Final Project Documentation
Ehsan Mehryaar

In this Notebook we want to evaluate three classification techniques of Random Forest, Support Vector Machine, and Naive Bayes on prediction of diabetes in python. For the purpose of Modeling, Sklearn library in python has been used.

The Zip file includes this documentation file, one Main.py which is the the raw python code and one main.ipynb which is the python notebook and also used dataset.

Data was acquired from the hospital Frankfurt, Germany which is available in Kaggle website (Link). Data consists of of 8 features including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The output is binary, 0 for negative result and 1 for diagnosis of diabetes.

As it was observed the data was imbalanced in favor of negative results. The results was then balanced by over sampling of positive results.

This process Consist 4 steps

- Loading data
- Cleaning data
- Modeling
- Evaluating the models

In this section performance metrics for each step of cross-validation and also test date for each model is acquired.

The performance metrics given in this function includes:

- Accuracy
- Precision
- f1
- ROC and AUC
- Jaccard
- True Positive
- False Positive
- True Negative
- False Negative
- True Positive Rate
- True Negative Rate
- False Positive Rate
- False Negative Rate
- Error Rate
- Balanced Accuracy
- True Skill Statistics

- Heidke Skill Score

Models was developed for both imbalanced and balanced data. It was observed that balancing increased the performance of the models in the cross-validation phase however it was not showing substantial changes. In general SVM and Random forest showed better performance than Naïve bayes algorithm.