Report

Sampling Techniques Used

Oversampling

- **Random Oversampling**: Randomly duplicated minority classs to achieve balance
- **SMOTE**: Created synthetic samples of the minority class using nearest neighbours

Undersampling

- **Random Undersampling**: Randomly removed majority class samples
- **Cluster Centroids**: Reduced majority class by replacing clusters with their centroids

Class Distribution Results

Original Dataset

- Initial distribution: 357 (class 1) vs 212 (class 0)
- Train split: 287 vs 168
- Test split: 70 vs 44

After Resampling

- **Oversampled**: Balanced : 357 vs 357
  - Train split: 289 vs 282
  - Test split: 75 vs 68
- **Undersampled**: Reduced to 212 vs 212
  - Train split: 172 vs 167
  - Test split: 45 vs 40

Performance Metrics

Original Dataset

- Accuracy: `93.86%`
- Precision: [0.930, 0.944]
- Recall: [0.909, 0.957]

Oversampled Dataset

- Accuracy: `88.81%`
- Precision: [0.894, 0.883]
- Recall: [0.868, 0.907]

Undersampled Dataset

- Accuracy: `92.94%`
- Precision: [0.976, 0.886]
- Recall: [0.889, 0.975]

In summary

The original imbalanced dataset for some reason performed best with 93.86% accuracy. The undersampled dataset showed comparable performance at 92.94%, while oversampling resulted in slightly lower accuracy at 88.81%. This shows that for this particular dataset, the original class distribution might be optimal for the classifier's performance.