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

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

Roll No: ED21S001,EE20S051

Name: ROHAN PADHY,KUMARI RASHMI

- 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.)]:

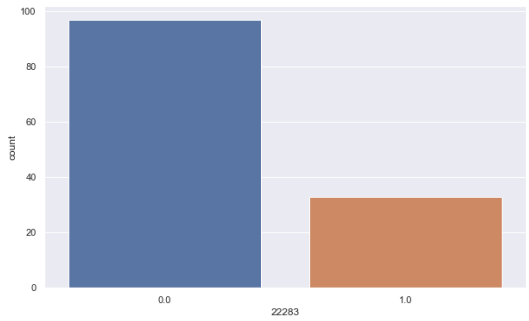
Solution:
The following are the models selected for our data-contest and their respective paradigms are also mentioned un the table below.

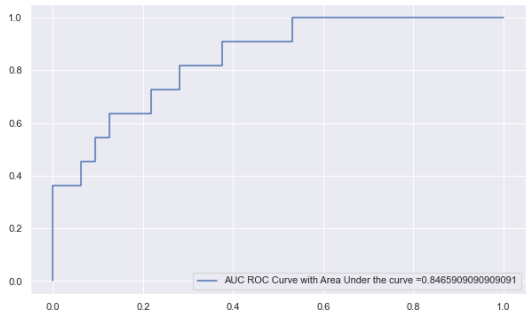
DECRIPTOR	MODEL	PARADIGM
CO:1	Logistic regression	Linear
CO:2	Random forest	Non-Linear
CO:3	Logistic regression	Linear
CO:4	Random forest	Non-Linear
CO:5	Adaboost	Non-Linear
CO:6	Random forest	Non-Linear

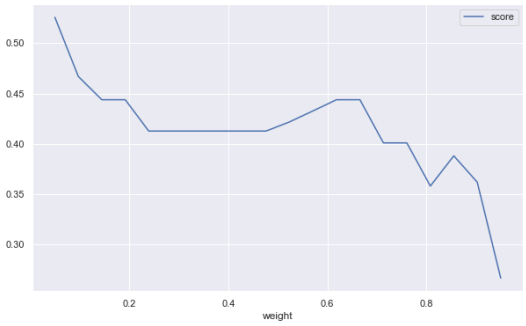
2. (points) [Brief description on the dataset: could show graphs illustrating data distribution or brief any additional analysis/data augmentation/data exploration performed]

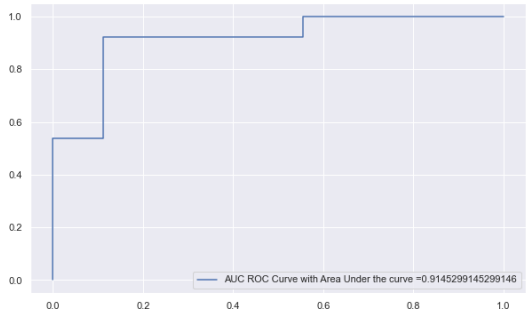
Solution:

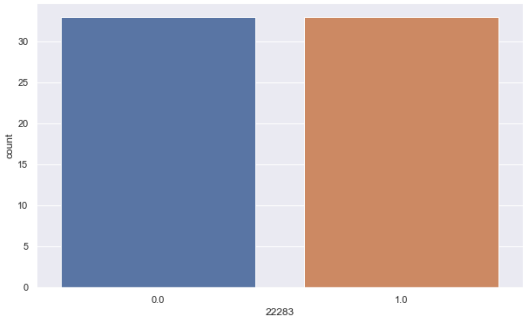
As the data set was imbalanced so i tried undersampling to balance out the outputs and also the area under the ROC curves increased from 84.65 to 85.51 but as during that time other libraries were not allowed so i used a manual code by implementing of my own.,still i didnot get any improvements.







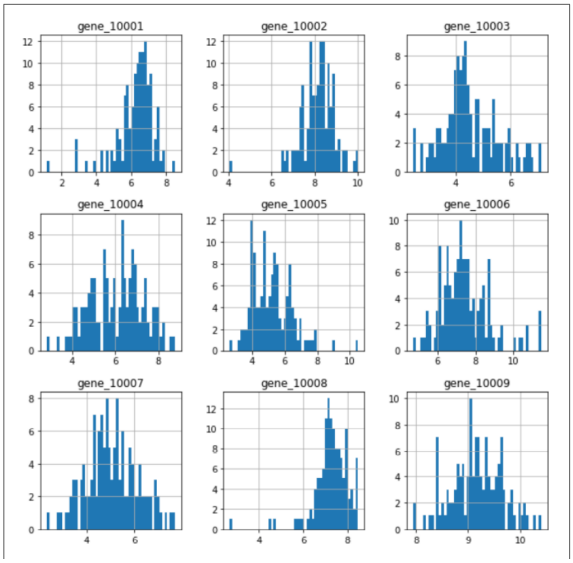


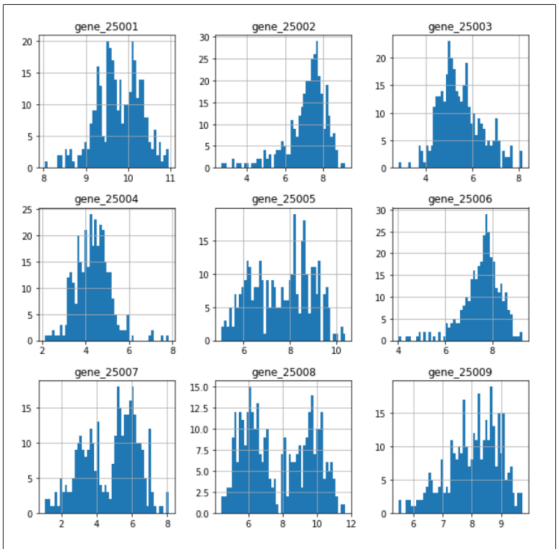


Dataset 1 contains 22283 genes as features and 2 clinical descriptors 'CO: 1' and 'CO: 2'. It contains 130 training samples and 100 test samples. While visualizing the training set we can observe that in general, genes approximately follow a normal distribution. We used StandardScaler to normalize the data where we fitted the scaler on Training dataset and further transformed both the training and the test datasets.

On the other hand, Dataset 2 contains 54675 genes as features and 4 clinical descriptors 'CO: 3','CO: 4','CO: 5' and 'CO: 6'. It contains 340 training samples and 214 test samples. While the genes in dataset 1 approximately followed a normal distribution, several genes in dataset 2

had multi-modal distributions. Both the datasets are complete, without any missing or 'NAN' values, and hence we need not do any data cleaning to compensate for the missing values





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:

MODEL	CO:1	CO:2	CO:3	CO:4	CO:5	CO:6
Logistic regression	0.752	-	0.856	0.259	0.658	0.062
SVM	0.463	0.459	0.325	0.259	0.639	0.040
Adaboost	0.752	-	-	-	0.892	-
Random forest	-	–0.636	0.414	0.309	0.677	0.241

- Initially, we adopted the same model for all the six clinical descriptors to arrive at a baseline model. We used SVM , Logistic regression and random forest to predict the descriptors and the MCC score was not better, since all the six of them used the same model.
- Then we tried to visualize the training data better to realize that certain descriptors were very different from the others. Later we tried a plethora of classification models for each of the descriptors with KFoldValidation and the mean values of the MatthewsCorrelationCoefficient(MCC) of the same can be seen from the above table of the methods which we have tried.

- Also we tried Linear discriminant analysis and Quadratic discriminant analysis but their MCC scores were not as expected.
- So we tried logistic-regression for CO:1 and CO:3,Adaboost for CO:5 and Random forest for CO:2 , CO:3 , CO:6 .We came to this method by applying random forest in all the classifiers and then replaced some with adaboost and Logistic-regression.
- Then we performed hypertuning on the models and obtained the desired MCC score as predicted by our output attached pythoncode

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:

- Because of the data's nature being from a clinical setting, there are thousands of features/attributes responsible for classifying the clinical descriptors. So, in this case it would be better to determine the significant genes by obtaining the features that are more important to the clinical descriptors, through the feature importances method of the RandomForestClassifier that we train.
- Also in the beginning we performed EDA and some feature engineering techniques.In the beginning we checked the null values to replace it as all the datas are coming from a clinical environment but nontheless we didn't find any null values.
- In the next step we performed PCA to reduce the redundant dimensins which were much corelated but PCA didnot improve our MCC score.
- Feature selection then came to our attention as the number of independent columns were high. we performed some methods like removing constant values , thresholding technique by using PEARSON'S CORELATION COEFFICIENT for classifier CO:1,but that didn't improved our score though but we got certain insights about our data.
- Finally we tried L-1 based feature selection which made us reach our output.L1 based feature selection is used for all descriptors.

- For CO:1, feature selection is done using Logistic regression and that classifier is used in model for prediction as well. For CO:2, feature selection is done using LinearSVC and Random Forest is used as classifier. For CO;3, feature selection is done using Logisic regression and Logisic regression is used as classifier. For CO:4, CO:5 and CO:6, classifier used in feature selection and classifier used for model both are same only

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

Solution:

DECRIPTOR	DIFFICULTY
CO:1	Easy
CO:2	Easy
CO:3	Easy
CO:4	DIFFICULT
CO:5	Easy
CO:6	DIFFICULT

The descriptors from the first dataset gave a consistent performance with more than 80 percentage accuracy.The same was observed with CO:3 and CO:5 as the dataset was schewed. So, specially we applied undersampling techniques to balance the data.Especially we had to apply brute-force in CO:6 by applying in all classifiers and also hypertuning them.

Then we referred the following papers to do feature selection(L-1 based selection):

1."Boosting for tumor classification with gene expression data" by M. Dettling
2. "Support vector machine based feature selection for land cover classification" by M.Pal
3. "High Dimensional data classification" by Vijay Pappu and M.Pardalos.

We were then able to visualise features independently but according to our understanding it was not helpful to select the perfect model but it gave initial start for understanding the behaviour of the data.Since data was high dimensional, so first we applied feature selection then applied these classifiers to get the predictions

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:
After trying the ensemble methods : boosting algorithms such as Xg-boost,Ada-boost given above as well as bagging algorithms such as random forest classifier, it became clear that both bagging and boosting was much more suitable to make predictions in this particular problem.

The models using RandomForestClassifier required extensive Hyper parameter tuning using different parameters like max depth,leaf nodes,n estimators,max features and the SVM kernel required kernel para-metrics, also in the case of logistic regression the solver values and "C" posed a certain difficulty.
Also as the data was high dimensional data visualization was pretty difficult, so we had to spend a majority of our time and effort by doing EDA (finding co-relations,checking null values) and dropping certain features by using Feature engineering techniques.

Initially we tried hyper parameter optimisation by using Random search CV and then to do the through analysis we shifted to Grid search CV., afterwards we implemented both which saved our time as Random search CV maximises the probability of the location of best parameters.