**CS 584- Data Mining and Applications**
**Kaustubh Prashant Karanjkar**
Identifier: **Kaustubh**
**#G01314326**

# HW2: Drug Prediction

## On Miner:
## Rank: 102
## Accuracy: 73%

Aim:

- Use/implement a feature selection/reduction technique.
- Experiment with various classification models.
- Think about dealing with imbalanced data.
- Use F1 Scoring Metric

Approach:

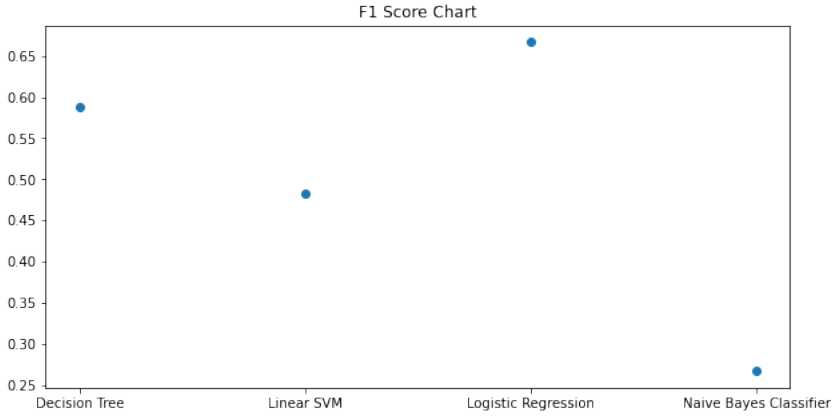
The training data given was an imbalance data, which means that the data which belongs to class 0 (inactive) was 722 and the data which belongs to class 1 (active) was 78.
The first step was to convert the training data into a sparse matrix, in the training data only the columns with class 1 were given.
The training data had 100,000 features, so the next task was to eliminate the unwanted features. So to eliminate the unwanted features, I have tried and used Chi-Square, SVD and PCA feature selection techniques. These feature selection techniques help reduce the dimensions of the data and help us select the most relevant features to train our model. I have imported these from the sklearn library.

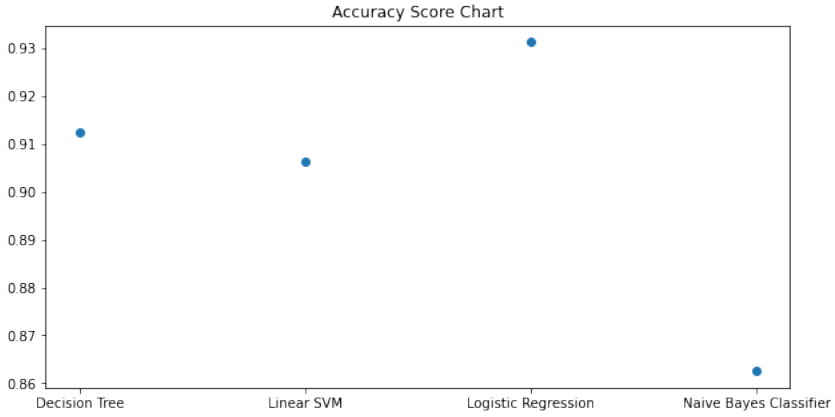
I used chi-square feature selection to reduce the dimensionality of the data and applied it to the 4 classifiers, namely,

1. Decision Tree Classifier
2. Linear SVM Classifier
3. Logistic Regression Classifier

4. Naïve Bayes Classifier

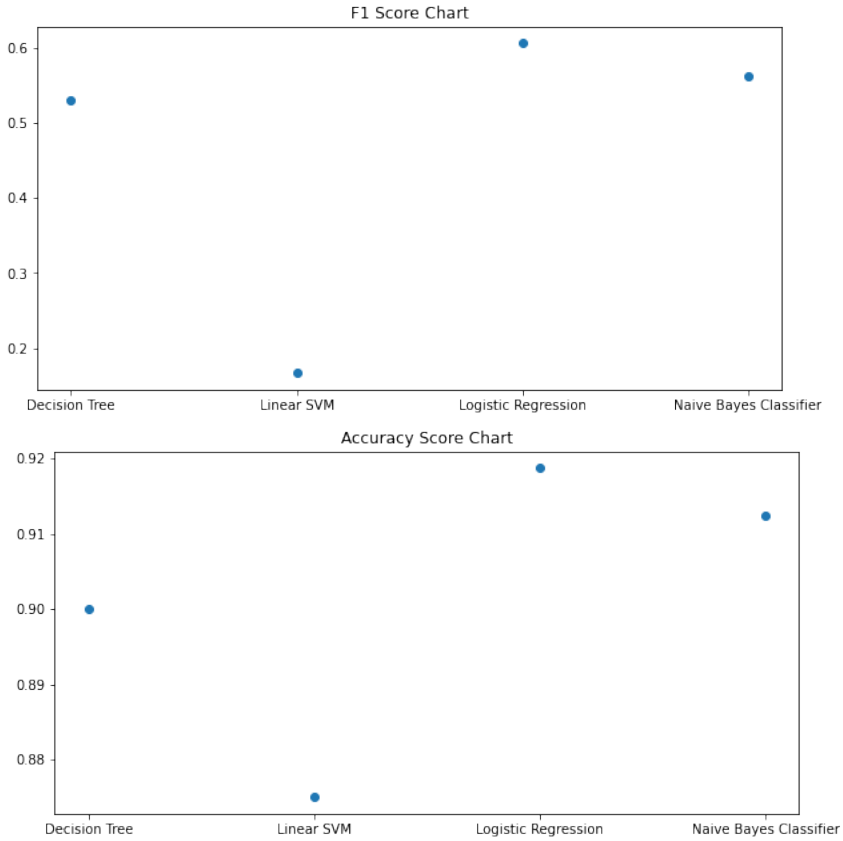
I used the Train Test Split function from sklearn model selection to split the training data in the ratio of 80:20, training data being 80% and testing data being 20%. I had done this to train my model and check the accuracy before using the model directly on the actual test data. I have used f1 Score and Accuracy metrics to check the accuracy of the model.
The below chart shows the result of the f1 Score and Accuracy metrics, Logistic Regression classifier performed best with an accuracy of 66%. I have used the class weight of {0:1, 1:9} to balance the imbalanced nature of the training data.



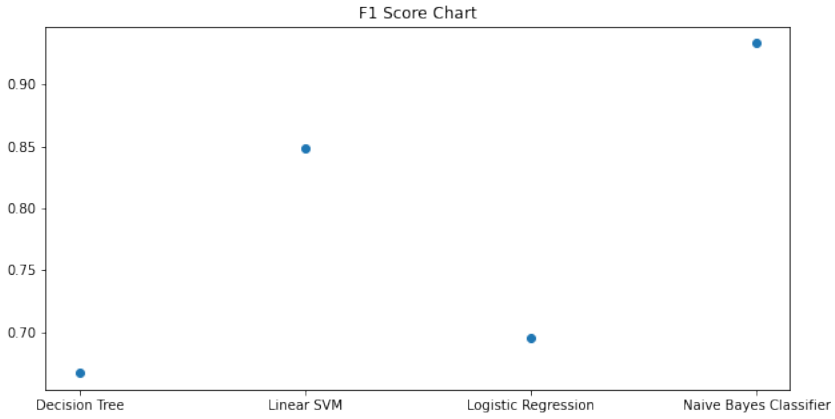

Later, I tried SVD feature selection method on these 4 classifiers, I have kept n_component=800 and n_iter=30. This time Naïve Bayes classifier also performed

well along with the Logistic regression classifier with the f1 score of 56.25% and accuracy score of 91.25% . Below chart shows the result.

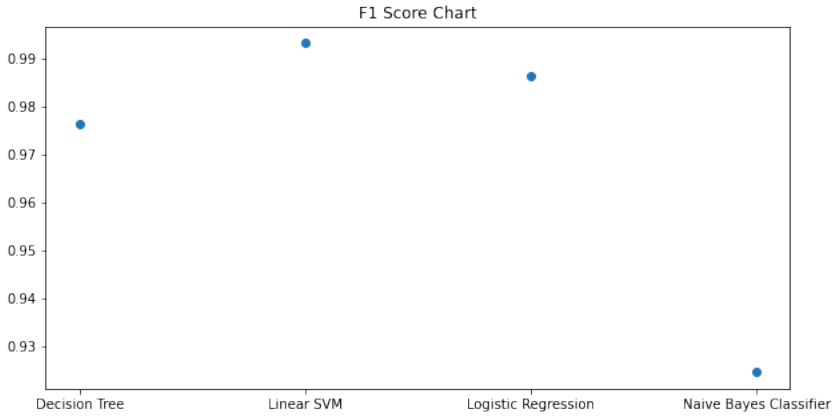



Later, to improve my accuracy score I had used sampling techniques to balance the imbalanced training data. I have tried both the Oversampling and Undersampling techniques. I have used Random OverSampler, Random UnderSampler and SMOTE from imblearn.

Firstly, I tried undersampling the training data and applied chi square feature selection technique and found that Naïve Bayes Classifier performed well among all the classifiers with the accuracy of 93%




Then, I applied Oversampling and used SVD feature selection method on all the classifiers and this time Linear SVM performed well with an accuracy of 99.3%, this time also Naïve Bayes classifier performed well along with other classifiers.




I have observed that after performing sampling on the training data to make balanced data, the f1 score accuracy has been increased on all the classifiers.

Later, uploading on the miner, the **Naïve Bayes classifier** received the highest accuracy of 73% among all the classifiers and the **Decision Tree Classifier** was the second with an accuracy of 63%.