IITM-CS5691 : Pattern Recognition and Machine Learning Release Date: 23 Oct, 2021
New Method Writeup Due Date: 12 Nov, 2021

Roll No: CS21M029 , CS21M055 Name: Keyur Bharatkumar Raval , Makwana Rutik M.

- 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:
we have used $\cdots$
For CO1 : logistic regression
paradigm : linear model

For CO2 : logistic regression
paradigm : linear model

For CO3 : logistic regression
paradigm : linear model

For CO4 : AdaBoost with decision stump
paradigm : non-linear model

For CO5 : AdaBoost with decision stump
paradigm : non-linear model

For CO6 : AdaBoost with decision stump
paradigm : non-linear model

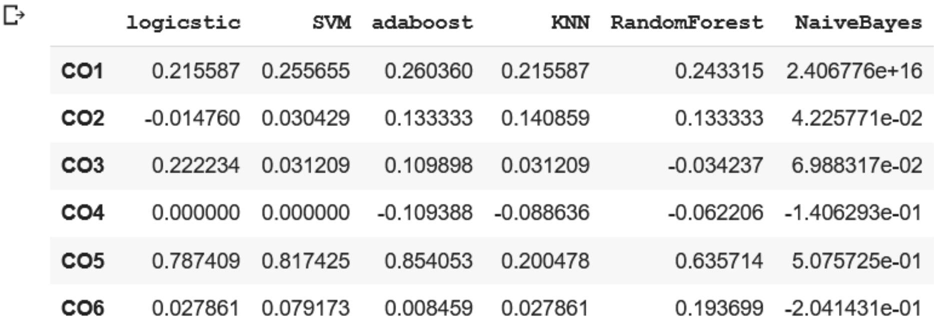
2. (points) [Brief description on the dataset: could show graphs illustrating data distribution or brief any additional analysis/data augmentation/data exploration performed]

Solution: we have used data augmentation to balance the data cause for CO1,CO3 and CO4 there are so many 1′s then 0′s . Hence we have added data which has 1′s in row for column CO1,CO3 and CO4 to the original data so that data should not be imbalanced , after that we have used imblearn to generate new data point for minority classes.

Hence we have build 2 models and that we have choose in Kaggle to one without taking care of imbalance data and one with imlearn which will take care of data.

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 implemented Logistic regression , SVM , adaboost , KNN , RandomForest and NaiveBayes for each co1 to co6 and we have submited at kaggle .
After submitting at kaggle we have realized that some of the classifier perform well in train data and bad in testing data.
Adaboost was giving overall good mcc score but for co1 and co2 Logistic regression was giving the best consistant result so we have used Logistic regression in co1,co2 and co3 and Adaboost in co4,co5and co6.
here given below table denotes mcc score not accuracy on Trainng data.



	logicstic	SVM	adaboost	KNN	RandomForest	NaiveBayes
CO1	0.215587	0.255655	0.260360	0.215587	0.243315	2.406776e+16
CO2	-0.014760	0.030429	0.133333	0.140859	0.133333	4.225771e-02
CO3	0.222234	0.031209	0.109898	0.031209	-0.034237	6.988317e-02
CO4	0.000000	0.000000	-0.109388	-0.088636	-0.062206	-1.406293e-01
CO5	0.787409	0.817425	0.854053	0.200478	0.635714	5.075725e-01
CO6	0.027861	0.079173	0.008459	0.027861	0.193699	-2.041431e-01

Here is the link of each classifier my google colab notebook from where we have created above table:
logicstic
SVM
adaboost
KNN
RandomForest
NaiveBayes

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: For restrict the number of genes or we can say restric the number of feature we could use the feature selection method from sklearn or if our goal is to restrict the number of features then we can do principal component analysis(PCA) and linear discriminant analysis(LDA) , we have done PCA and LDA but they did not gave better accuracy so we didn't use it in our model.But if we want to restrict the number of features/genes then we can do PCA,LDA or use feature selection method from sklearn.

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

Solution: CO1 : CO1 is we wolud say it is moderate to predict not too easy not too hard.
CO2 : CO2 is hard to predict
CO3 : CO3 is easy too predict
CO4 : CO4 is also very hard to predict cause of 2 reason cause of imbalanced data more 1's than 0's and class boundary is also complicated.
CO5 : CO5 is easiest to predict it is linearly separable it's giving good MCC score with all classifier.
CO6 : CO6 is also quite hard to predict.

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:

At First our strategie was to plot the feature vector but there are way too much features so we have done PCA on data and then apply the model but it was not giving the good accuracy so then we have directly trained different model on data there we have observed that CO2 and CO5 are easy to predict where as CO4,CO6 are hard to predict.Also there we have observed that some of the class have imbalance data . After applying different combiation of model and hyperparameters we have decided to go with co1 to co3 with logistic regression and Co4 to Co6 with adaboost.

After that we tried to oversample the minority class and add it back to original data and then train the model but that did not improve the mcc score the score was more or less same , then we have used imblearn to generate the new data point and that also did not increase our mcc score but we strongly believe that that model will do good on 70 percentage of private data.