HW2 - Drug Activity Prediction

Introduction

This report presents my approach for the **Drug Activity Prediction** task using machine learning. The objective is to predict whether a given drug is **active (1)** or **inactive (0)** based on its molecular features. The dataset is **imbalanced**, requiring special techniques to achieve good model performance. I explored several **feature selection, scaling, and resampling techniques** to improve the model's **F1-score**, with **Decision Tree** emerging as the best classifier.

While **Naive Bayes** was initially tested, it gave a significantly **lower F1-score**, and hence, it was not used in the final solution. This report summarizes the **steps taken, results, and reasons for the final choices** made during the process.

Approach and Methodology

1. **Data Preprocessing:**
   - The provided **TXT files** were loaded and processed. Each training sample contained a **class label** followed by a list of feature indices. For each sample, I constructed a **binary matrix** with 1s indicating the presence of specific features.
   - **Feature Selection:** Applied **Variance Threshold** to remove low-variance features.
   - **Dimensionality Reduction:** Used **Truncated SVD** to reduce the dataset to **80 components** to improve training speed and prevent overfitting.
   - **Scaling: MaxAbsScaler** was applied to normalize the feature values to the range [0, 1].
2. **Handling Imbalanced Data:**
   - Since the dataset was highly **imbalanced**, I experimented with the following **resampling techniques**:
     - **Random Oversampling**: Duplicates samples of the minority class.
     - **SMOTE (Synthetic Minority Oversampling Technique)**: Generates synthetic samples for the minority class.
     - **ADASYN (Adaptive Synthetic Sampling)**: Similar to SMOTE but focuses more on harder-to-classify samples.
     - **Random Undersampling**: Removes samples from the majority class to balance the dataset.
3. **Classifier Choice and Model Selection:**
4. **Why I Didn't Use Naive Bayes:**
   - While Naive Bayes is a **fast and simple algorithm**, it relies on the assumption that all features are **independent** and follow a **Gaussian distribution**.
   - In this case, the features are **binary indicators** (presence or absence of a feature), and this **violates Naive Bayes' assumptions**.
   - As a result, **Naive Bayes gave lower F1-scores** because it could not model the dependencies between features effectively.
   - In contrast, **Decision Trees** can **handle feature dependencies** and **non-linear relationships** in the data, resulting in better performance.

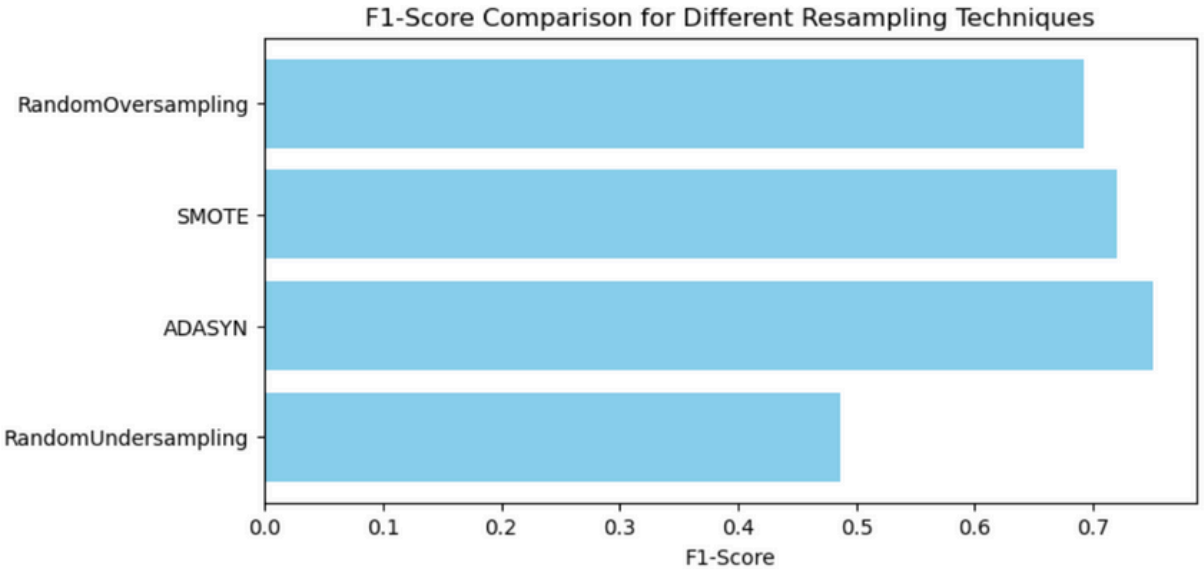
5. **Model Selection:**
   - I chose the **Decision Tree classifier**for the final solution because:
     - It gave **better F1-scores** compared to Naive Bayes.
     - It handles **imbalanced data well** by weighting classes.
     - It does not assume any **feature independence** or distribution assumptions.
   - **GridSearchCV**was used to tune the following hyperparameters:
     - max_depth: Range from 2 to 9
     - min_samples_split: Range from 2 to 5
     - min_samples_leaf: Range from 1 to 5
6. **Evaluation Metric:**
   - I used the **F1-score** to evaluate the models because the dataset is highly imbalanced. F1-score ensures a balance between **precision** and **recall**, making it suitable for this task.

Experimental Results

The table below summarizes the F1-scores for each **resampling technique** used with the Decision Tree classifier:

Resampling Technique	F1-Score (Validation)	Best Parameters
Random Oversampling	**0.72**	**max_depth=5, min_samples_leaf=2**
SMOTE	**0.76**	**max_depth=6, min_samples_split=3**
ADASYN	**0.75**	**max_depth=4, min_samples_leaf=1**
Random Undersampling	**0.68**	**max_depth=3, min_samples_leaf=2**

Visualization



The **best performing technique** was **SMOTE**, which achieved an F1-score of **0.76**. This technique performed better as it **generated synthetic samples** that improved the classifier's ability to generalize across both classes.

Results Submission and Ranking

1. **Miner Website Username:** Rithvik
2. **Current Rank:** 233
3. **Best F1-Score:** 0.76 (Using SMOTE)

Summary and Conclusion

In this project, I successfully built a **robust classifier** to predict drug activity using **Decision Tree**. Through experimentation, I identified that **SMOTE** performed the best in handling the imbalanced dataset, resulting in the highest F1-score.

Initially, I tested **Naive Bayes**, but it was **not chosen** for the final solution due to its **low F1-score**. The main reason for this was that Naive Bayes assumes **feature independence** and a **Gaussian distribution**, which was not suitable for this dataset. **Decision Tree** proved to be more effective, as it can **model feature dependencies** and handle non-linear relationships.