IITM-CS5691 : Pattern Recognition and Machine Learning
New Method Writeup

Release Date: 23 Oct, 2021
Due Date: 10 Nov, 2021

Roll No: CS21M023 Name: JAYESH MAHAJAN
Roll No: CS21M039 Name: NITIN BAHEKAR

- Dear Student, You may have tried different methods for predicting each of the clinical descriptors in the data contest. Submit a write-up of the methods chosen for the data contest in the template provided below. **You will have to add the details in your own words and submit it as a team in gradescope**.
- We will run plagiarism checks on codes/write-up, and any detected plagiarism in writing/code will be strictly penalized.

1. (points) [Name of the Method chosen for each descriptor, and its Paradigm: paradigm could be linear/non-linear models, etc.)]:

Solution:
CO1:
Name of method chosen: SVM
Paradigm: Linear model, Radial basis function kernel

CO2:
Name of method chosen: RandomForestClassifier
Paradigm: non-linear model, ensemble

CO3:
Name of method chosen: RandomForestClassifier
Paradigm: non-linear model, ensemble

CO4:
Name of method chosen: RandomForestClassifier
Paradigm: non-linear model, ensemble

CO5:
Name of method chosen: RandomForestClassifier
Paradigm: non-linear model, ensemble

CO6:
Name of method chosen: AdaBoost
Paradigm: non-linear model, ensemble

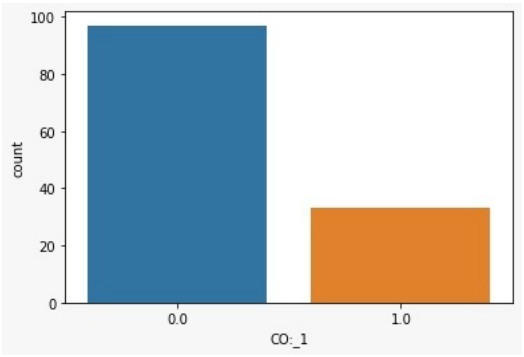
2. (points) [Brief description on the dataset: could show graphs illustrating data distribution or brief any additional analysis/data augmentation/data exploration performed]

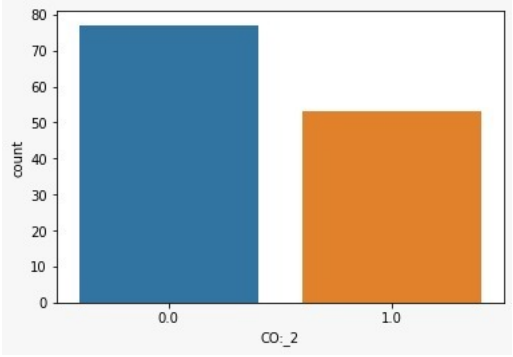
Solution:
The dataset provided for CO1 and CO2 had 130 samples and for CO3 to CO6 had 340 samples, which was very less data for training the models

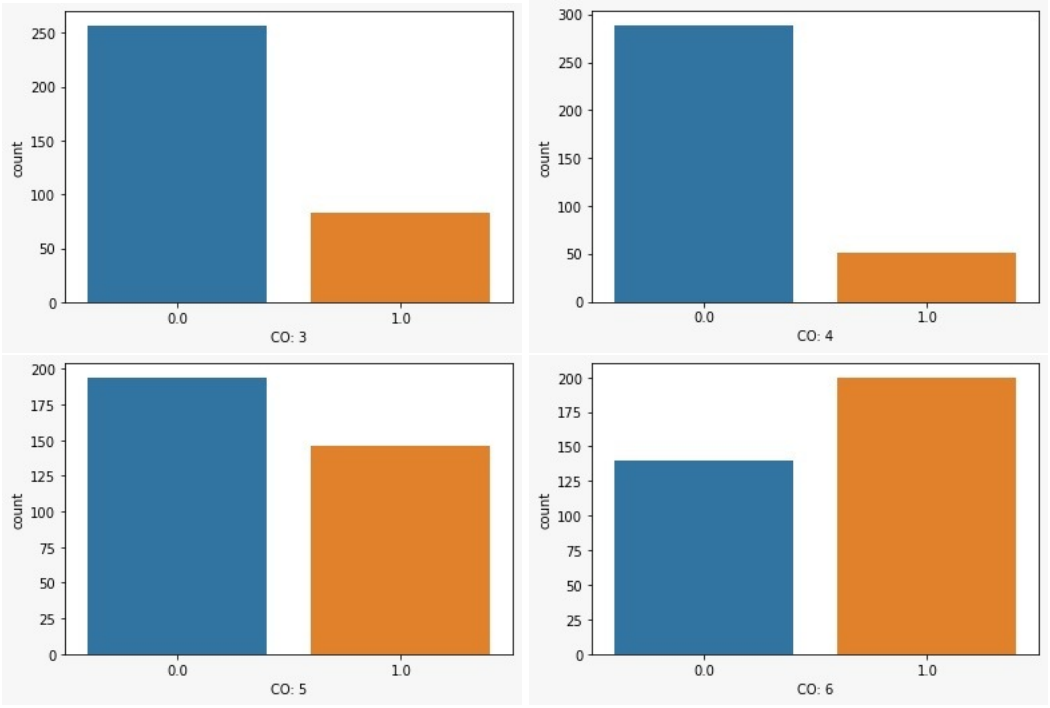
We also performed the Data Exploratory analysis. In which we found there was data imbalance problem. So, we used imblearn library to import **SMOTE** which is a type of data augmentation for minority class also called Synthetic Minority Oversampling Technique.

Also, the Dimensions or features of the data was very high. The dataset provided for CO1 and CO2 had 22283 dimensions for each sample and for CO3 to CO6 had 54675 dimensions for each sample, which was very high.

Below are the images showing Data-imbalance for all clinical descriptors:







3. (points) [Brief introduction/motivation: Describe briefly the reason behind choosing a specific method for predicting each of the descriptors (could show plots or tables summarizing the scores of training different model)]

Solution:
We have done hyper-parameter tuning on each model and found out the best parameter for each model

CO1 : We selected SVM as the score was highest among other models

model	**score**	**best params**
svm	0.903704	('C': 10, 'kernel': 'rbf')
random forest	0.896296	('max_depth': 3, 'n-estimators': 30, 'random-state': 1)
logistic regression	0.888889	('C': 1)

CO2 : We selected Random Forest as the score was highest among other models

model	score	best params
svm	0.786147	('C': 1, 'kernel': 'rbf')
random forest	0.787296	('max_depth': 5,'n-estimators': 30, 'random-state': 10)
logistic regression	0.776190	('C': 100)

CO3 : We selected Random Forest as the score was highest among other models

model	score	best params
svm	0.821987	('C': 1, 'kernel': 'linear')
random forest	0.846753	('max_depth': 5,'n-estimators': 30, 'random-state': 100)
logistic regression	0.819210	('C': 1)

CO4 : We selected Random Forest as the score was highest among other models

model	score	best params
svm	0.877685	('C': 5, 'kernel': 'rbf')
random forest	0.943086	('max_depth': 5,'n-estimators': 50, 'random-state': 1)
logistic regression	0.872747	('C': 1)

CO5 : We selected Random Forest as the score was highest among other models

model	score	best params
svm	0.875455	('C': 1, 'kernel': 'linear')
random forest	0.878047	('max_depth': 2,'n-estimators': 50, 'random-state': 1)
logistic regression	0.874411	('C': 50)

CO6 : We selected AdaBoost as the score was highest among other models

model	score	best params
svm	0.689286	('C': 1, 'kernel': 'linear')
random forest	0.657413	('max_depth': 3,'n-estimators': 10, 'random-state': 1)
logistic regression	0.678571	('C': 10)
AdaBoost	0.695962	(n-estimators=25,learning-rate=0.5,'random-state': 0)

4. (points) [As the data comes from a clinical setting, model interpretation is also an important task along with prediction. Describe briefly your preferred choice of methods if you are asked to determine the significant genes behind the regulation of the endpoints or restrict the number of features/genes.]:

Solution:
Yes, the model interpretation is also very important task along with prediction on the data. Our preferred choice of models will be the non-linear models such as tree based models such as

Random Forest or AdaBoost.
Because the models allows us to obtain the information on the feature/parameter importance. In Random Forest or AdaBoost, the feature importance is always there when training a model, so it is a great way to identify 'what' the model is learning.

5. (points) [Share your thoughts on each of the endpoints: whether easy/difficult to predict]:

Solution:
CO1: Was Easy to predict
After hyper-parameter tuning the SVM gave maximum accuracy among all classifiers

CO2: Was Moderate to predict
After hyper-parameter tuning the Random Forest gave maximum accuracy among all classifiers

CO3: Was Easy to predict
After hyper-parameter tuning the Random Forest gave maximum accuracy among all classifiers

CO4: Was Easy to predict
After hyper-parameter tuning the Random Forest gave maximum accuracy among all classifiers

CO5: Was Easy to predict
After hyper-parameter tuning the Random Forest gave maximum accuracy among all classifiers

CO6: Was Difficult to predict
After hyper-parameter tuning the AdaBoost gave maximum accuracy among all classifiers

6. (points) [Add any additional information: like challenges faced or some details that would help us to better understand the strategies that you have utilized for model development]:

Solution:
1. Not enough training data :
For clinical Descriptors CO:1 and CO:2 only 130 samples were present, and for CO:3, CO:4, CO:5, CO:6 only 340 samples were present, which made it difficult to increase the accuracy.

2. Imbalanced Data:
The data for the clinical descriptors provided came with imbalanced problem. We used smote to balanced the data. Smote is present in imblearn library. The Data Imbalance graphs have been shown in Question-2.

3. Used Ensemble Techniques :
First we used logistic regression, Gaussain Naive bayes and Support vector machines, but the accuracy was low for these classifiers, So, we switched to Ensemble models, out of which Random Forest and AdaBoost performed the best in terms of accuracy.

4. Very High dimensional Data:
The first two clinical descriptors had features above 22283 and last four clinical descriptors had features above 54675, so which made it difficult to understand the data.