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

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

Roll No:CS21M005 Name:(Ankit Kharwar)
Roll No:CS21M059 Name :(Sarvagya Sriram Mamgain)

- 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: After applying all the classifiers ,we have come to the best one which is best in predicting each of the descriptors.

We have build the Matthews Correlations table after applying different classifiers in our models.

The Matthews Correlations table will looks like the followings.

	CO 1 (0-99)	**CO 2 (100-199)**	**CO 3 (200-413)**	**CO 4 (414-627)**	**CO 5 (628-841)**	**CO 6 (842-1055)**
Logestic Regression	0.583	0.54	0.31	0.229	0.55	0.08
Ada Boost	0.364	0.4	0.02	0.11	0.77	0.153
Naive Bayes	0.58	0.41	0.02	0.104	0.15	0.12
Decision Tree	0.233	0.42	0.26	-0.19	0.69	0.11
SVC	0.583	0.54	0.3	0.229	0.71	0.016
Random Forest	0.3	0.32	0.19	0.0	0.33	0.01

Matthews Correlation Coefficients

As, from the above,Figure which we calculated using Matthews correlation table we can see that

(i) For descriptor "CO 1": 0-99 :- The Best suited Classifier will be the **"Logestic Regression"** , because for the logestic regression we get the predicted score of **0.583**.

All the other than Logestic regression are smaller than others classifiers.

Thus, Logestic regression is best for descriptor for "CO: 1" :0-99.

(ii) For descriptor "CO 2": 100-199 :- The Best suited Classifier will be the "Logestic Regression" because for the logestic regression we get the predicted score of **0.54.**
All the other than Logestic regression are smaller than others classifiers.

Thus, Logestic regression is best for descriptor for "CO 2" :100-199.

(iii) For descriptor "CO 3": 200-413 :- The Best suited Classifier will be the **"Logestic Regression"** because for the logestic regression we get the predicted score of **0.31**.
All the other than Logestic regression are smaller than others classifiers.

Thus, Logestic regression is best for descriptor for "CO 3" 200-413.

(iv) For descriptor "CO 4": 414-627 :- The Best suited Classifier will be the **"Logestic Regression"** because for the logestic regression we get the predicted score of **0.299**.
All the other than Logestic regression are smaller than others classifiers.

Thus, Logestic regression is best for descriptor for "CO 4": 414-627.

(v) For descriptor "CO 5" :628-841 :- The Best suited Classifier will be the **"Ada- Boost"** because for the Ada- Boost we get the predicted score of **0.77**.
All the other than "Ada- Boost" are smaller than others classifiers.

Thus, "Ada- Boost" is best for descriptor for "CO 5": 628-841.

(vi) For descriptor "CO 6" :842-1055 :- The Best suited Classifier will be the **"Ada- Boost"** because for the Ada- Boost we get the predicted score of **0.153**
All the other than Ada- Boost are smaller than others classifiers.

Thus, Ada- Boost is best for descriptor for "CO 6": 842-1055.

Now,in the **CO1,CO2,CO3.CO4** we use Logistic regression, which is a linear model.We train it using Sklearn library and using classifier as following.

model name = LogisticRegression()

for the **CO5,CO6**,we have use the Ada-Boost classifier which is the non-linear model and we have train it using the following nestimators , learningrate.

n-estimators = 20

learning-rate = 0.4578938985893898

We, use the model as
model name =AdaBoostClassifier(n_estimators=n_estimators, learning_rate=learning_rate)

Thus,after applying all this we get the score in the data contest as **0.53349**, which is the best Score we get in all the attemps.

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

Solution: Datasets are as follows :-

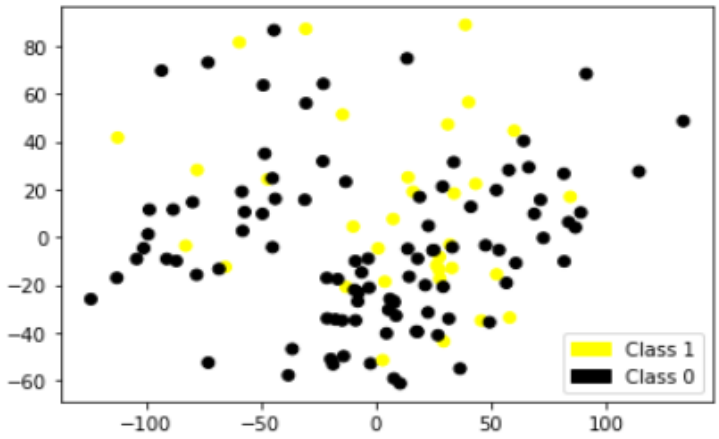
For the "Dataset1Training.csv":- We have 22285 rows × 130 columns, where all the Columns actually represent the Data Samples and the first 22283 Rows are the genes(ie, the features given to us).
The Second last row is actually the descrpitor "CO: 1" and the last row is actually the descrpitor "CO: 2".
For its respective test data, we don't have these CO: 1 and CO: 2, and actually these only we have to find out.

Similarly ,
For the "Dataset2Training.csv":- , We have 54679 rows × 341 columns where all the Columns actually represent the Data Samples and the first 54675 Rows are the genes(ie, the features given to us).
Now the last 4 rows are the descriptors CO: 3,CO: 4,CO: 5 and CO: 6 respectively.
For its respective test data, we don't have these CO: 3,CO: 4,CO: 5 and CO: 6, and actually these only we have to find out.

Scaling the data:-We have used the StandardScaler() for scaling our data for buildng the models. It actually standardize features by removing the mean and scaling to unit variance.

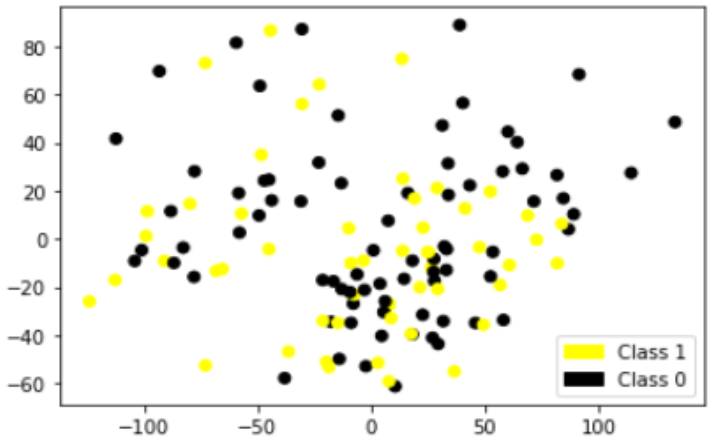
Also as we have so many features, thus for Plotting the graph to show the data distribution, we have used **Singular Value Decomposition** to get Best 2 Features and then using those we have plotted the graphs and Each descriptor has 2-class values , ie, either class 1 or class 0. Thus , we show these two classes with different colors respectively.
We have analyse following and we have describe by ploting the two class class 1 and class 2

(i) For descriptor "CO 1": 0-99 :- The Best suited Classifier will be the **"Logestic Regression"** , because for the logestic regression we get the predicted score of **0.583**.All the other than Logestic regression are smaller than others classifiers.The ploting of the discriptor Co:1 with two class will look like the following:-



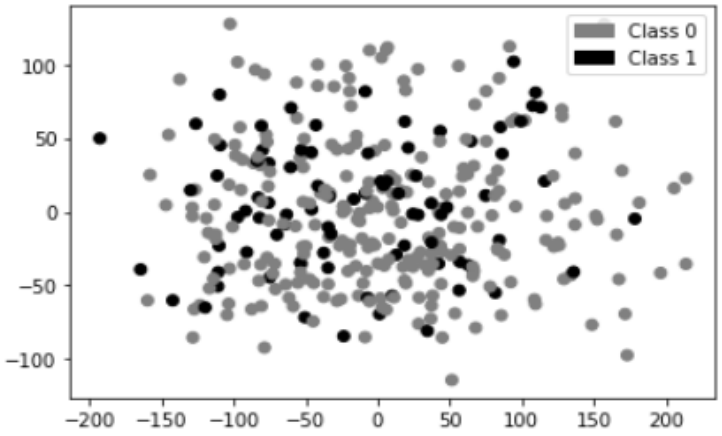
Descriptor "CO 1" :0-99

(ii) For descriptor "CO 2": 100-199 :- The Best suited Classifier will be the "Logestic Regression" because for the logestic regression we get the predicted score of **0.54.** All the other than Logestic regression are smaller than others classifiers. The ploting of the discriptor Co:1 with two class will look like the following:-



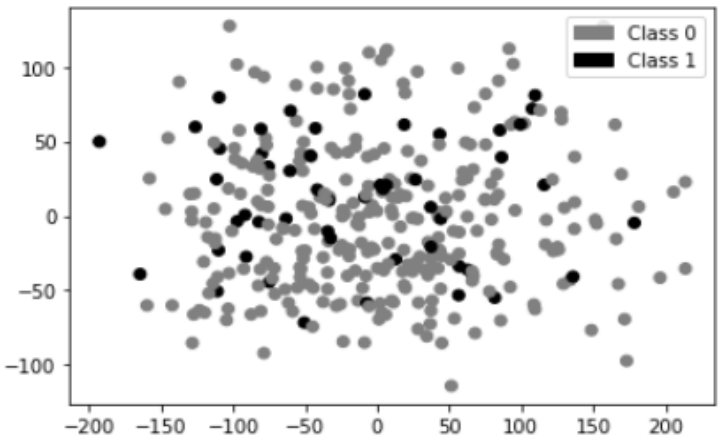
Descriptor "CO 2" :100-199

(iii) For descriptor "CO 3": 200-413 :- The Best suited Classifier will be the **"Logestic Regression"** because for the logestic regression we get the predicted score of **0.31**. All the other than Logestic regression are smaller than others classifiers. The ploting of the discriptor Co:1 with two class will look like the following:-



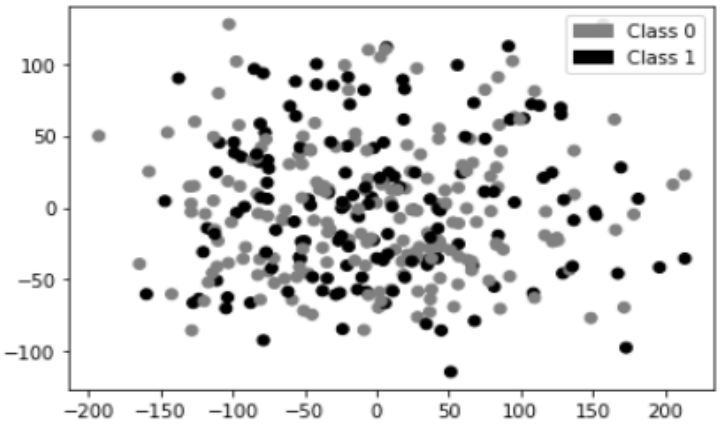
Descriptor "CO 3" :200-413

(iv) For descriptor "CO 4": 414-627 :- The Best suited Classifier will be the **"Logestic Regression"** because for the logestic regression we get the predicted score of **0.299**.All the other than Logestic regression are smaller than others classifiers.The ploting of the discriptor Co:1 with two class will look like the following:-



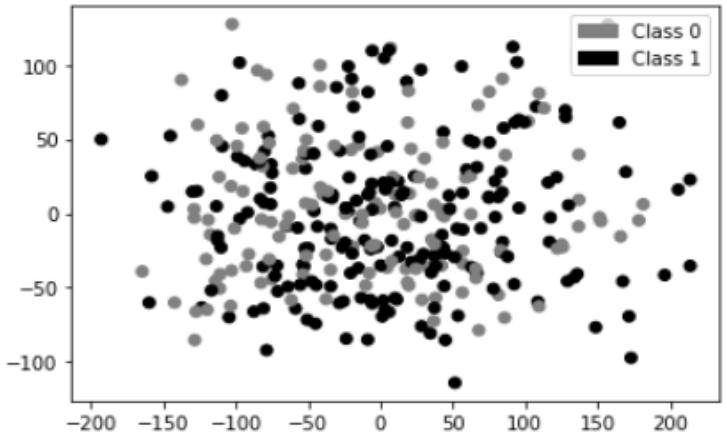
Descriptor "CO 4" :414-627

(v) For descriptor "CO 5" :628-841 :- The Best suited Classifier will be the **"Ada- Boost"** because for the Ada- Boost we get the predicted score of **0.77**.All the other than "Ada- Boost" are smaller than others classifiers.The ploting of the discriptor Co:1 with two class will look like the following:-



Descriptor "CO 5" :628-841

(vi) For descriptor "CO 6" :842-1055 :- The Best suited Classifier will be the **"Ada- Boost"** because for the Ada- Boost we get the predicted score of **0.153**.All the other than Ada- Boost are smaller than others classifiers.The ploting of the discriptor Co:1 with two class will look like the following:-



Descriptor "CO 6" :842-1055

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: As our solutions in Kaggle Data Contest are evaluated based on "Matthews Correlation Coefficient" , thus we made 80-20 splitting on the training dataset , training on 80 percent and and found the Matthews Correlation Coefficient on 20 percent . So, "Matthews Correlation Coefficient" for different models being analysed and best one being chosen for each model was our intension behind choosing the models. Different models with thier "Matthews Correlation Coefficient" were as follows :-

The Matthews Correlations table will looks like the followings.

	CO 1 (0-99)	**CO 2 (100-199)**	**CO 3 (200-413)**	**CO 4 (414-627)**	**CO 5 (628-841)**	**CO 6 (842-1055)**
Logestic Regression	0.583	0.54	0.31	0.229	0.55	0.08
Ada Boost	0.364	0.4	0.02	0.11	0.77	0.153
Naive Bayes	0.58	0.41	0.02	0.104	0.15	0.12
Decision Tree	0.233	0.42	0.26	-0.19	0.69	0.11
SVC	0.583	0.54	0.3	0.229	0.71	0.016
Random Forest	0.3	0.32	0.19	0.0	0.33	0.01

Matthews Correlation Coefficients

Code for it is simple , "from sklearn.metrics import matthews_corrcoef" we have to print the matthews_corrcoef value for any model between Actual and Predicted Output using : "print(matthews_corrcoef(Actual Output, Predicted Output))"

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: The Methods I will opt for if asked to restrict the number of features will be surely Principal Component Analysis(PCA) or Singular Value Decomposition(SVD) as they are Use to reduce the Number of Features and find the best K features that are most useful to us.

Also, In the code we made, We have tried to Apply SVD,PCA , but as we found no improvement in the accuracy thus we finally not opted for it.

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

Solution:

It was bit difficult as When we found out the Classes for each of the descrptors, we could find out as discussed above that there was class imbalancing . In CO:1 , CO:2 , CO:5 and CO:6 we can see that the number of zeros and Ones are not too much deviated from each other, but in CO:3 and CO:4 we see that Number of Zeros Overpower the number of Ones, thus class Imbalancing is over here.Hence a problem.

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: There were various challenges that were here , like the Number of Features are too many and number of data samples are too less, thus it is also a point to be noted, it is called "Big p , Little n" Problem. For this we tried to Reduce the Features first by trying Singular Value Decompostion, but We found out no Improvement, thus not went for it. We also did some preprocessing on the Data as it gave better results after doing it(using StandardScaler()). We then tried out various models , starting with the basic ones finding the results and then slowly moving on to other models . We Started with like LogisticRegression , Naive Bayes ,etc. then

went to SVM , kNN, etc. then some tree models , like Decision Tree and Random Forest , and finally Ensemble methods like bragging , boosting , etc. was also used by us.Also , on seeing the Part(3) Results we can tell that none of the Models gave a good result for CO:6 , but still going for the best of those we had to use Ada-Boost for it.