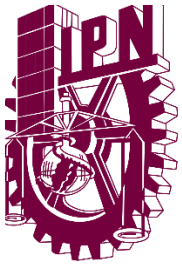


INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

MACHINE LEARNING

5BV1

Complejidad de datos

PROFESOR: ANDRÉS GARCÍA FLORIANO

ALUMNOS:

MIGUEL ANGEL OCAMPO PORCAYO,

GERARDO MARTINEZ AYALA

octubre 2024

INTRODUCCIÓN

El presente informe tiene como objetivo describir el proceso que se llevó a cabo para limpiar y preparar el dataset "Adult" con el fin de realizar un análisis más preciso y eficiente. Este dataset contiene información sobre características demográficas y laborales de personas, y la variable de interés es "income", la cual clasifica a las personas según si sus ingresos anuales son superiores o inferiores a $50,000.

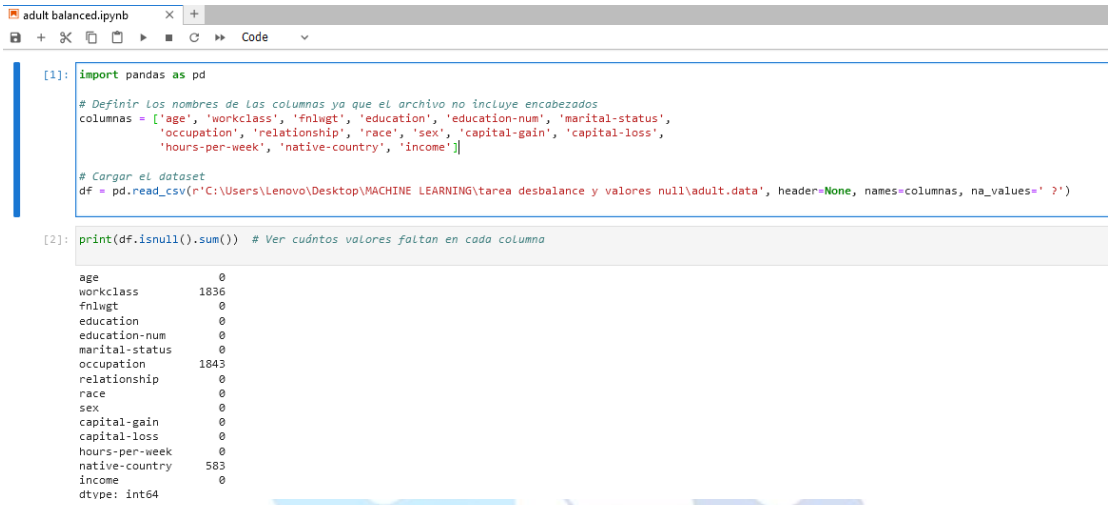
Antes de poder utilizar estos datos en modelos de machine learning, era necesario realizar una serie de pasos para mejorar su calidad. Entre los principales problemas que encontramos se incluyen valores faltantes en varias columnas, así como un notable desbalance entre las clases de la variable objetivo, ya que la mayoría de los registros pertenecen a personas con ingresos menores o iguales a $50,000.

Para abordar estos problemas, se aplicaron técnicas de limpieza de datos y balanceo de clases. Primero, se eliminaron las filas con valores faltantes y, posteriormente, se utilizó la moda de cada columna para completar los datos que faltaban en algunas variables categóricas. Luego, se aplicó la técnica de SMOTE (Synthetic Minority Over-sampling Technique) para balancear la cantidad de ejemplos de cada clase y asegurar que el análisis posterior no estuviera sesgado.

Este informe detalla cada uno de los pasos realizados, desde la identificación de los valores faltantes hasta la aplicación de SMOTE para el balanceo de clases, y muestra cómo estos cambios mejoraron la calidad del dataset y lo dejaron listo para ser utilizado en futuros modelos de machine learning.

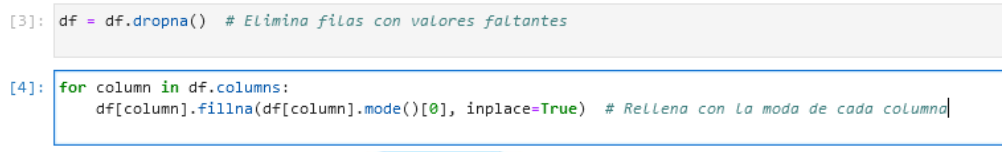
DESARROLLO

Primero, se cargó el dataset y se definieron las columnas debido a que el archivo no incluía encabezados. La siguiente tabla muestra los valores faltantes en cada columna:



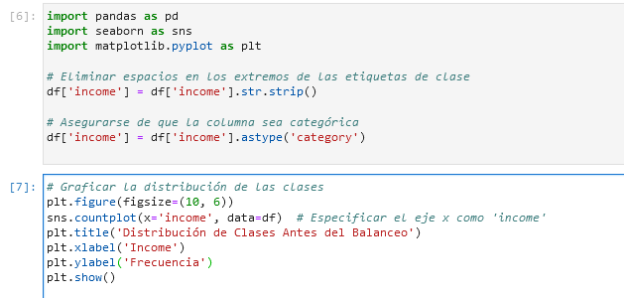
```
[1]: import pandas as pd

     # Definir los nombres de las columnas ya que el archivo no incluye encabezados
     columnas = ['age', 'workclass', 'fnlwgt', 'education', 'education-num', 'marital-status',
                 'occupation', 'relationship', 'race', 'sex', 'capital-gain', 'capital-loss',
                 'hours-per-week', 'native-country', 'income']

     # Cargar el dataset
     df = pd.read_csv(r'C:\Users\Lenovo\Desktop\MACHINE LEARNING\tarea desbalance y valores null\adult.data', header=None, names=columnas, na_values=' ?')

[2]: print(df.isnull().sum())  # Ver cuántos valores faltan en cada columna

age                0
workclass       1836
fnlwgt             0
education          0
education-num      0
marital-status     0
occupation      1843
relationship       0
race               0
sex                0
capital-gain       0
capital-loss       0
hours-per-week     0
native-country   583
income             0
dtype: int64
```

Se decidió eliminar las filas con valores faltantes utilizando el método dropna() de pandas para mantener la integridad del análisis. Además, las columnas restantes se rellenaron con la moda para asegurar que no quedaran vacíos en las variables categóricas.

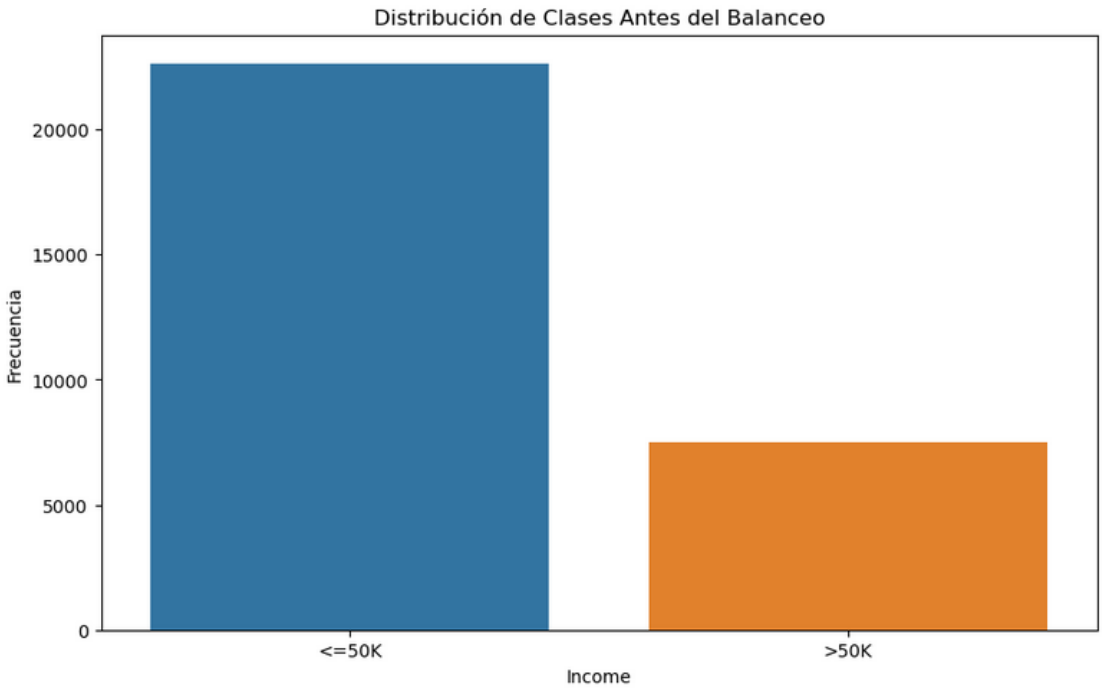


```
[3]: df = df.dropna()  # Elimina filas con valores faltantes

[4]: for column in df.columns:
         df[column].fillna(df[column].mode()[0], inplace=True)  # Rellena con la moda de cada columna
```

Una vez eliminados los valores faltantes, se verificó la distribución de las clases en la variable objetivo "income". El gráfico a continuación muestra un claro desbalance en la cantidad de ejemplos de cada clase.



```
[6]: import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt

     # Eliminar espacios en los extremos de las etiquetas de clase
     df['income'] = df['income'].str.strip()

     # Asegurarse de que la columna sea categórica
     df['income'] = df['income'].astype('category')

[7]: # Graficar la distribución de las clases
     plt.figure(figsize=(10, 6))
     sns.countplot(x='income', data=df)  # Especificar el eje x como 'income'
     plt.title('Distribución de Clases Antes del Balanceo')
     plt.xlabel('Income')
     plt.ylabel('Frecuencia')
     plt.show()
```

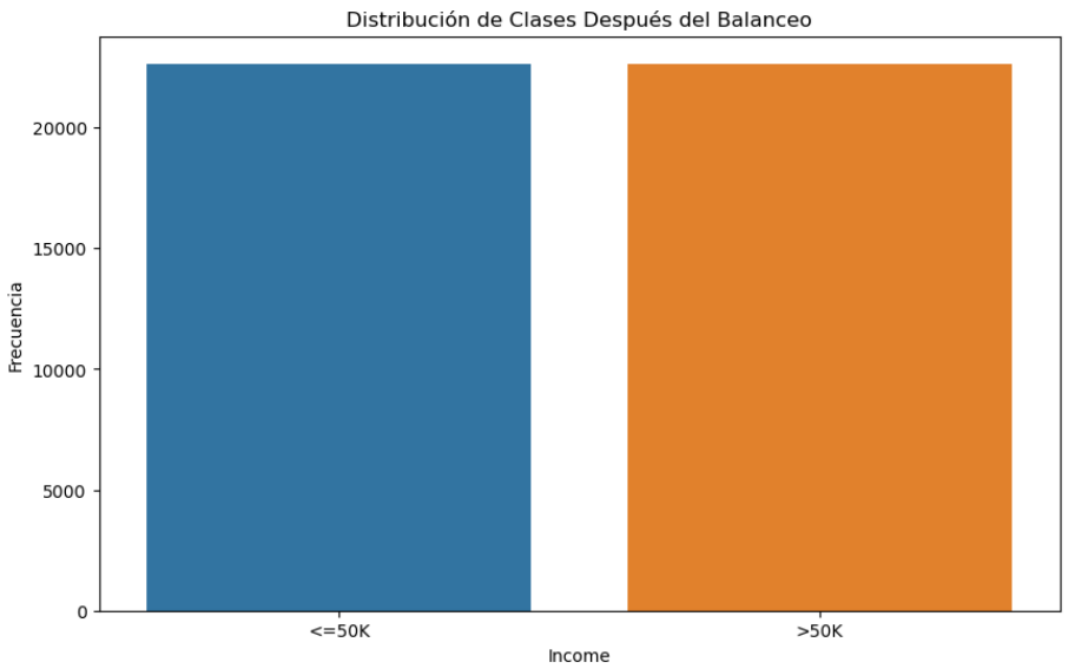


Para balancear las clases, se aplicó el método de sobremuestreo SMOTE (Synthetic Minority Over-sampling Technique), con el fin de aumentar el número de ejemplos en la clase menos representada. El código utilizado para aplicar SMOTE es el siguiente:

[8]:
```
#aumentar numero de instancias de la clase menos representada
from imblearn.over_sampling import SMOTE

X = df.drop('income', axis=1)  # Características
y = df['income']  # Variable objetivo

# Convertir variables categóricas en variables dummy
X = pd.get_dummies(X)

# Aplicar SMOTE para hacer balance de clases
smote = SMOTE(random_state=42)
X_res, y_res = smote.fit_resample(X, y)
```

[10]:
```
# Convertir el dataset resampleado en un DataFrame
df_resampled = pd.concat([pd.DataFrame(X_res), pd.DataFrame(y_res)], axis=1)

# Guardar el dataset transformado
df_resampled.to_csv(r'C:\Users\Lenovo\Desktop\MACHINE LEARNING\tarea desbalance y valores null\adult_transformed.csv', index=False)
```

Después de aplicar SMOTE, se generó el siguiente gráfico para visualizar la distribución de las clases balanceadas:



El dataset transformado fue verificado nuevamente para asegurarse de que no hubiera valores faltantes:

```
[13]: print(df_resampled.isnull().sum())  # Ver cuántos valores faltan en cada columna

age                                0
fnlwgt                             0
education-num                      0
capital-gain                       0
capital-loss                       0
                                  ..
native-country_ Trinadad&Tobago    0
native-country_ United-States      0
native-country_ Vietnam            0
native-country_ Yugoslavia         0
income                             0
Length: 105, dtype: int64

[ ]:
```

CONCLUSION

En este informe se ha detallado el proceso completo de preprocesamiento de datos aplicado al dataset "Adult", centrándonos en la eliminación de valores faltantes y el balanceo de clases. Estos pasos son fundamentales para asegurar que cualquier análisis posterior, como el uso de modelos de machine learning, se realice sobre datos de calidad y sin sesgos.

Primero, identificamos que el dataset contenía una cantidad considerable de valores faltantes, especialmente en columnas como "workclass", "occupation" y "native-country". Para resolver esto, eliminamos las filas que contenían estos valores faltantes y completamos las variables categóricas restantes con la moda, lo que permitió mantener la consistencia en los datos sin perder demasiada información.

Después de la limpieza, se observó un desbalance importante en la variable objetivo "income", donde la mayoría de los ejemplos pertenecían a personas con ingresos menores o iguales a $50,000. Para corregir este desbalance, utilizamos la técnica de SMOTE, que genera nuevos ejemplos sintéticos para la clase menos representada. Este proceso nos permitió equilibrar la cantidad de ejemplos en cada clase, lo que es clave para que los modelos no se inclinen injustamente hacia la clase más común.

Gracias a este preprocesamiento, el dataset ahora está listo para ser utilizado en modelos de machine learning sin los problemas iniciales de desbalance y valores faltantes. Esto no solo mejorará la precisión de los modelos, sino que también permitirá que las predicciones sean más justas y representativas de ambas clases. En resumen, este trabajo de preprocesamiento ha sido crucial para asegurar que los datos sean fiables y estén en óptimas condiciones para su análisis y modelado.

REFERENCIAS

Becker, B. & Kohavi, R. (1996). Adult [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.

Tratamiento de clases desbalanceadas. (2020, January 20). Machinelearningparatodos.com. https://machinelearningparatodos.com/tratamiento-de-clases-desbalanceadas/