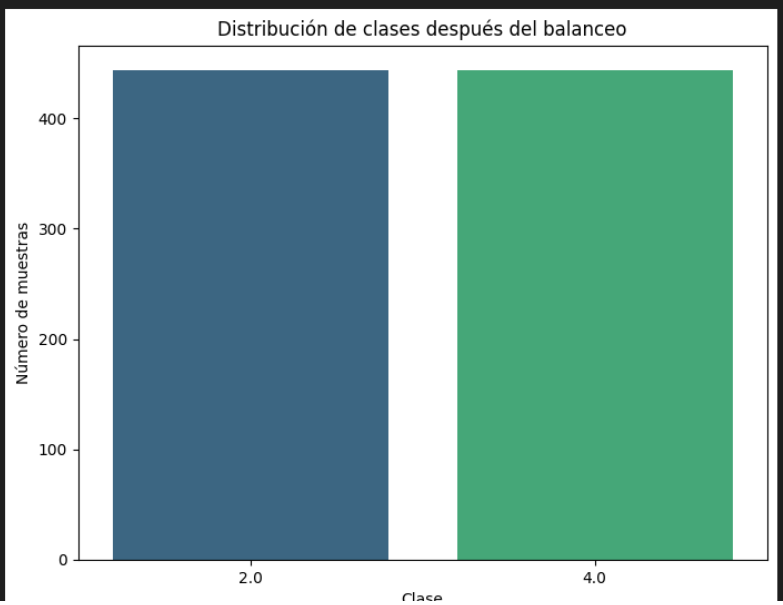
```
Distribución de clases antes del balanceo:
class
2.0    444
4.0    239
Name: count, dtype: int64
C:\Users\vicda\AppData\Local\Temp\ipykernel_2960\468185436.py:9: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=distribution.index, y=distribution.values, palette='viridis')
```



 Distribución **Después del Balanceo**:

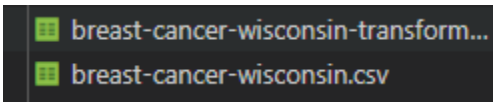
- Tras aplicar SMOTE, se equilibró la cantidad de datos entre las clases:
  - Clase 0: 444 muestras.
  - Clase 1: 444 muestras.

```
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=distribution_after.index, y=distribution_after.values, palette='viridis')

Distribución de clases después del balanceo:
class
2.0    444
4.0    444
Name: count, dtype: int64
```



 Visualización:

- Las gráficas de barras presentadas muestran claramente el cambio en la distribución de clases antes y después del balanceo.
- Además, el dataset transformado no contiene valores faltantes, como se evidenció tras la imputación.



3 Conclusiones

En este trabajo, se logró transformar y mejorar un dataset que presentaba varios problemas comunes en ciencia de datos, como valores faltantes y un desbalance significativo entre las clases. A través del uso de técnicas de preprocesamiento como la imputación de valores y el balanceo de clases con SMOTE, ahora contamos con un conjunto de datos mucho más limpio y equilibrado.

El balanceo de clases fue una de las partes más importantes, ya que al tener cantidades desiguales de ejemplos por clase, los modelos de machine learning tienden a aprender más de la clase mayoritaria y olvidarse de la minoritaria. Con la técnica SMOTE, logramos igualar las clases sin perder información importante, y esto asegura que cualquier modelo que entrene con estos datos tendrá una mayor probabilidad de predecir ambas clases con precisión.

En general, este proyecto me permitió aprender y poner en práctica conceptos clave como la limpieza de datos, el manejo de valores faltantes y técnicas para mejorar datasets desbalanceados. Además, me di cuenta de lo importante que es revisar y preparar bien los datos antes de entrenar un modelo. Si no hacemos un buen trabajo con esta parte, los resultados del modelo pueden ser completamente erróneos o inútiles.

Este dataset transformado ahora puede ser usado para entrenar modelos de machine learning con confianza, sabiendo que el análisis y preprocesamiento previos aseguraron que la información esté bien distribuida y completa. Este tipo de trabajo es esencial en cualquier proyecto de machine learning y me ayudó a entender mejor cómo mejorar los datos para obtener mejores resultados.