**PROJECT ON CLINVAR CONFLICTING**

**EXECUTIVE SUMMARY**

This project is aimed at data cleaning and mining of a given dataset to extract useful insight and create a machine learning model. The following questions will be answered in the course of this work:

* Work on this dataset, bringing out all the useful insight and information
* Build a model for this data set, the accuracy score should not be less than 96%

**INTRODUCTION**

ClinVar is a public resource containing annotations about human genetic variants. These variants are classified on a spectrum between benign, likely benign, uncertain significance, likely pathogenic, and pathogenic. Variants that have conflicting classifications (defined above) can cause confusion when clinicians or researchers try to interpret whether the variant has an impact on the disease of a given patient.

The CLASS feature in clinvar\_conflicting.csv is a binary representation of whether or not a variant has conflicting classifications where 0 represents consistent classifications and 1 represents conflicting classifications.

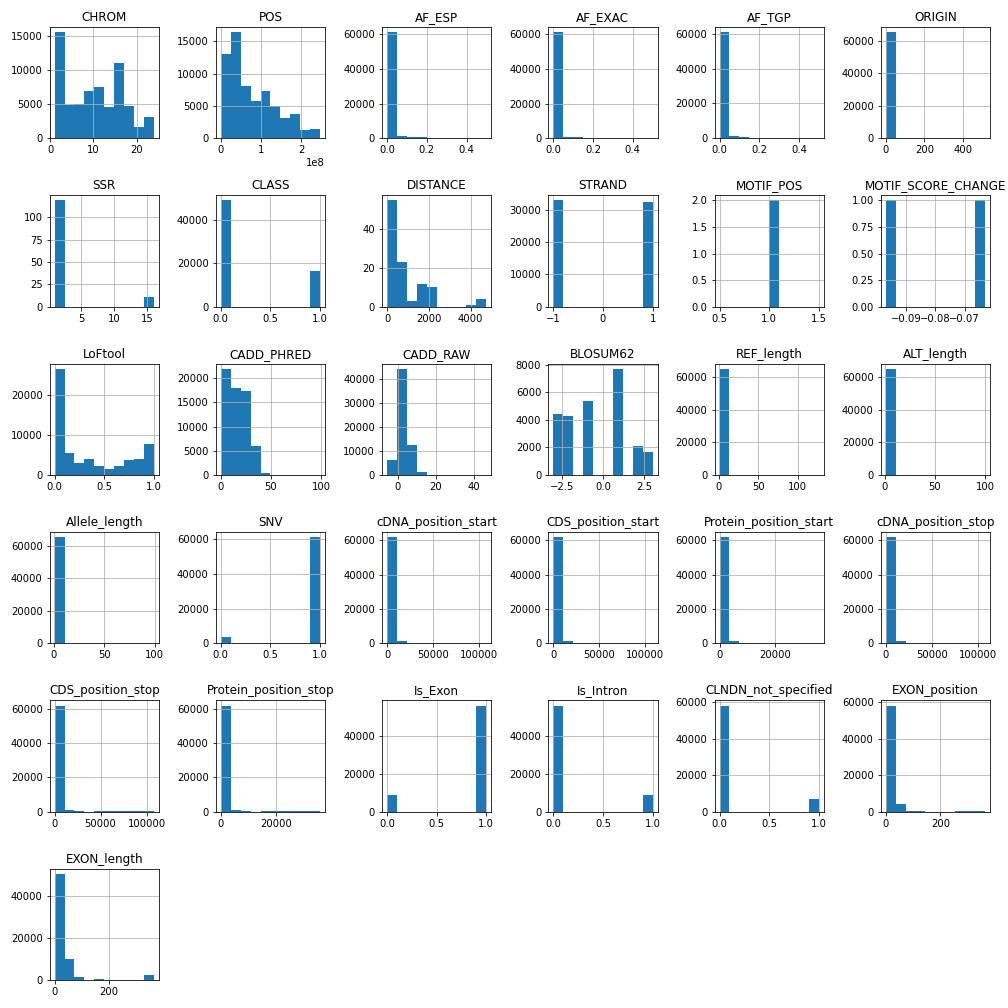
**METHODOLOGY**

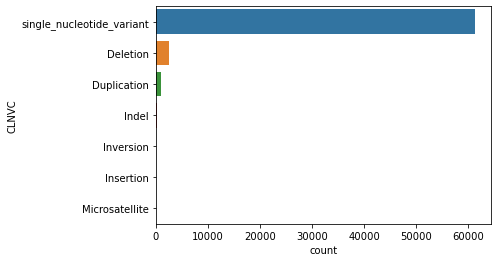
Python programming language was used throughout the work which was done using Jupyter notebook.

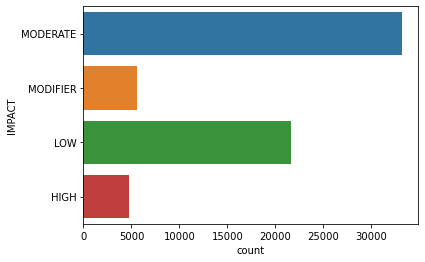
Different EDA techniques were deployed in order to achieve the desired goal of the questionnaire, but first data mining and cleaning was deployed to enable accuracy of facts pulled from the given datasets. Then I used best\_params\_, best estimators and then used confusion matrix to check accuracy of result in the machine learning.

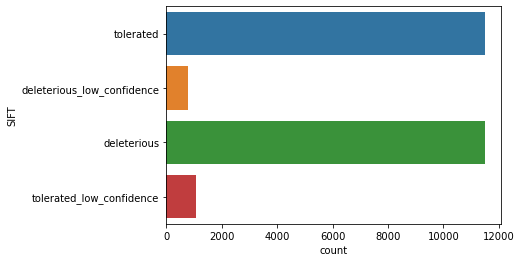
**RESULTS**

1. **Work on this dataset, bringing out all the useful insight and information:**

****

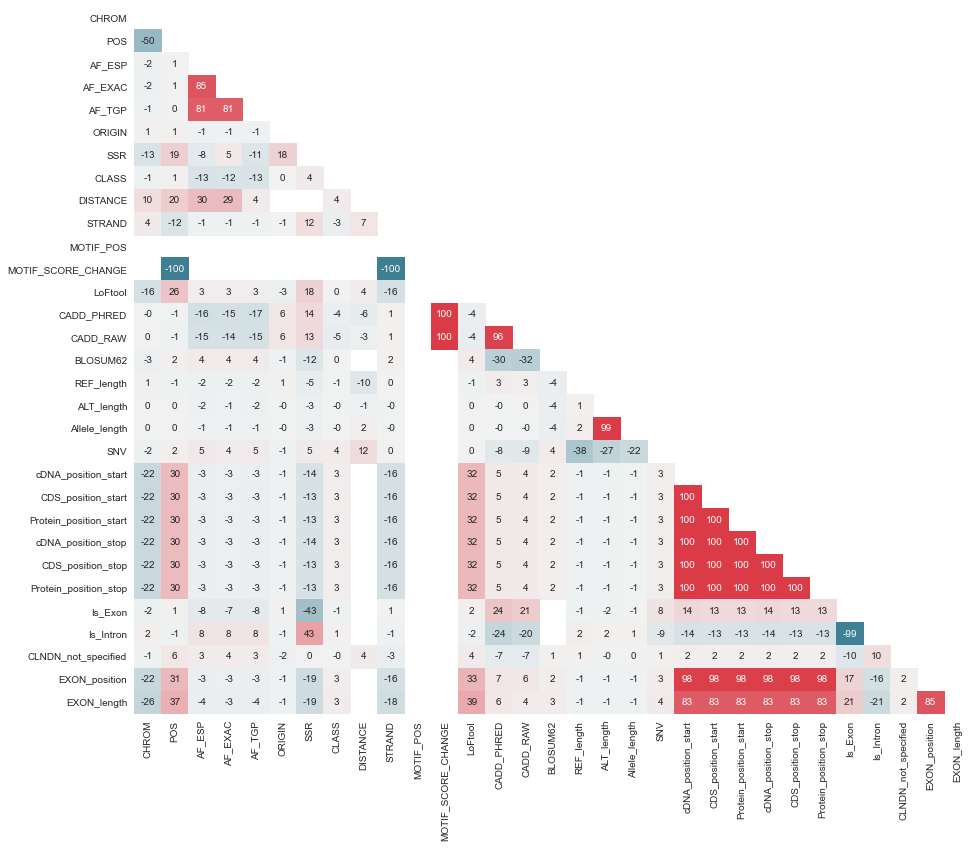
****

****

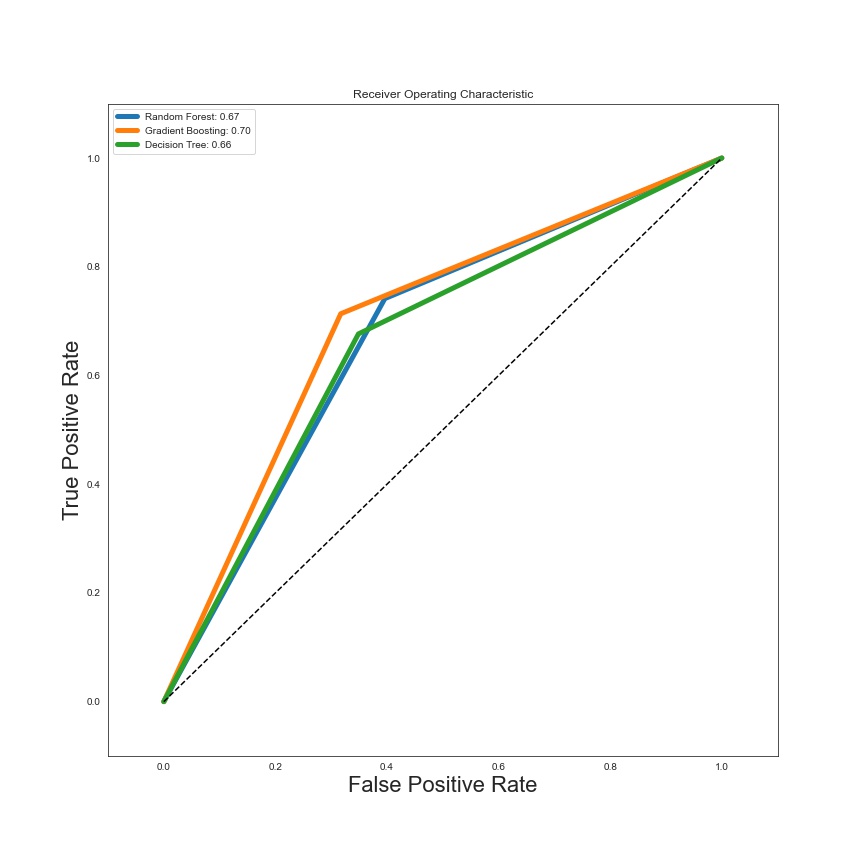
****

****

**Calculating the correlations between numeric features:**

****

**MACHINE LEARNING:**



**decision\_tree**

**precision recall f1-score support**

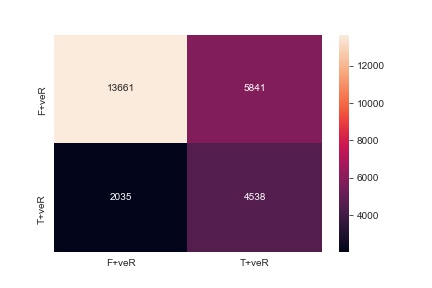
**0 0.87 0.70 0.78 19502**

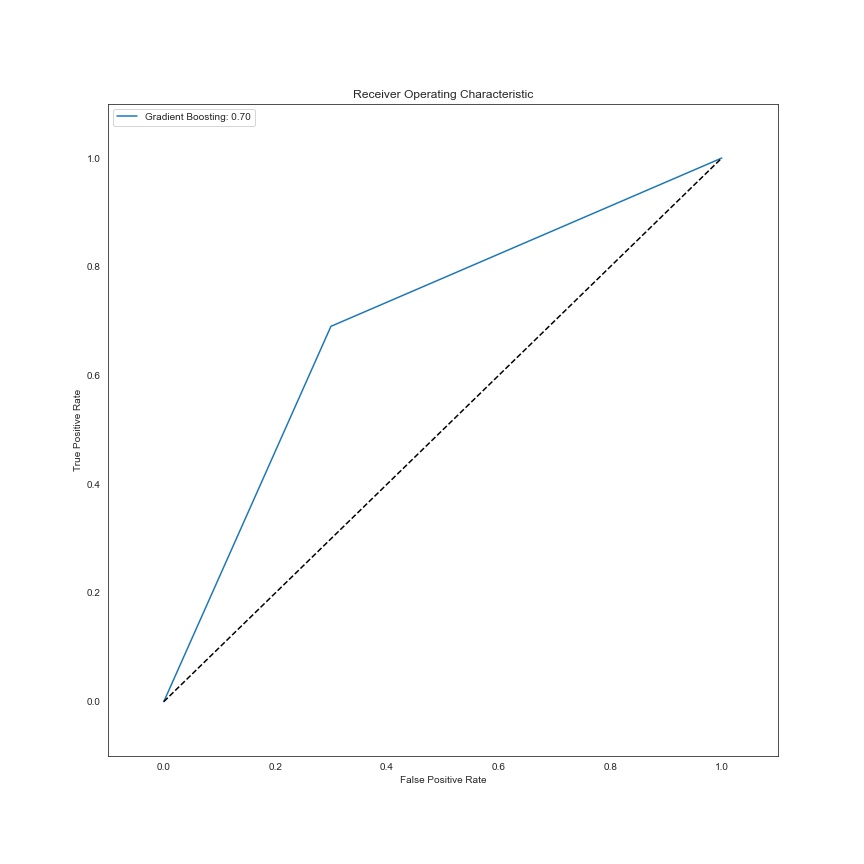
**1 0.44 0.69 0.54 6573**

**accuracy 0.70 26075**

**macro avg 0.65 0.70 0.66 26075**

**weighted avg 0.76 0.70 0.72 26075**





**CONCLUSION**

The final model yields an average Recall of 0.69, and a Recall of the positive class of 0.73. This means that out of the 6573 genetic variants with conflicting assessment in the test set our final model captures ca. 4800. Given the problem setting described above, we hope that this finding may be helpful for researchers and clinicians in identifying conflicting assessments of genetic variants. To further improve the prediction of genetic variants, one might try Support Vector Machines, Artificial Neural Networks, and Deep Learning. In addition, further preprocessing steps based on better domain knowledge might lead to additional performance improvements.

In general, we hope that some of the ideas presented in this notebook will encourage people to contribute and advance solution to the problem at hand'''