# **TECHNOCOLABS DATA ANALYSIS INTERNSHIP**

**PROJECT REPORT**

**TITLE:** Predicting the Severity of Adverse Drug Reaction

**AIM:**

The main focus of our project is to perform data analysis and to train the model using the Machine Learning algorithms in-order to find out the Adverse Drug Reaction.

**ABSTRACT:**

Clinicians use database in-order to determine the severity of the side effects from the prescribed drug combinations and also there are many drug combinations which are generally not found within such databases, though the side effects of those drugs are reported. So based on those things we used some Machine Learning Models – SVC, Random Forest, Voting Classifier and Xgboost to compare the accuracy and select the best suitable model for prediction.

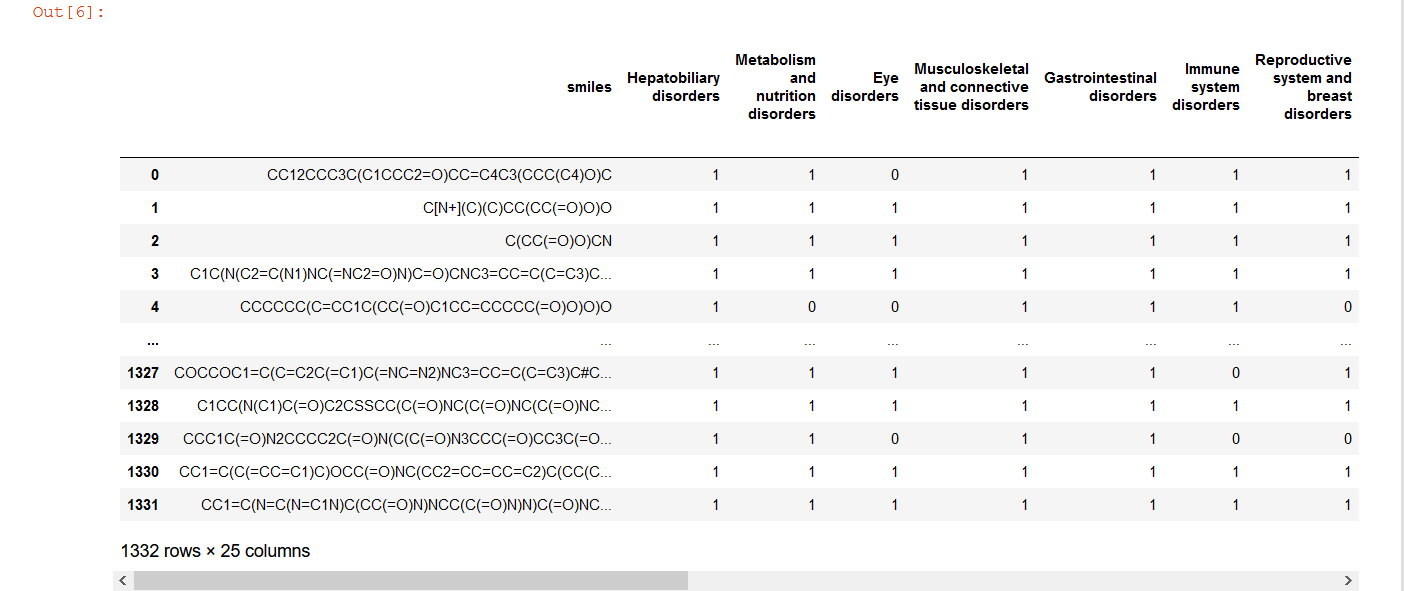
**INTRODUCTION:**

Polypharmacy, co-prescribing multiple drugs is generally common and it often leads to drug interaction that can have adverse side effects. Currently, to aid the doctors in prescribing treatments, the clinical decision systems gave alerts when these drug combination prescribed that have the known reactions and these alerts are based on the drug interaction severity stored in the databases such as Lexi-interact. But these databases like lexi-interact severity is based on clinical trials and literature reviews but still there are many other drug combinations which does not include all the drug interactions though there are many prescribed drug combinations that are not been formerly covered by literature. So for that we propose to use some Machine Learning models SVC, Random Forest, Voting Classifier and Xgboost.

**OVERVIEW:**

* Merging of the datasets.
* Adding the descriptors and fingerprints to the merged dataset.
* Implementing the Machine Learning model for prediction.

There were generally three datasets and those datasets were in CSV format i.e Comma Separated Value (.csv). The name of those datasets were sider, offsides\_socs and offsides\_socs\_modified. Among those the two datasets sider and offsides\_socs\_modified have unique smiles value and same column names except three different columns in the sider dataset which were not needed and those column names were product-issue, social-circumstances and investigations and after removing those columns the dataset were merge into one particular dataset which contains all unique smiles value with some disorders of each smiles.



Dataset before being merged.

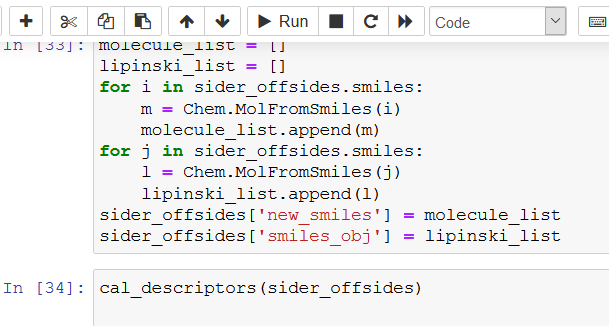
Rows = 1332, Columns = 25

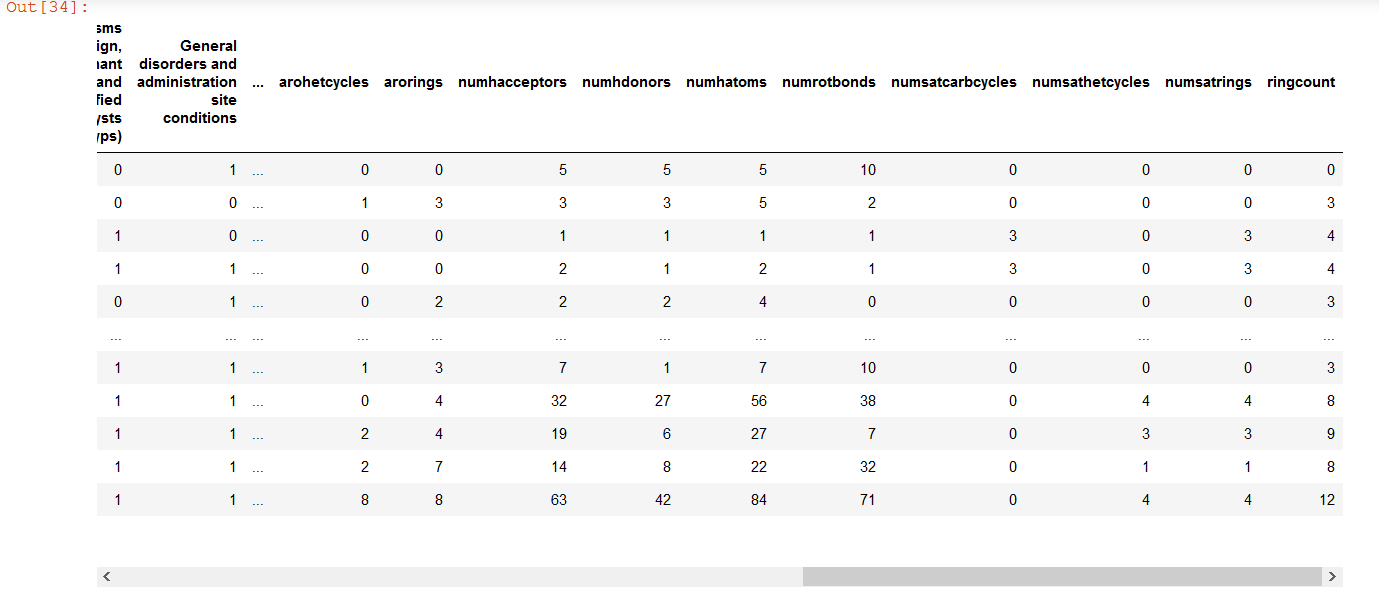
Then after that there was also a column related to the molecular structure of chemical formula in the combined dataset, below is the image given:



Fingerprints or descriptors are an abstract representation of certain structural features of a molecule. These descriptors may represent a structural key within a molecule. This might be as simple as a count of a particular atom type, S, N etc, or halogen, or sp3. It might be the presence of a particular ring system e.g. Phenyl, Pyridyl, Naphthyl, or a functional group e.g. Amide, Ester, Amine. It might be a calculated property Hydrogen Bond donor, Polar Surface area, LogP. Fingerprints are more abstract than a structural key but have the advantage of being more general since they do not represent pre-defined patterns.

Unlike a structural key with its pre-defined patterns, the patterns for a molecule's fingerprint are generated from the molecule itself. The fingerprinting algorithm examines the molecule and generates the fingerprint based on a set of rules.





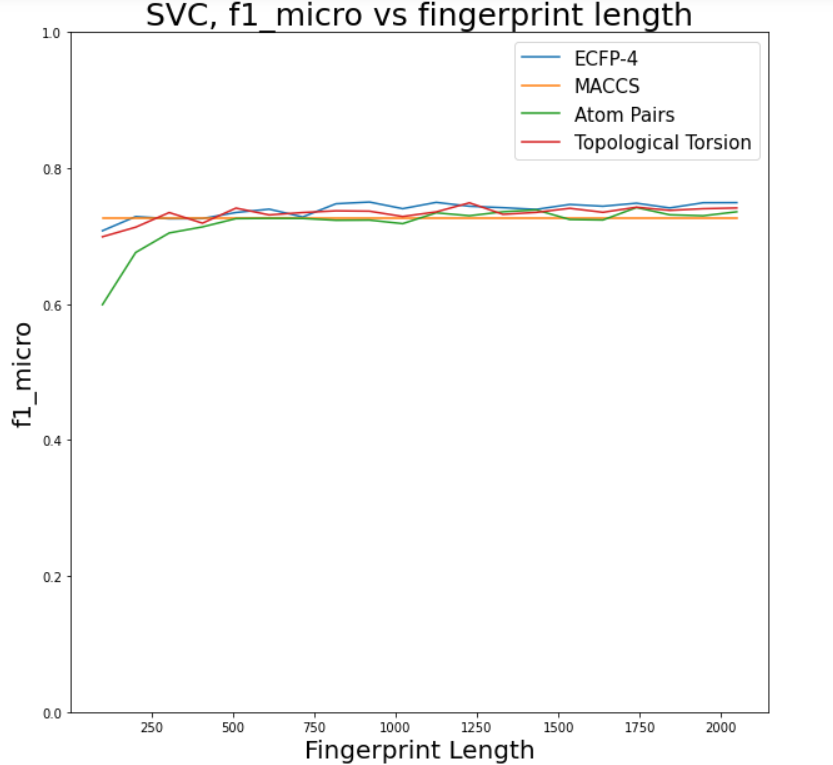
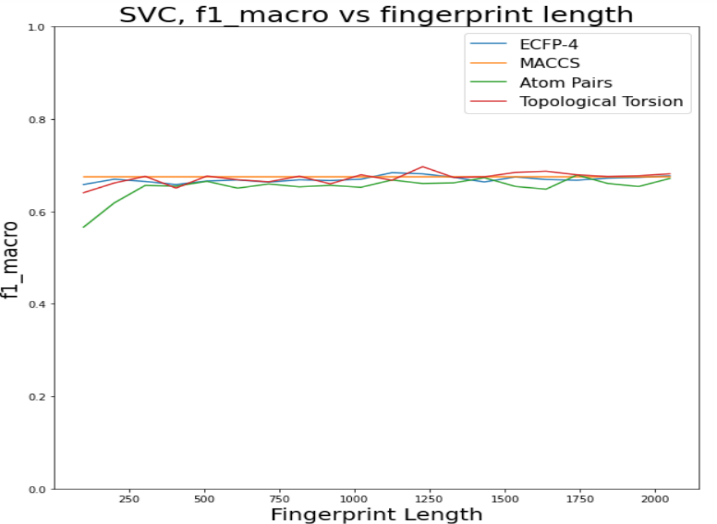
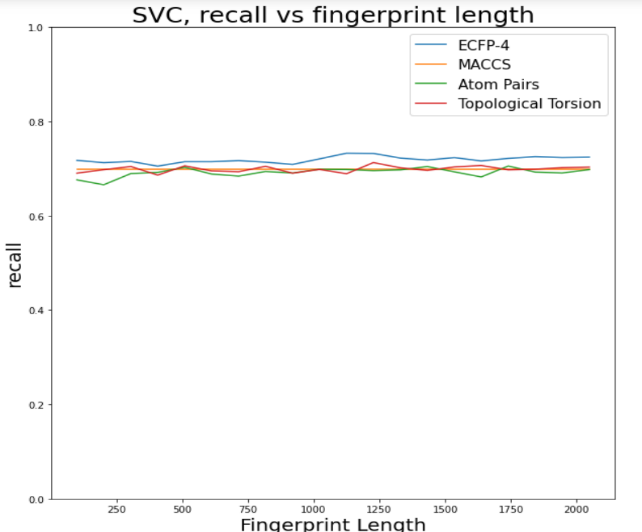
All the images shown above shows how the descriptors and fingerprints are created and then added into that merged dataset in-order to proceed further in the project.

**FINGREPRINT SELECTION AND COMPARISON WITH EVALUATION METRICS:**

Before we get into the modeling part first we created the function related to fingerprints and descriptor which would be used in the training of the model and so different function for fingerprints were created:

* Extended-Connectivity Fingerprints (ECFPs) are circular topological fingerprints designed for molecular characterization, similarity searching, and structure-activity modeling. They are among the most popular similarity search tools in drug discovery and they are effectively used in a wide variety of applications.
* Molecular Access System (MACCS) substructure fingerprints are 2D binary fingerprints (0 and 1), with each of 166 bits indicating the presence or absence of particular substructure keys . Daylight fingerprints and extended connectivity fingerprints (ECFP) extract chemical patterns of up to a specified length or diameter from a chemical graph. In comparison to the predefined substructure keys of MACCS, these fingerprints can dynamically index features using hash functions and often yield higher specificity when searching complex structures.
* The Morgan Fingerprint is basically a reimplementation of the extended connectivity fingerprint(ECFP). In essence you go through each atom of the molecule and obtain all possible paths through this atom with a specific radius.

And more related to these fingerprints and descriptors were created and then after creating the these function we generated some features and selection based on the evaluation metrics like f1-micro, precision, recall, f1-macro and compared them with the length of the fingerprint to get which evaluation metric is best suited with the fingerprint length. The images are shown below:

**Precision** is the ability of the classifier not to label as positive a sample that is negative and is defined by:

*𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛*=TPTP + FP

**Recall** is the ability of the classifier to label as positive a sample that is positive and is defined by:

*𝑅𝑒𝑐𝑎𝑙𝑙* = TPTP + FN

**Average Precision** summarizes a precision-recall curve as the weighted mean of precisions archived at each threshold (*𝑃𝑛*)

. The increase in recall (*𝑅𝑛*) from the previous threshold (*𝑅𝑛*−1)

is used as the weight. It is defined by:

*𝐴𝑃* = ∑(*𝑅𝑛*−*𝑅𝑛*−1)*𝑃𝑛*

**AUROC** is the area under the receiving operating characteristic (ROC) curve. This curve is created by plotting the fraction of TP out of the actual positives against the fraction of FP out of the actual negatives, at different thresholds.

**F1 Score** is a weighted average of the precision and recall and is defined by:

*𝐹*1 = 2⋅*𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛*⋅*𝑟𝑒𝑐𝑎𝑙𝑙𝑝𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛*+*𝑟𝑒𝑐𝑎𝑙𝑙*

In this work, three types of F1-score were used F1 binary, also represented as F1, is the F1 Score with respect only to the positive label. F1 Macro Score is the unweighted mean between both positive and negative labels. F1 Micro Score uses global TP, FN and FP and is equivalent to the accuracy metric in a binary classification task.

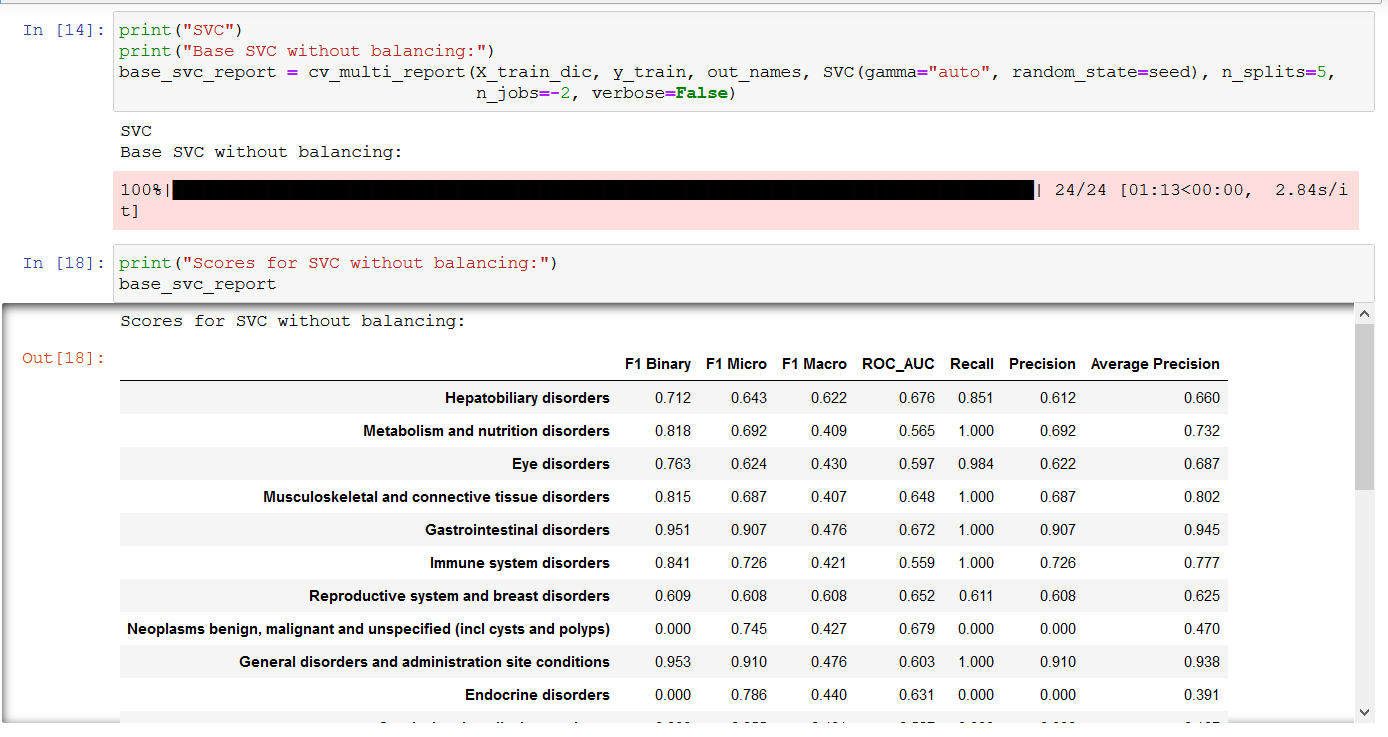
During this work, Average Precision, Recall and the different F1 Scores will be the main metrics used to evaluate and develop the model since they deal with imbalanced datasets better than AUROC.

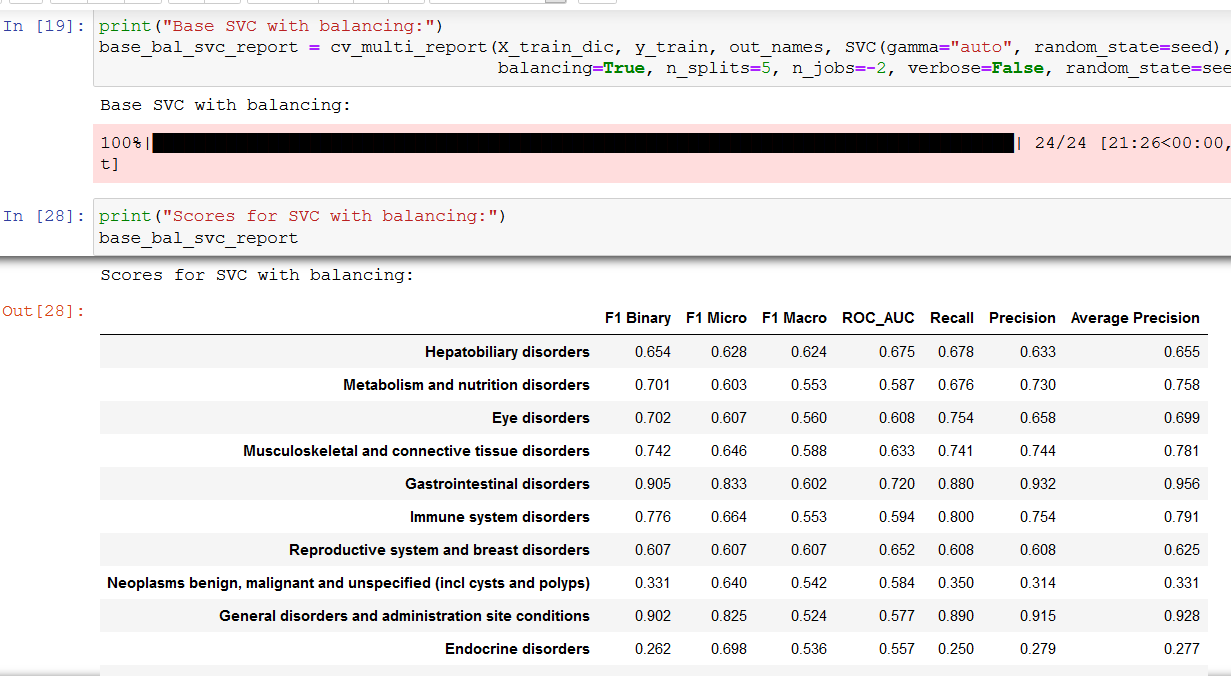
**MODEL DEVELOPMENT AND SELECTION:**

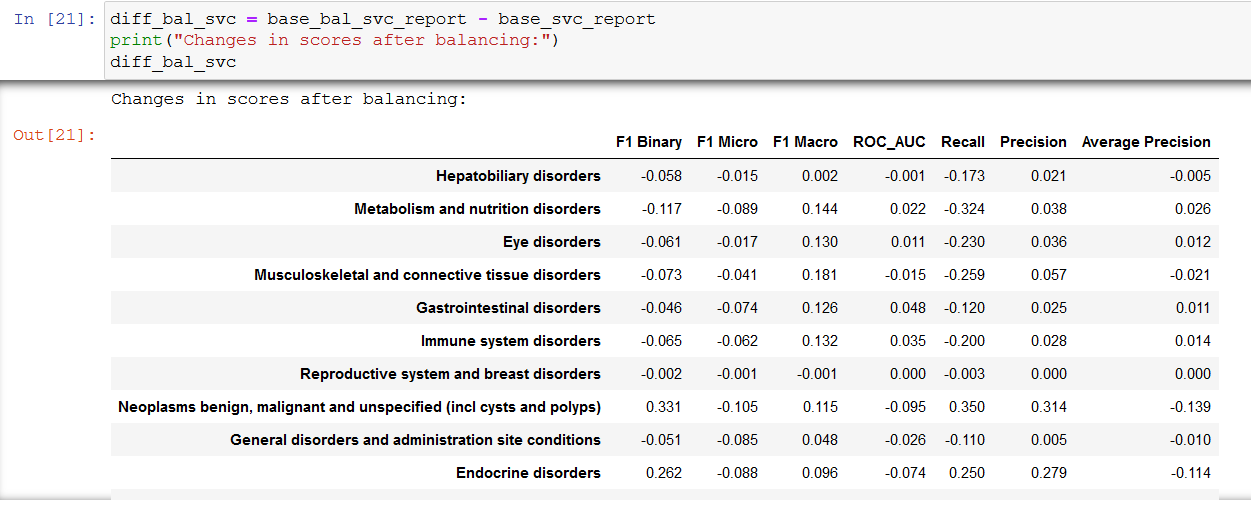
1. **SVC MODEL**

The first step was to test if the oversampling of the dataset improved the model.

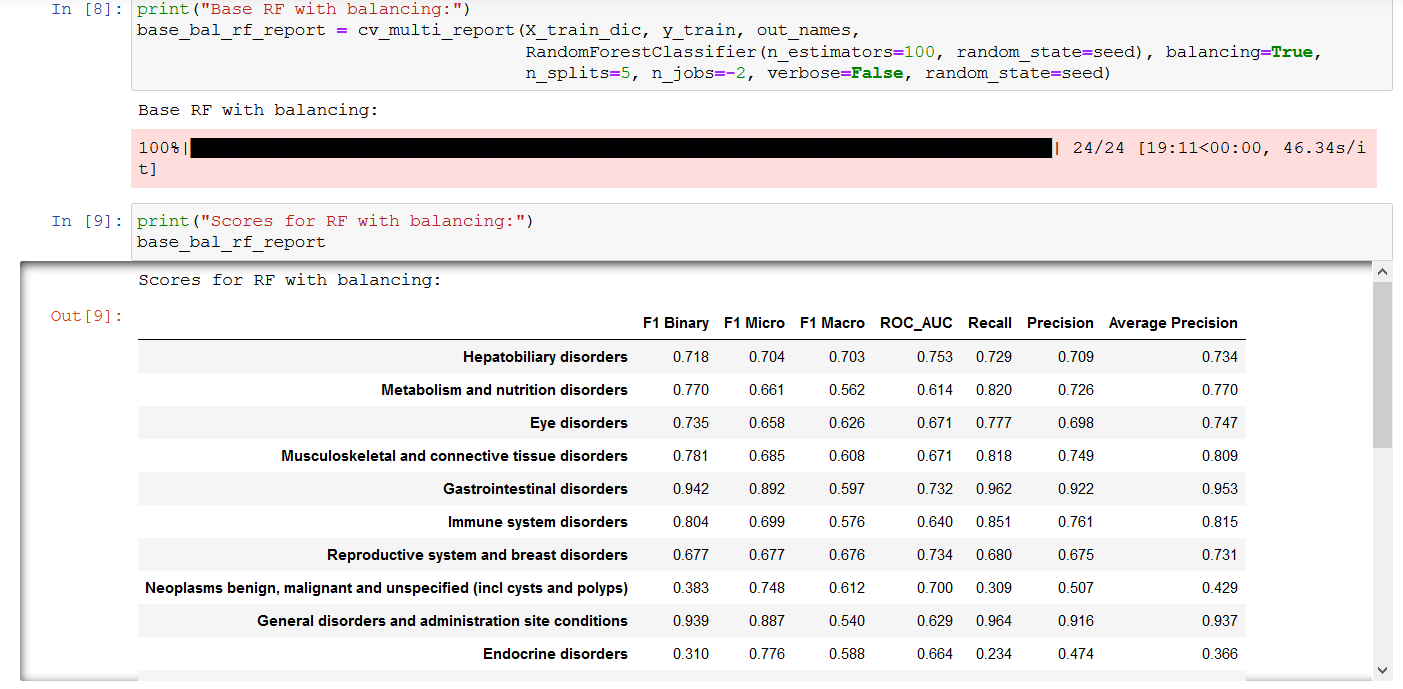
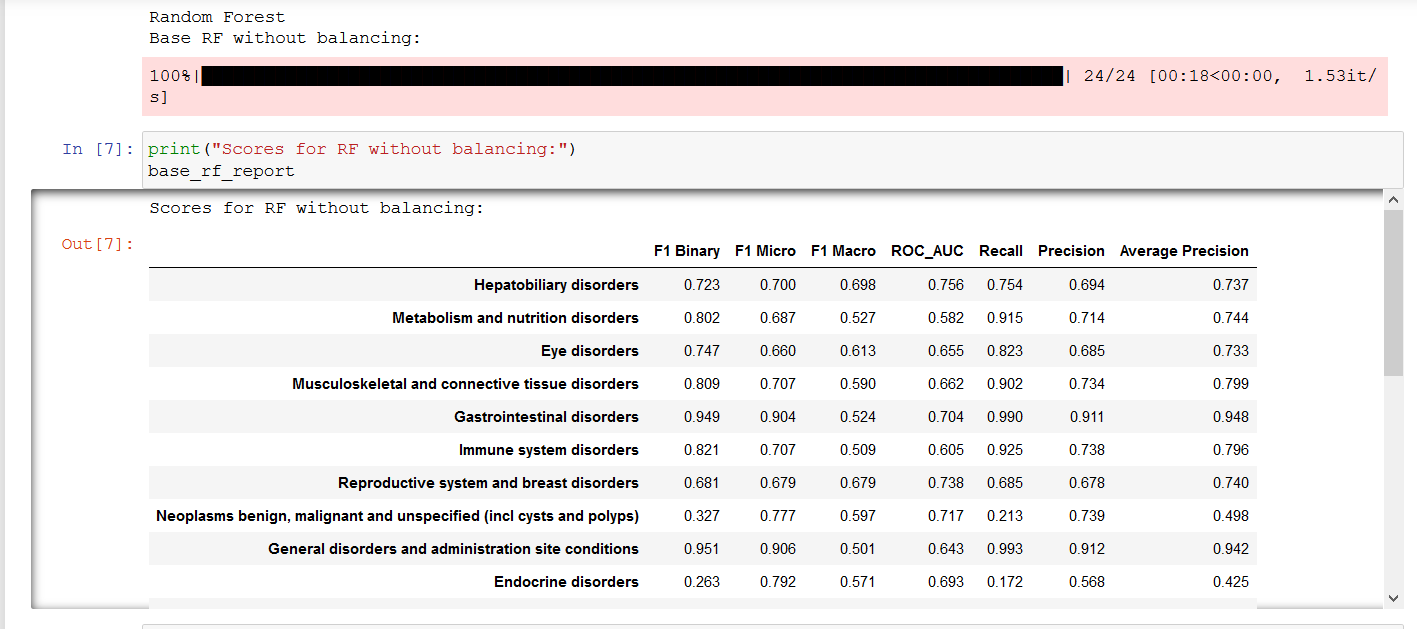
In this the images below tells about the changes in the score with balanced base SVC and unbalanced base SVC:





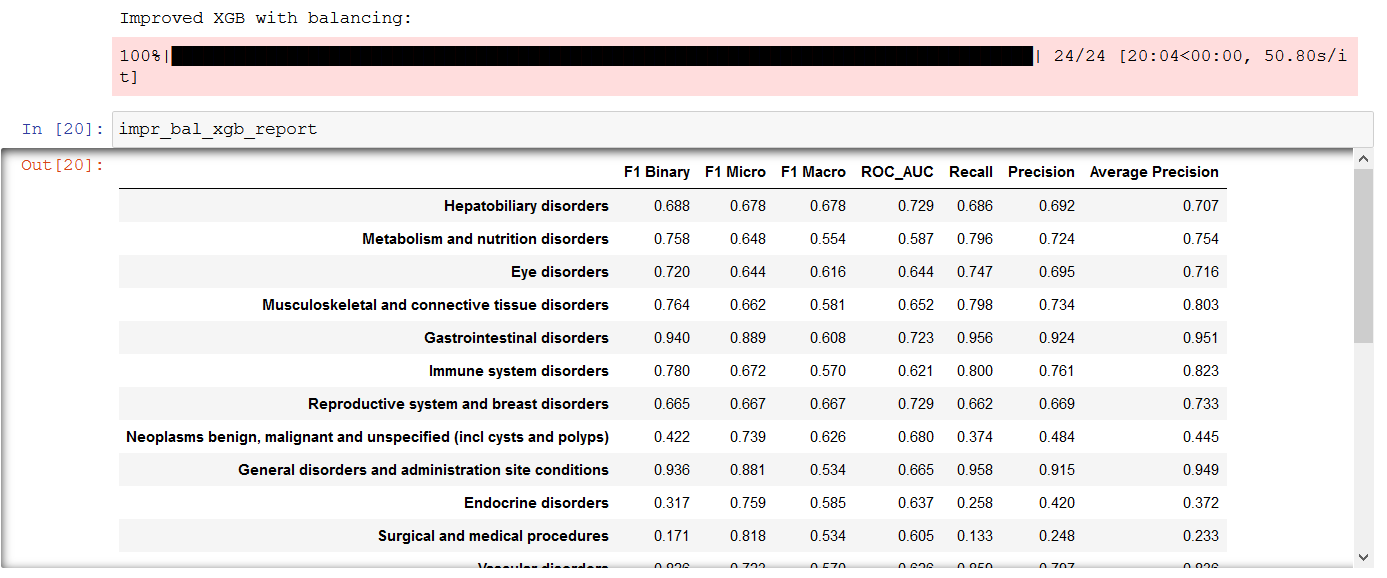


1. **RANDOM FOREST**

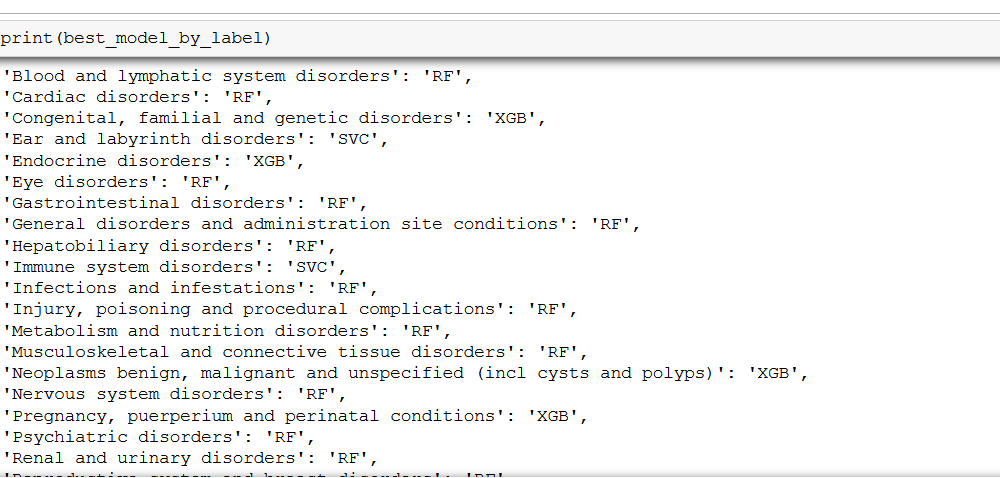


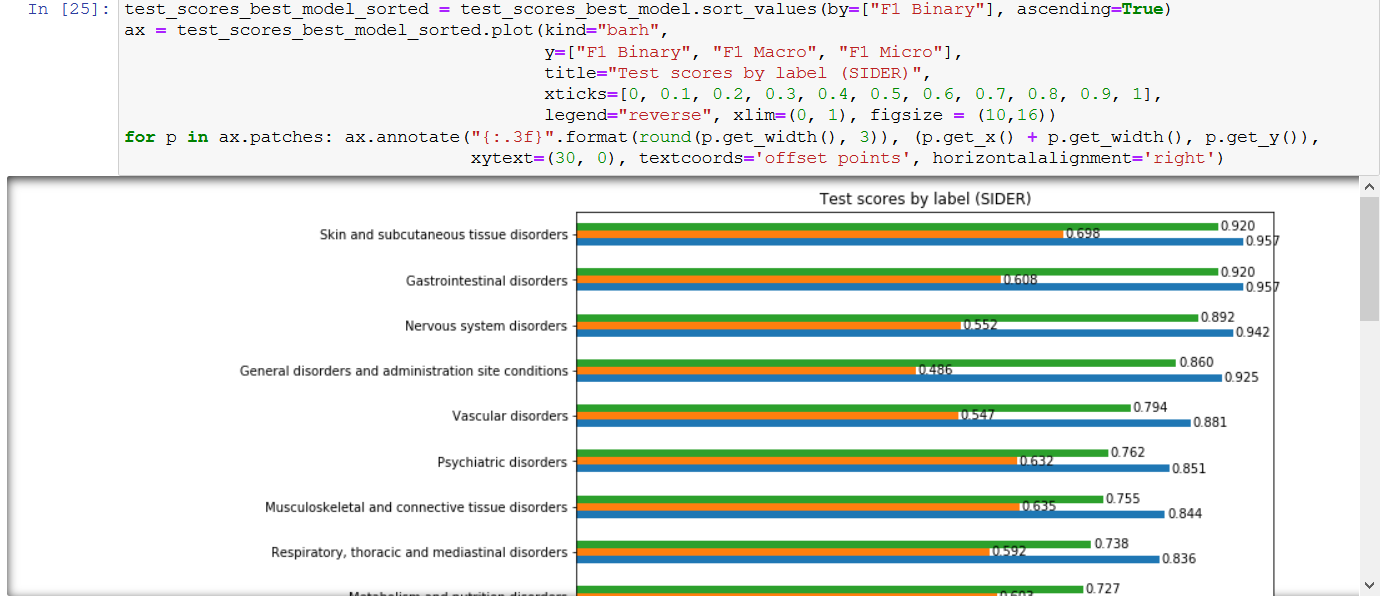
1. **XGBOOST**

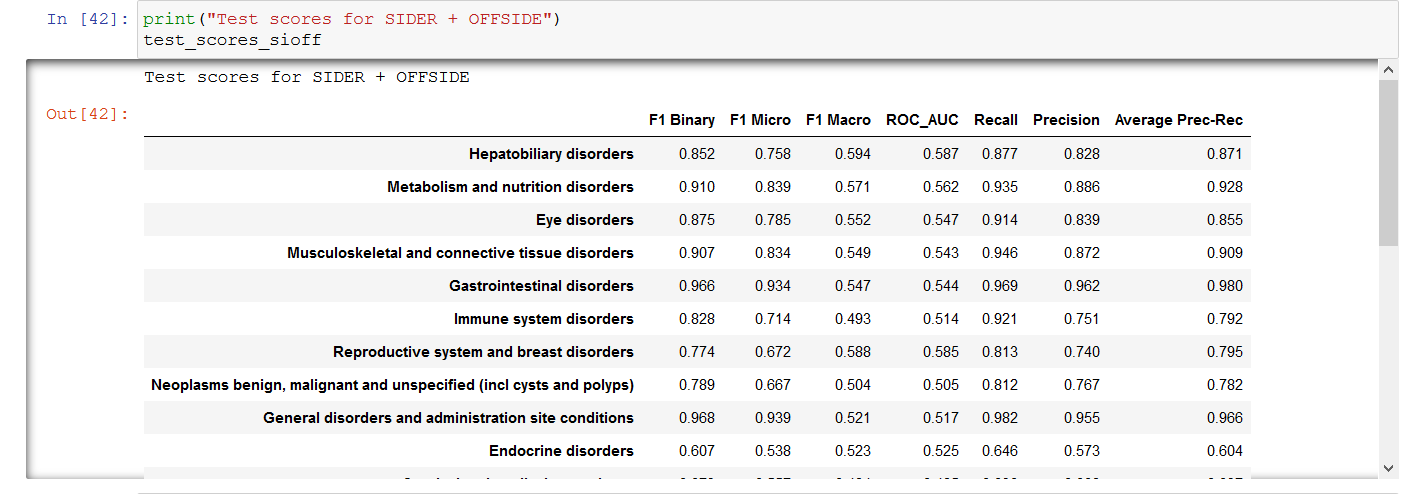


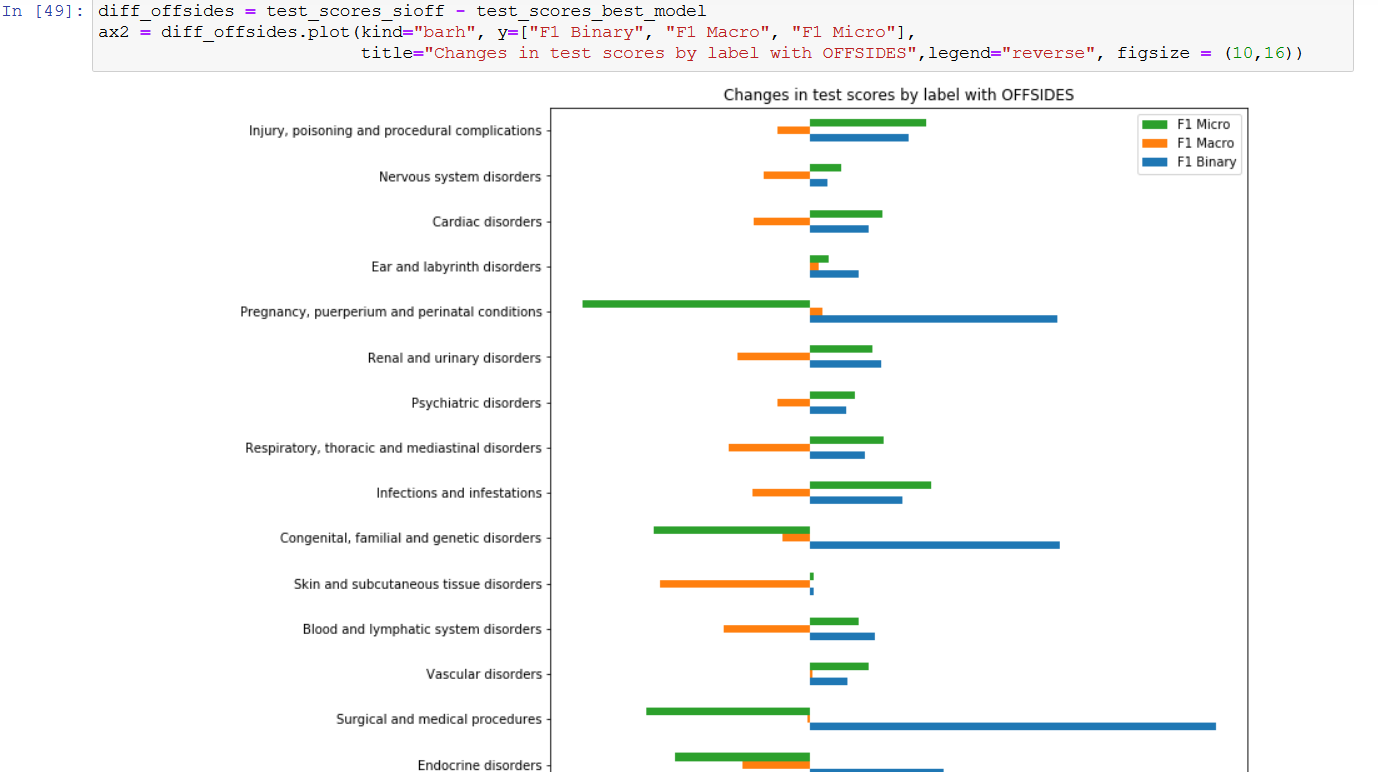


After developing these three models for each label, it was possible to observe that different models perform better for different tasks. As such, after observing the different metrics, it was manually selected, mainly by comparing Average Precision, F1 Binary and Macro, and Recall, the best model for each. Below are the images related to model selection and result of those model when applied with those datasets:









**TEAM MEMBERS NAME:**

**Vishnu**

**Shreyansh**

**Anushka**

**Tousif**

**Abhishek**

**Zeno**

**Bhuvaneshwari**