Project Report: Data Preprocessing and Balancing Using SMOTE

# 1. Introduction

This project involves preprocessing three datasets—loan approval data, credit card data, and kidney disease data—and applying the Synthetic Minority Over-sampling Technique (SMOTE) to balance the datasets. The primary objectives of the project are to:  
- Preprocess the datasets by handling missing values, encoding categorical features, and splitting the data into training and testing sets.  
- Apply SMOTE to handle class imbalance, ensuring that minority classes are well-represented in the training data.

# 2. Dataset Overview

Three datasets were used in this project:  
1. \*\*Loan Approval Data\*\*: Contains information about loan applicants and their approval status.  
2. \*\*Credit Card Data\*\*: Includes details on credit card users, their demographics, and payment histories.  
3. \*\*Kidney Disease Data\*\*: Provides medical information used to classify patients as having kidney disease or not.

# 3. Preprocessing Steps

The following steps were applied to each dataset:  
1. \*\*Loading and Cleaning Data\*\*:  
 - Datasets were loaded into pandas DataFrames.  
 - Column names were stripped of leading and trailing whitespace.  
  
2. \*\*Handling Missing Values\*\*:  
 - Rows with missing values were removed to ensure data integrity.  
  
3. \*\*Encoding Categorical Features\*\*:  
 - Categorical variables with fewer than 10 unique values were one-hot encoded.  
 - Variables with more than 10 unique values were label encoded to reduce dimensionality.  
  
4. \*\*Splitting Data\*\*:  
 - The datasets were split into training and testing sets using a stratified split to preserve the distribution of the target variable.

# 4. Addressing Class Imbalance with SMOTE

To address the class imbalance, SMOTE was applied to the training datasets. However, challenges arose when some classes had too few samples to support the generation of synthetic data. To handle this:  
  
1. \*\*Class Filtering\*\*:  
 - Classes with fewer than two instances were removed before splitting.  
  
2. \*\*SMOTE Application\*\*:  
 - SMOTE was applied with `k\_neighbors=1` to accommodate datasets with fewer samples.  
  
3. \*\*Error Handling\*\*:  
 - A custom function was created to apply SMOTE and handle cases where SMOTE could not be applied due to the class size constraints.

# 5. Results and File Outputs

After preprocessing and applying SMOTE, the following balanced datasets were generated:  
1. \*\*Loan Approval Balanced Dataset\*\*  
2. \*\*Credit Card Balanced Dataset\*\*  
3. \*\*Kidney Disease Balanced Dataset\*\*  
  
These datasets were saved as CSV files for further analysis or model training.

# 6. Conclusion

This project successfully demonstrates the application of preprocessing techniques and SMOTE for handling class imbalance in various datasets. The methods implemented ensure that the datasets are clean, balanced, and ready for machine learning model training. Error handling mechanisms were crucial in managing edge cases where SMOTE could not be applied directly.