Qamar Uddin

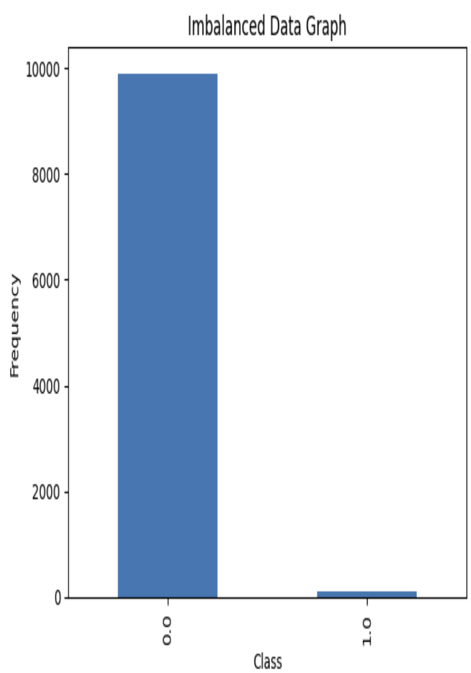
Ex6 Report: Data Augmentation for Imbalanced Dataset

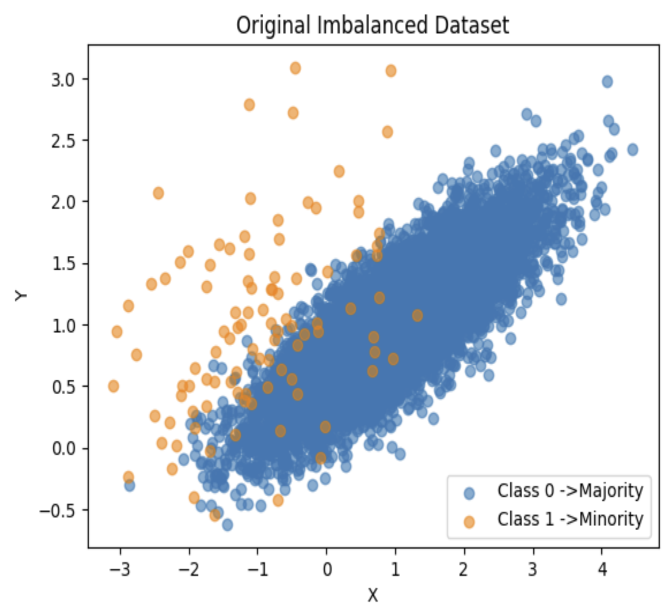
Objective

The aim of this exercise is to handle class imbalance in a binary classification dataset using numerical data augmentation techniques, specifically **SMOTE** and **Borderline-SMOTE**.

Dataset Description

The imbalanced dataset (`Imbalanced_data.csv`) was uploaded to Google Colab. The first two columns represent coordinate features (**X** and **Y**), and the last column represents the class label (**Class**) with values 0 and 1. The dataset is highly imbalanced, with approximately **99% samples belonging to class 0** and **1% belonging to class 1** shown in the graph below taken before any augmentation.





Methodology

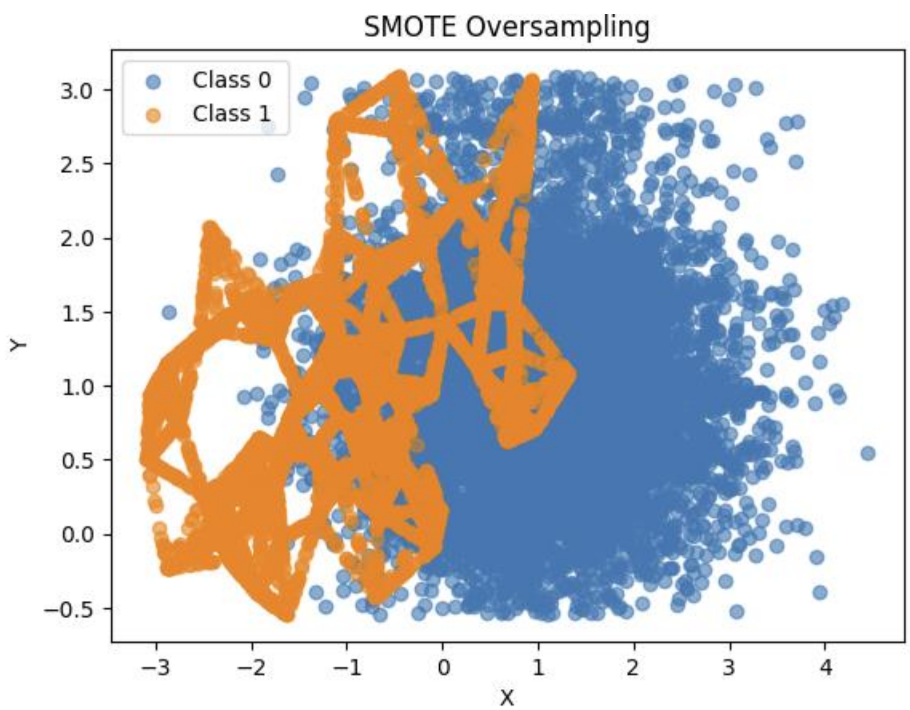
After loading the dataset, appropriate column headers were assigned, and the data was analyzed and visualized to confirm the imbalance. The dataset was then oversampled using **SMOTE** and **Borderline-SMOTE** methods from the *imbalanced-learn* library.

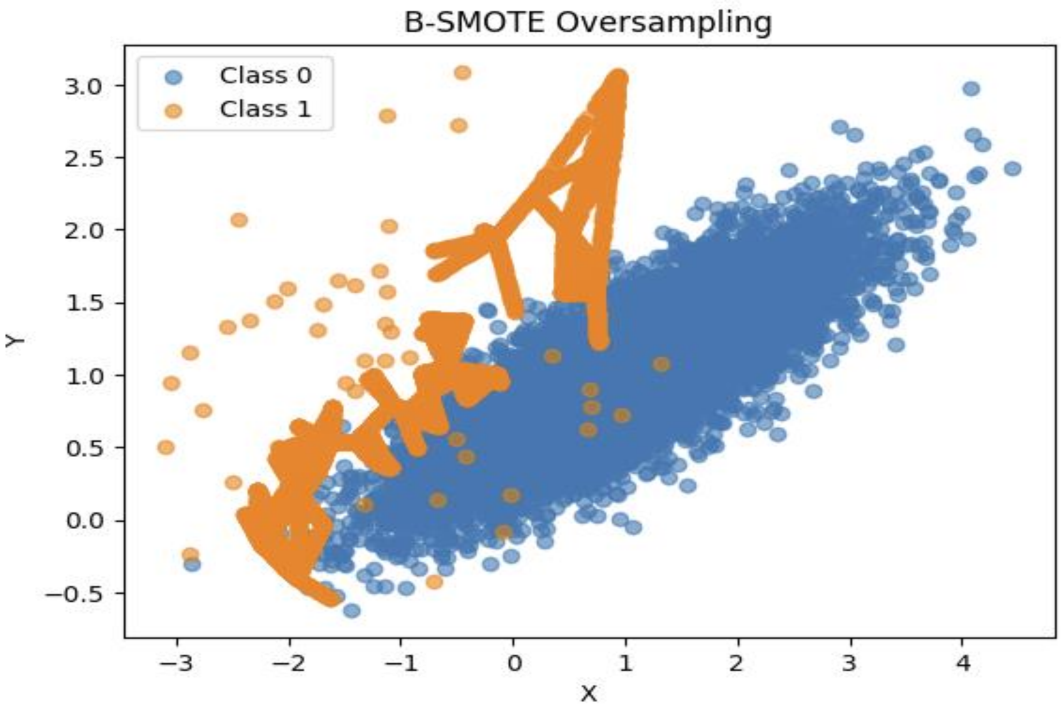
Results

SMOTE generated synthetic minority class samples uniformly across the feature space, resulting in a balanced dataset. Borderline-SMOTE generated synthetic samples mainly near the class boundary, focusing on difficult-to-classify regions. Scatter plots were used to visually compare the original imbalanced data with the oversampled datasets.

Conclusion

Both SMOTE and Borderline-SMOTE effectively addressed the class imbalance problem. Borderline-SMOTE provides a more focused oversampling strategy near decision boundaries and can be more effective for improving classification performance. Both SMOTE and B/SMOTE comparison graphs are given below.




B-SMOTE Oversampling
Class 0
Class 1
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
Y
−3
−2
−1
0
1
2
3
4
X