Machine Learning Classification on PIMA Indian Diabetes Dataset

CS677 Final Project by Vindhya Balasubramanyam

Instructor: Dr. Eugene Pinsky

Facilitator: Nadeem Khan

Aug 10th, 2021

Contents

[1. Objectives 3](#_Toc79535120)

[2. Data Preparation 3](#_Toc79535121)

[3. Run Classifiers and compare performance 11](#_Toc79535122)

[4. Shapely Analysis of feature significance 15](#_Toc79535123)

[5. Conclusion 17](#_Toc79535124)

# Objectives

The PIMA dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and is also available on Kaggle. The data is from women patients of Pima Indian heritage that are 21 years of age or older. The objective of the project is to

* predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.
* Determine feature contribution towards accuracy

# Data Preparation

Dataset has 8 predictor features and 1 target variable.

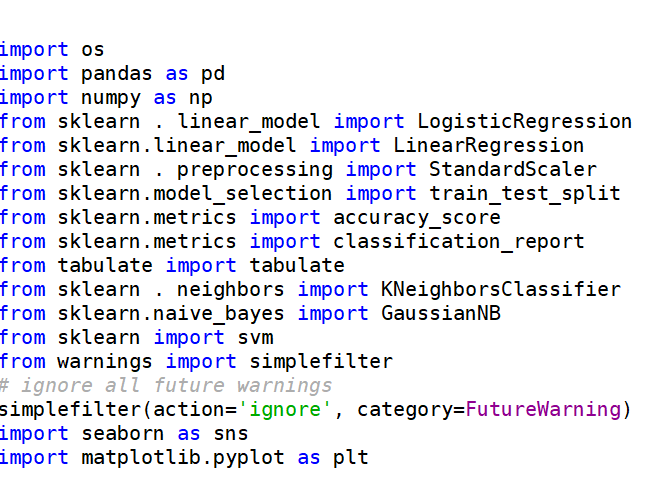
#### Target Variable

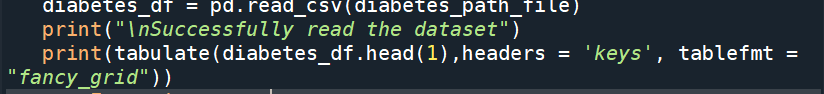
* Outcome: 0 (no diabetes) or 1 (has diabetes)

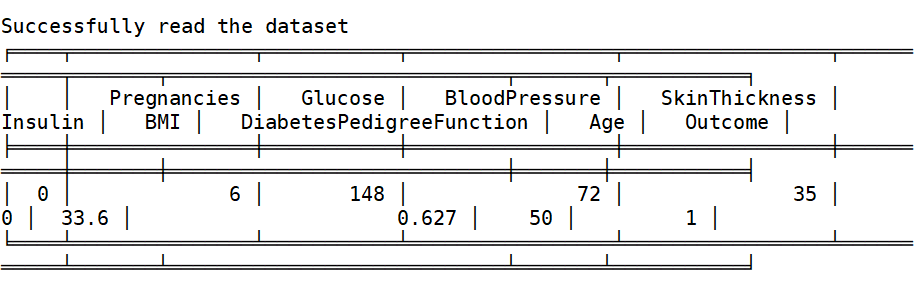
#### Features

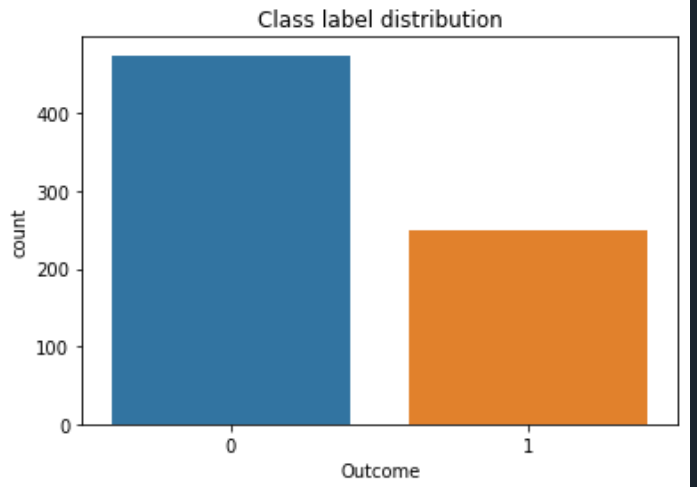
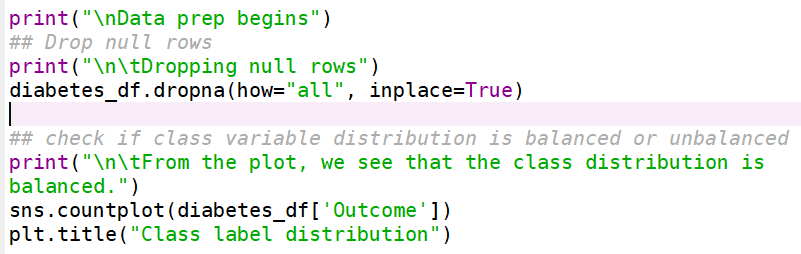
1. Pregnancies - Number of times pregnant
2. Glucose - Plasma glucose concentration
3. BloodPressure - Diastolic blood pressure (mm Hg)
4. SkinThickness - Triceps skin fold thickness (mm)
5. Insulin - 2-Hour serum insulin (mu U/ml)
6. BMI - Body mass index
7. DiabetesPedigreeFunction
8. Age

Libraries used:

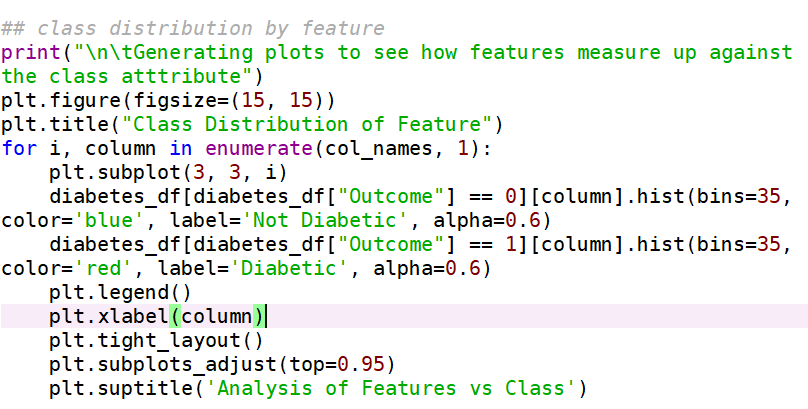


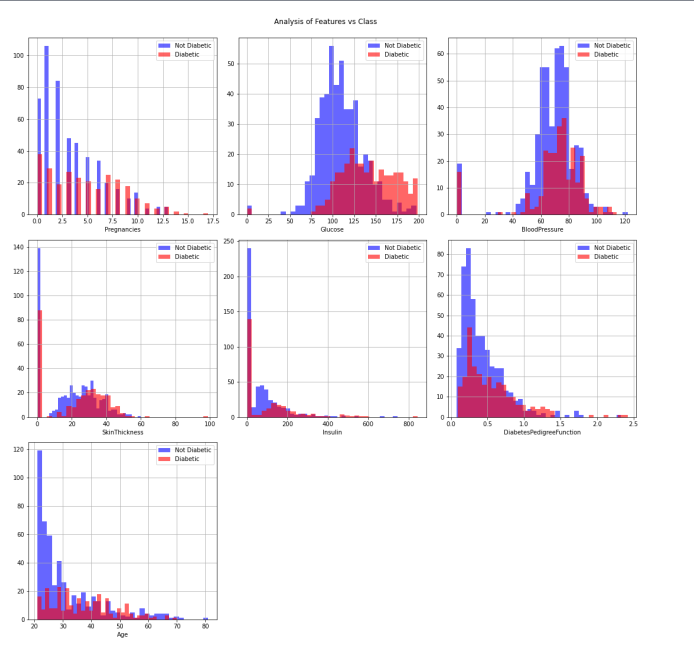




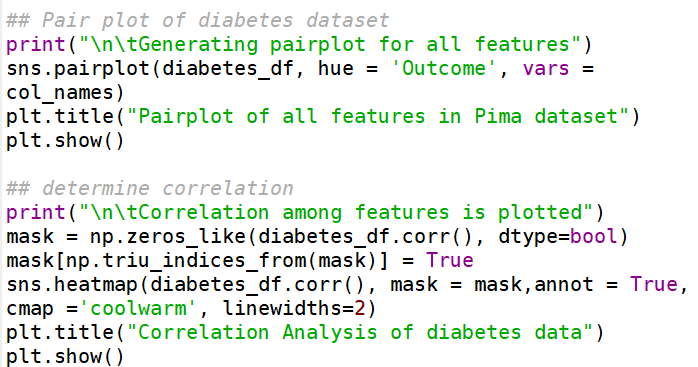


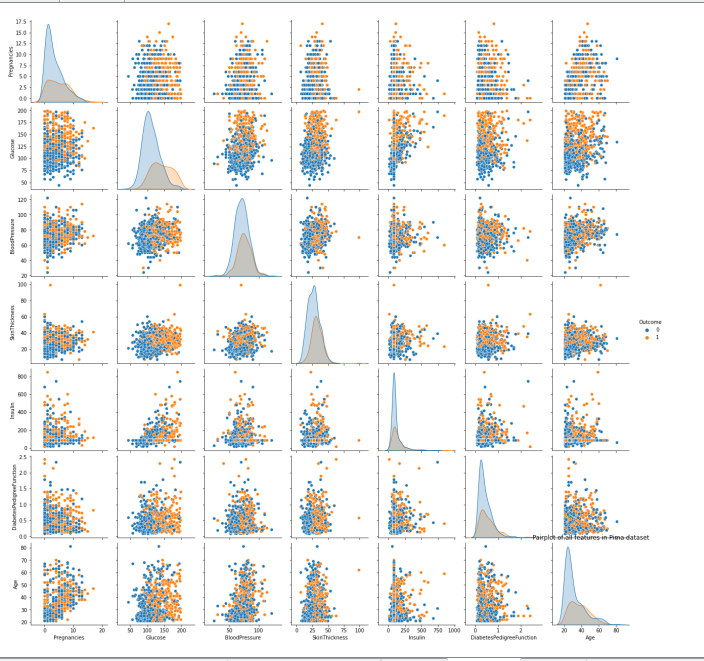
From the plot, we see that the class distribution is balanced.

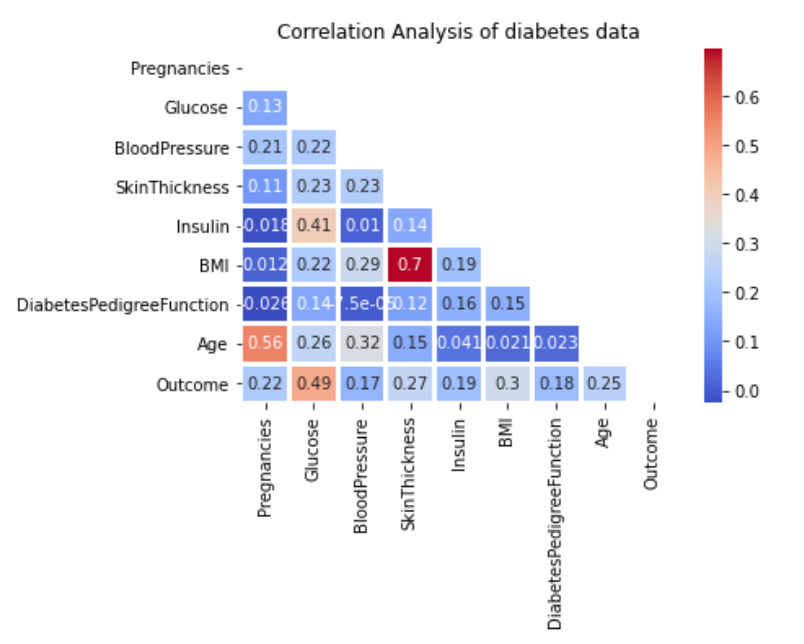




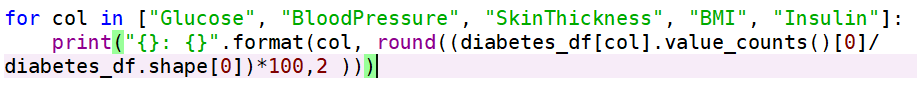
From the plot, we see that the class labels are well distributed across all features.

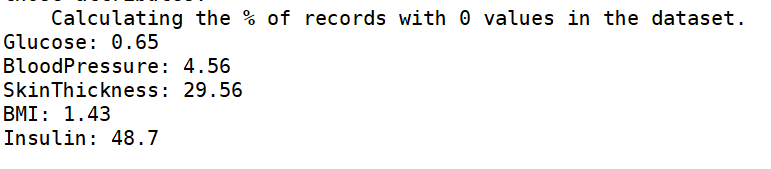


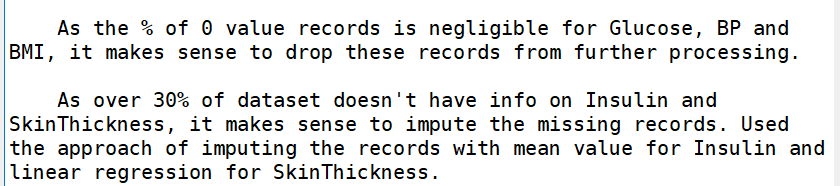


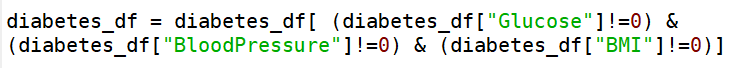


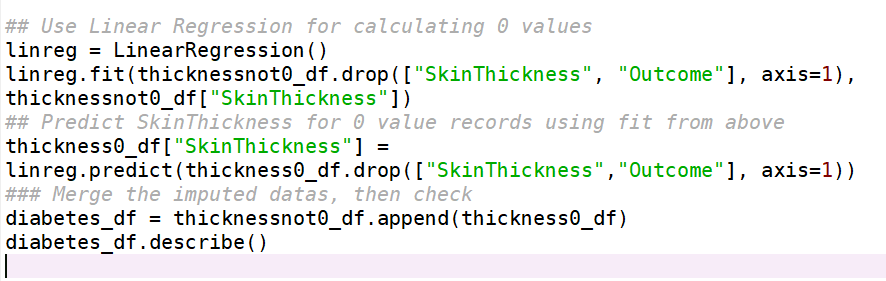
From the above plots, you see that there is not a strong correlation of features against the target variable. The feature with the highest correlation to Outcome variable is Glucose.

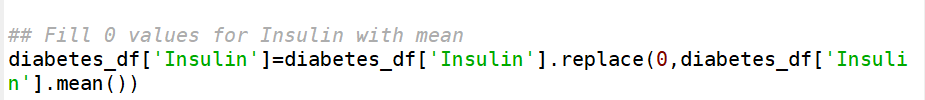






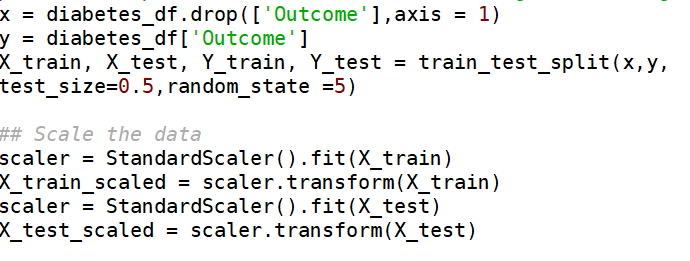




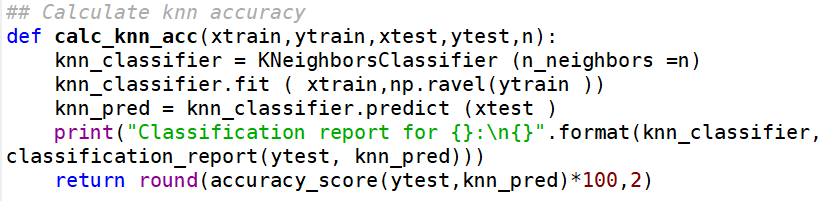


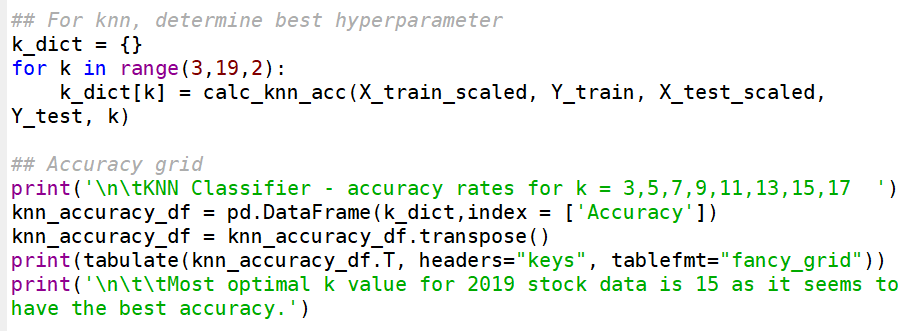
# Run Classifiers and compare performance

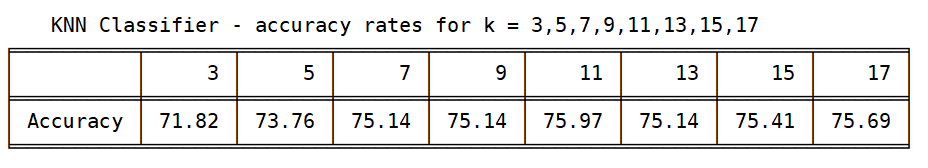
Split the dataset in to test and train sets and scale the data.



Run Knn classifier and determine which most optimal n value

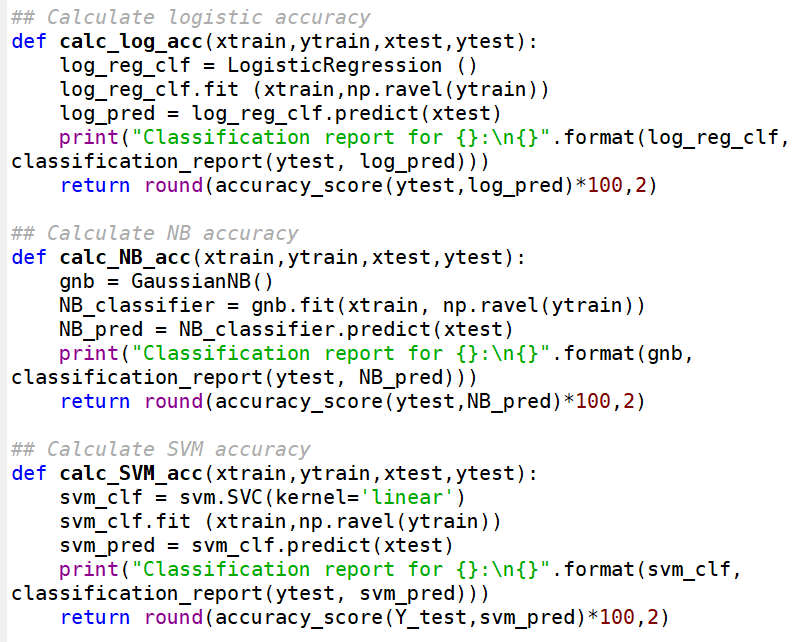


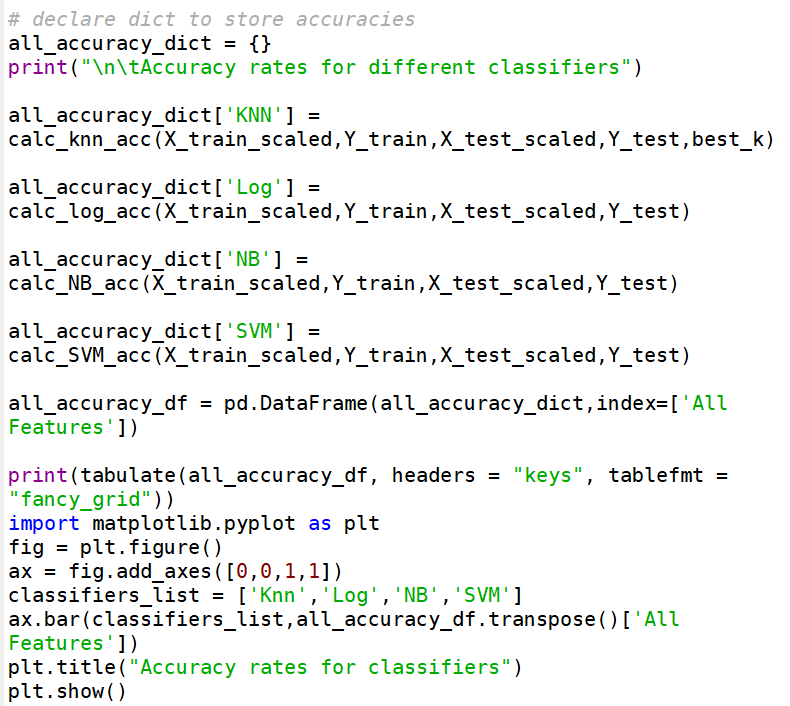


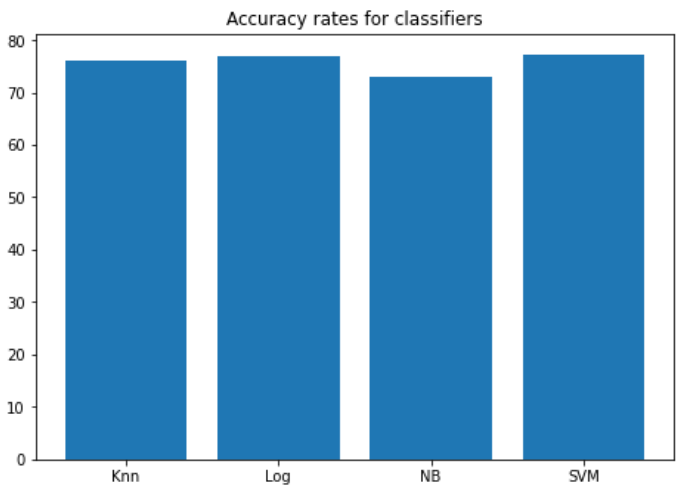


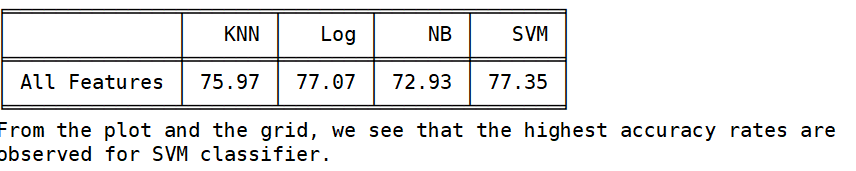
Most optimal k value for 2019 stock data is 15 as it seems to have the best accuracy.

Run Logistic, Naïve Bayes and SVM classifiers on test and train datasets and compile accuracy, precision and recall metrics.







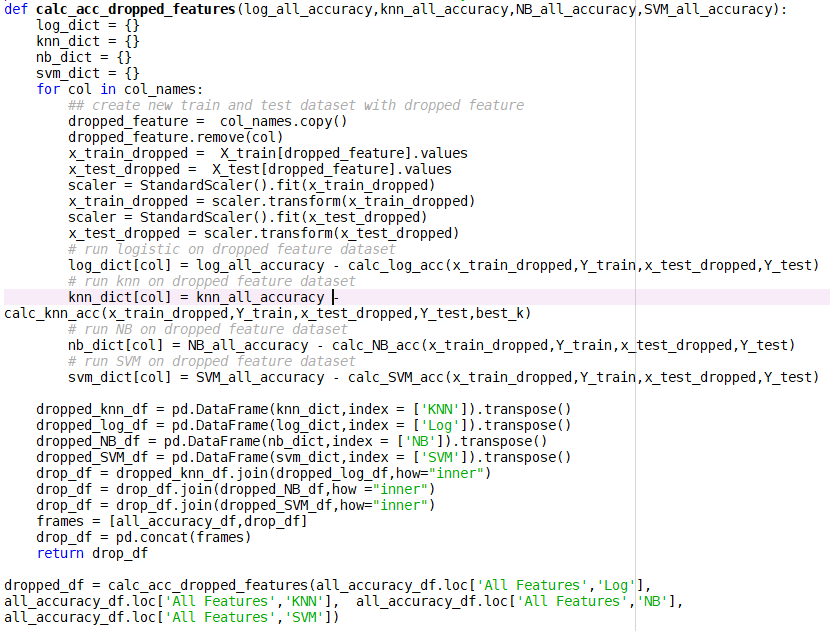


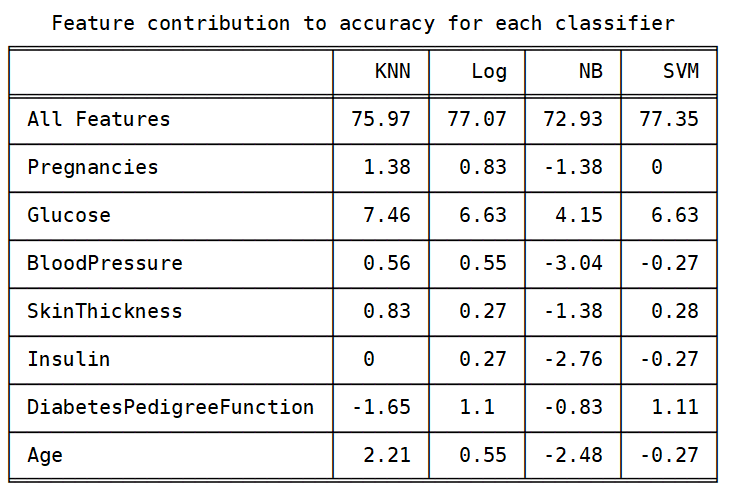
|  |  |  |  |
| --- | --- | --- | --- |
| Classifier | Accuracy | Precision | Recall |
| Knn | 76 | 76 | 48 |
| Logistic | 77 | 72 | 58 |
| NB | 73 | 64 | 54 |
| SVM | 77 | 73 | 58 |

* All 4 classifiers accuracy rate was above 73%
  + SVM had the highest accuracy at 77.35% followed by Logistic at 77.07%
  + The pre-process step helped improve accuracy of each classifier by 3-5%
* For Precision, Knn had the highest score and NB has the lowest.
* For Recall, SVM has the highest score and Knn has the lowest.

# Shapely Analysis of feature significance

For this part of the project, all the above classifiers are run by dropping one feature at a time. Here’s the code to do that.





* From the above grid, one can conclude that Glucose is the most significant contributor towards the accuracy rates of a classifier. We also see that the feature contribution varies for each classifier and there’s no common pattern. Accuracy of all classifiers would improve by excluding certain features and NB would have the most significant impact.

# Conclusion

Through this analysis, we see that the data exploration and cleanup is an important step in any machine learning project. From running a couple different iterations, I saw the accuracy rates of all classifers go up by 2-3% due to the cleanup steps. As the accuracy rates for all classifiers was in a similar range, I would recommend using SVM as the recall measure was the highest. With more data and some more fine tuning, the performance of all classifiers can be further improved.