Report

Sampling Techniques Used

Oversampling Methods

1. **Random Oversampling**
   - Randomly duplicated minority class samples
   - Balanced classes from 212:357 to 357:357
2. **SMOTE (Synthetic Minority Over-sampling Technique)**
   - Created synthetic samples for minority class
   - Achieved same balance as random oversampling (357:357)

Undersampling Methods

1. **Random Undersampling**
   - Randomly removed majority class samples
   - Reduced classes to 212:212
2. **Cluster Centroids**
   - Used clustering to reduce majority class
   - Achieved same balance as random undersampling (212:212)

Class Distribution Results

Pre-train-test

```
Original class distribution: Counter({1: 357, 0: 212})
Oversampled class distribution using RandomOversampling: Counter({0: 357, 1: 357})
Oversampled class distribution using SMOTE: Counter({0: 357, 1: 357})
Undersampled class distribution using RandomUndersampler: Counter({0: 212, 1: 212})
Undersampled class distribution using ClusterCentroids: Counter({0: 212, 1: 212})
```

Train-Test Splits

- **Original**: Train (288:167), Test (69:45)
- **Random Oversampling**: Train (286:285), Test (71:72)
- **SMOTE**: Train (286:285), Test (71:72)
- **Random Undersampling**: Train (172:167), Test (45:40)
- **Cluster Centroids**: Train (172:167), Test (45:40)

Performance Metrics

Accuracy Comparison

1. Random Oversampling: 95.80%
2. SMOTE: 95.80%
3. Random Undersampling: 95.29%
4. Original: 94.74%
5. Cluster Centroids: 91.76%

Precision & Recall

- **Original**:
  - Precision: [0.915, 0.970]
  - Recall: [0.956, 0.942]
- **Random Oversampling**:
  - Precision: [0.934, 0.985]
  - Recall: [0.986, 0.930]
- **SMOTE**:
  - Precision: [0.946, 0.971]
  - Recall: [0.972, 0.944]

In Summary

- Both oversampling methods (Random and SMOTE) achieved the best performance with 95.80% accuracy. Random undersampling followed closely at 95.29%.
- While the original imbalanced dataset performed well (94.74%), the balanced datasets generally showed better results.
- Cluster Centroids undersampling showed the lowest performance at 91.76%, maybe that this method might have lost important information during the undersampling process.