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

Release Date: 22 Oct, 2021
Due Date: 10 Nov, 2021

Roll No: CL21M015,CL21M004

Name: Different combinations of ensemble

- 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.

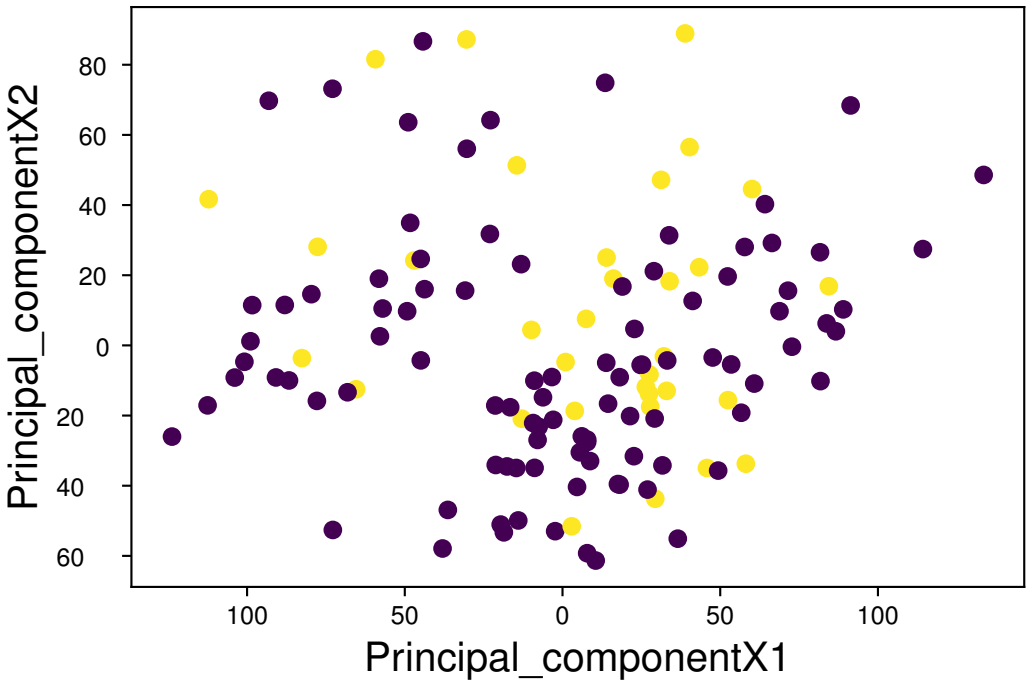
1. (points) [Name of the Method chosen for each descriptor, and its Paradigm: paradigm could be linear/non-linear models, etc.)]:

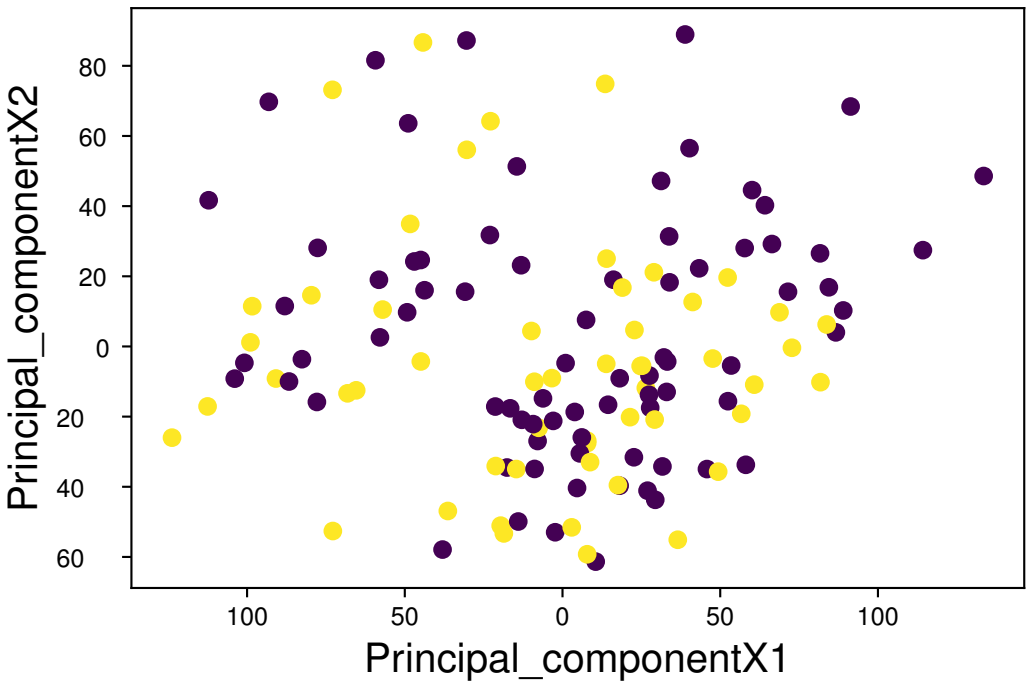
Dataset	Descriptor	Method	Paradigm
Dataset-1	Descriptor-1	Bagging Classifier	Non-linear
Dataset-1	Descriptor-2	Bagging Classifier	Non-linear
Dataset-2	Descriptor-3	Bagging Classifier	Non-linear
Dataset-2	Descriptor-4	Bagging Classifier	Non-linear
Dataset-2	Descriptor-5	Bagging Classifier	Non-linear
Dataset-2	Descriptor-6	AdaBoost Classifier	Non-linear

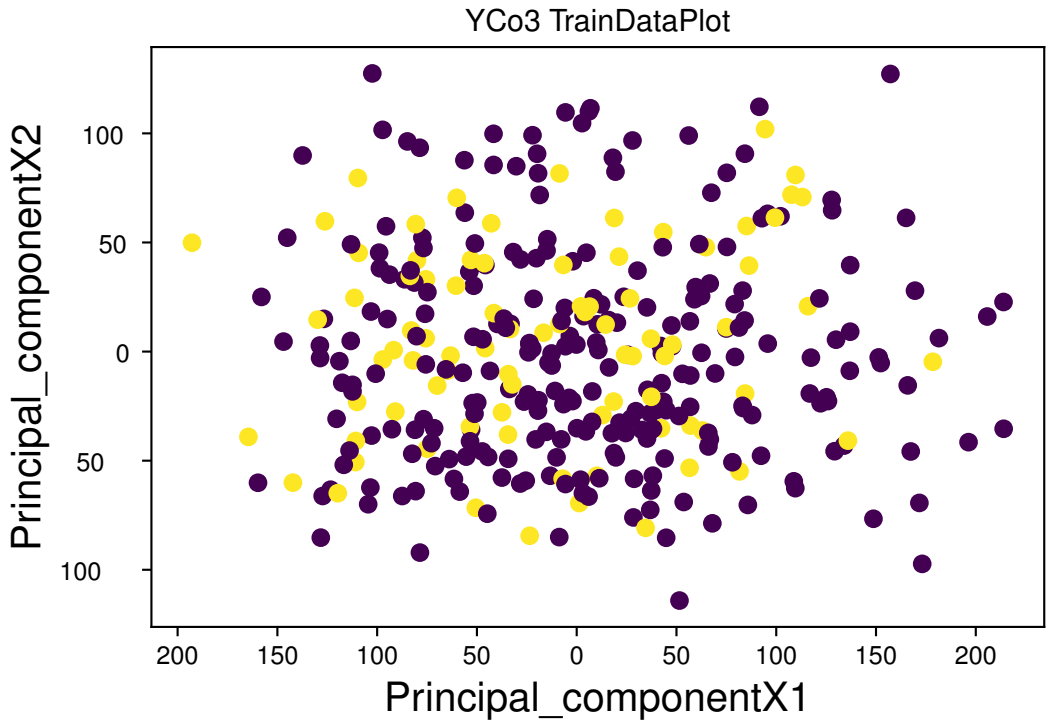
   The descriptor 6 uses Adaboost classifier with SVM as weak classifer and Kernel used for SVM is 'RBF'.

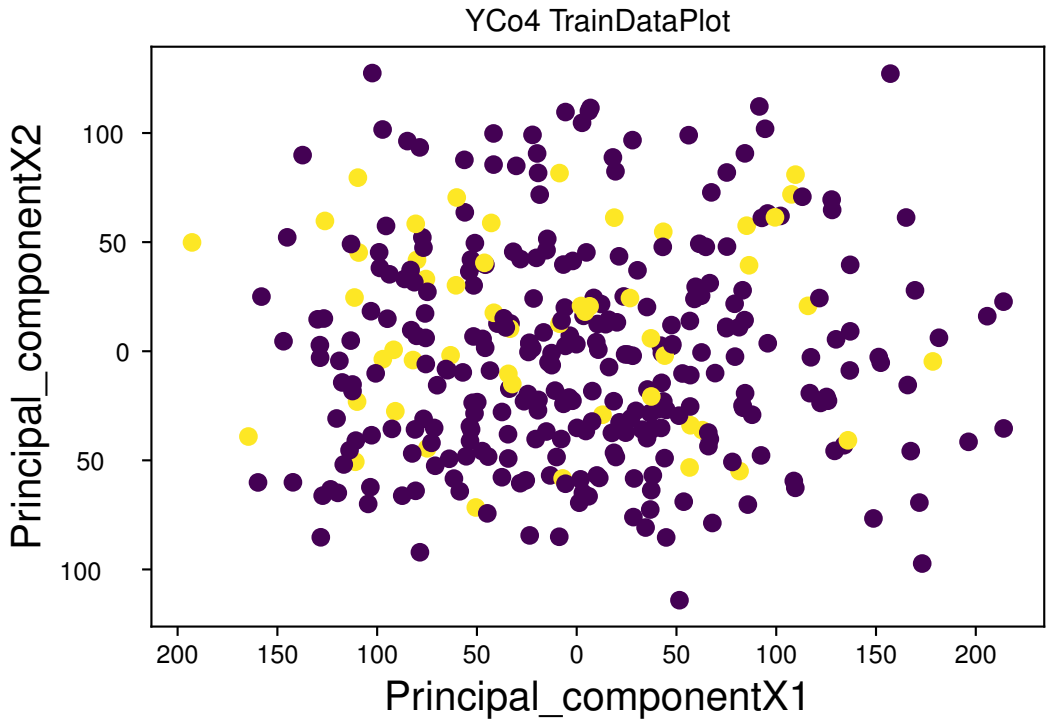
2. (points) [Brief description on the dataset: could show graphs illustrating data distribution or brief any additional analysis/data augmentation/data exploration performed]

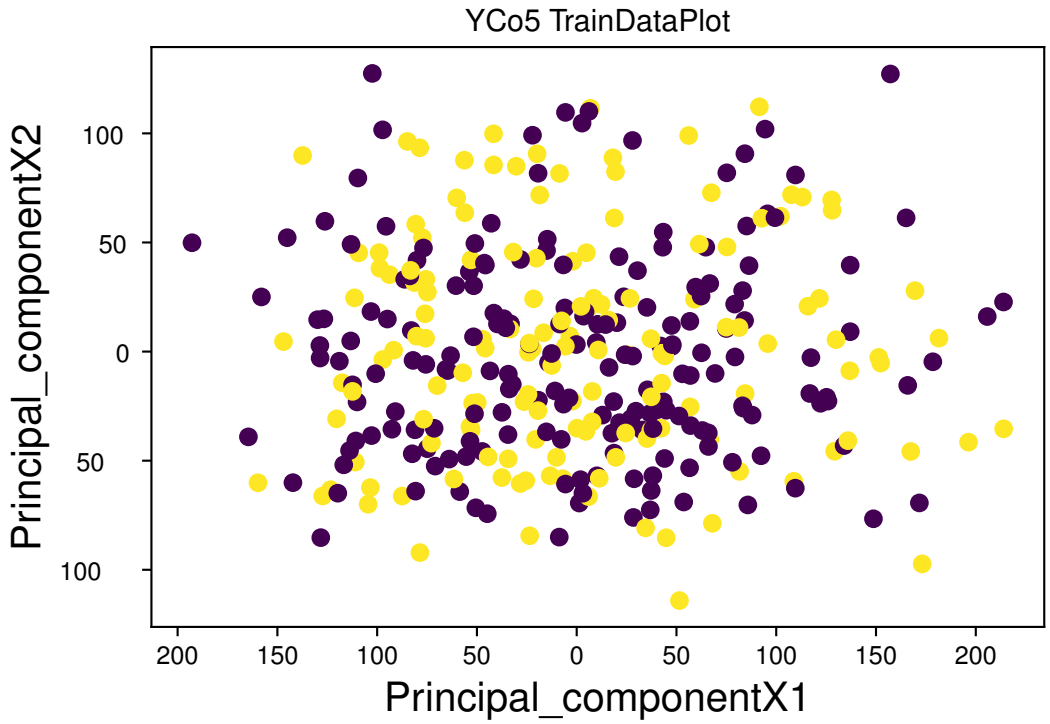
Solution:

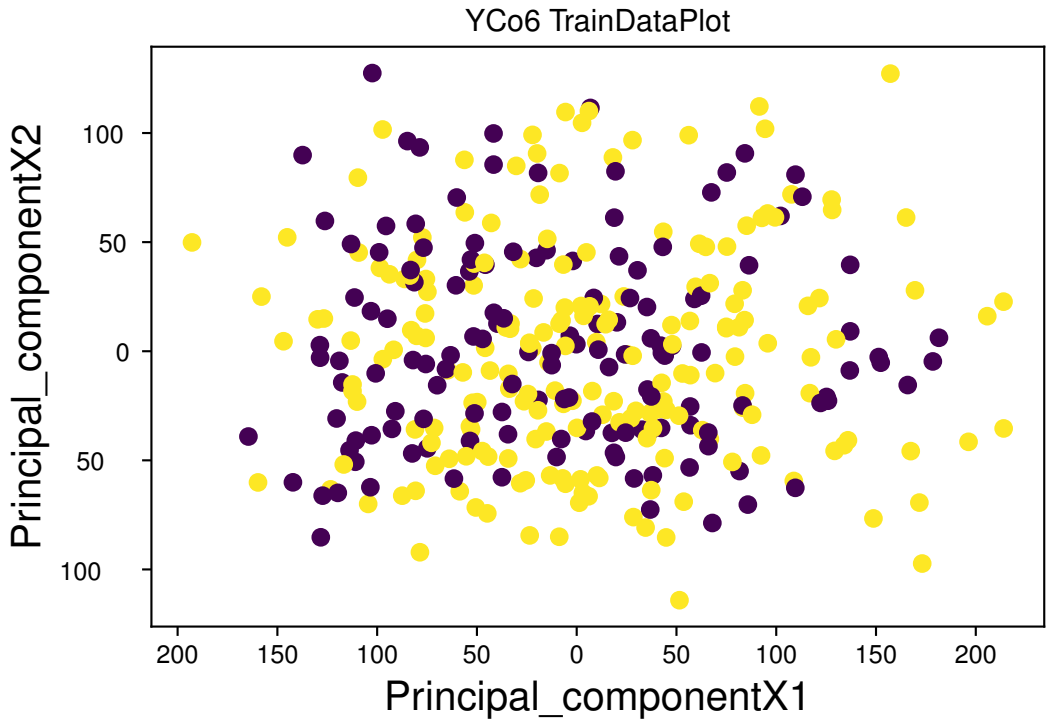

Principal_componentX2
80
60
40
20
0
20
40
60
100
50
0
50
100
Principal_componentX1


Principal_componentX2
80
60
40
20
0
20
40
60
100
50
0
50
100
Principal_componentX1


YCo3 TrainDataPlot
Principal_componentX2
100
50
0
50
100
200
150
100
50
0
50
100
150
200
Principal_componentX1


YCo4 TrainDataPlot
Principal_componentX2
100
50
0
50
100
200
150
100
50
0
50
100
150
200
Principal_componentX1


YCo5 TrainDataPlot
Principal_componentX2
100
50
0
50
100
200
150
100
50
0
50
100
150
200
Principal_componentX1


YCo6 TrainDataPlot
Principal_componentX2
100
50
0
50
100
200
150
100
50
0
50
100
150
200
Principal_componentX1

The plots show class(1) with yellow data points and the class(0) with blue data points.
There is large imbalance in data in yCo4 with class '1' occurring only 0.15 times in whole data.
Above plots are obtained after doing principal component analysis into 2 components.
This was done to understand the data spread among both classes.
number of 1's and 0's in yco1=33 and 67
number of 1's and 0's in yco2=53 and 47
number of 1's and 0's in yco3=83 and 257
number of 1's and 0's in yco4=51 and 289
number of 1's and 0's in yco5=146 and 194
number of 1's and 0's in yco6=200 and 140

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]

   - out of 6 descriptors,1st,2nd,3rd,5th and 6th have somewhat balanced classes. only 4th descriptor has class imbalance as only 15 percentage of data belong to one of the class which can bias our model.
   - So to choose which method is best for each descriptors we used k-fold cross validation score, accuracy scores for the models for class-balanced descriptors and F-1 score metrics for class imbalanced descriptor.
   - Given tables give a comparative performance of models for each descriptor for k-fold cross validation calculation , we have considered k=5 then average of cross validation scores for all the folds are calculated.

Descriptor	GaussianNB	LogiRegression	RandForest	SVM	GradBoost	Bagging	AdaBoost
Descriptor1	0.753	0.738	0.761	0.746	0.707	0.79	0.77
Descriptor2	0.707	0.669	0.669	0.592	0.669	0.727	0.720
Descriptor3	0.652	0.744	0.773	0.755	0.747	0.82	0.773
Descriptor4	0.726	0.844	0.85	0.85	0.805	0.883	0.85
Descriptor5	0.582	0.791	0.794	0.570	0.920	0.932	0.894
Descriptor6	0.511	0.538	0.561	0.588	0.55	0.598	0.642

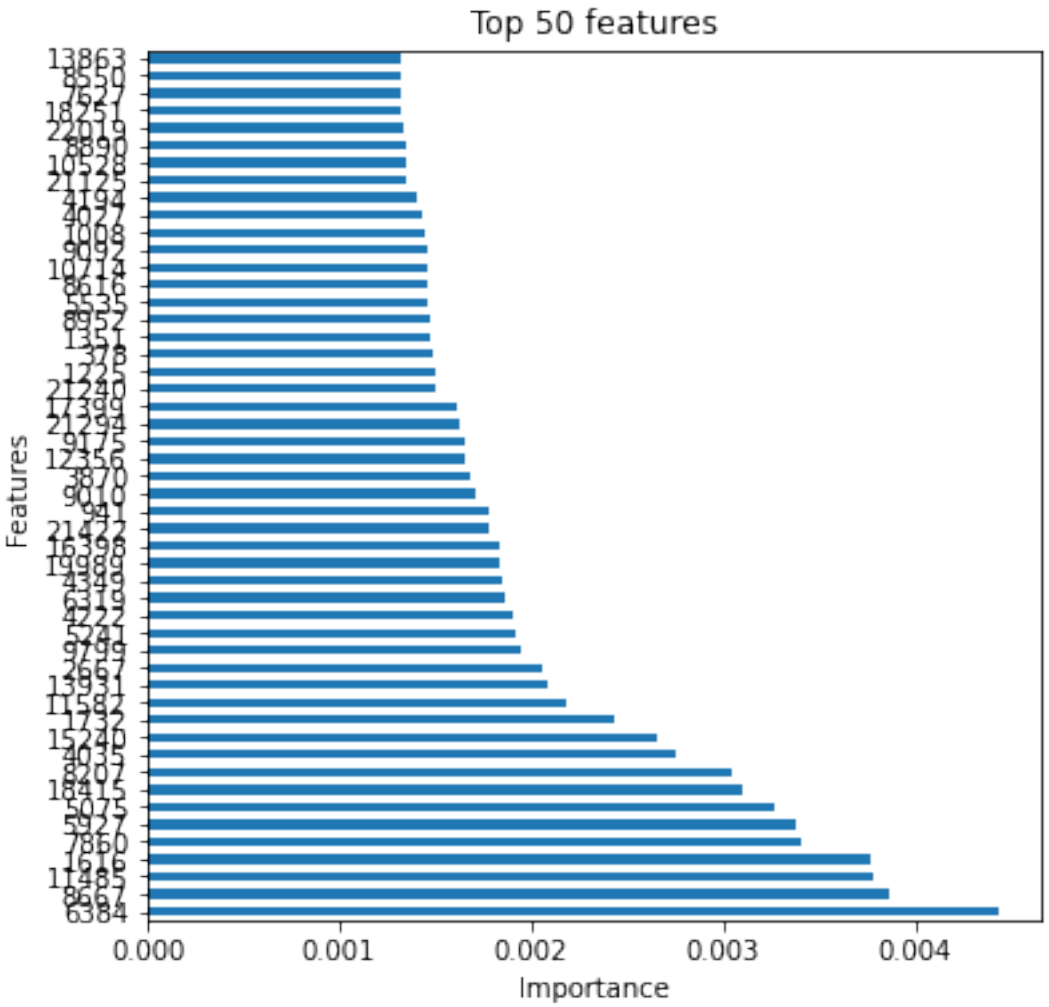
Model evaluation metrics "Accuracy" is used for all the descriptors with balanced classes and F1 score mterics is used for class imbalanced descriptor.

Descriptor	GaussianNB	RandForest	SVC	GradBoost	AdaBoost	Bagging	LogiRegression
Descriptor1	0.820	0.846	0.820	0.743	0.871	0.887	0.743
Descriptor2	0.589	0.589	0.564	0.692	0.615	0.703	0.641
Descriptor3	0.735	0.784	0.774	0.784	0.774	0.83	0.784
Descriptor5	0.662	0.734	0.764	0.570	0.575	0.845	0.674
Descriptor6	0.551	0.553	0.602	0.584	0.645	0.608	0.592

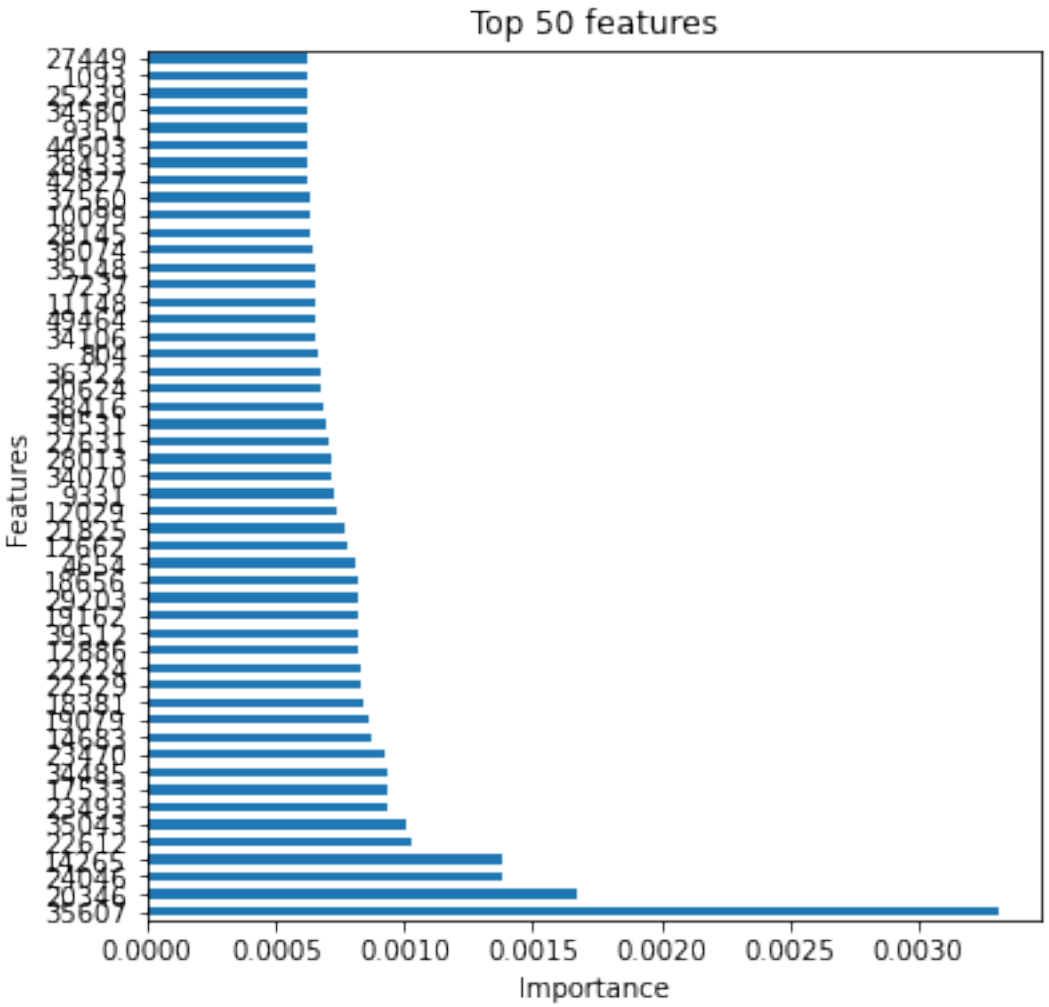
Descriptor	GaussianNB	RandForest	SVC	GradBoost	AdaBoost	Bagging	LogiRegression
Descriptor4	0.89	0.88	0.90	0.90	0.92	0.92	0.88

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.]:

   - Model interpretation is important to trust the model. and we always do a trade off between model interpretability and performance. Global interpretation of the model can be observed using feature importance. feature importance tells, how much a feature is important for final prediction.
   - Using sklearn libraries top 50 features for Train dataset1 and descriptor1 were plotted for Extremely randomized tree classifier model.



- Now the below figure shows top 50 features for train dataset2 and descriptor3 combination for same Extremely randomized tree classifier model.



- Now from these figures we get an idea about top 50 feature's individual imporatnce for final prediction. we can top 100 or top 5000 features also. Then we can negelect the less important features which eventually make our model training slow and less accurate. using these best features for training of our model we will be able to get good accuracy for final test data.
- Another method we have partial dependence plot. this PDP plot show the relationship between one or two input features and output.all the data points are forced to have same feature values to get the PDP estimates. Local interpret-ability can be observed using LIME(local interpret-able model-agnostic explanation).Again ELI5 python tool kit can also be used to know how individual features are important for the final prediction.

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

Solution:

- The endpoints Yco1 and Yco2 were easy to predict
- The endpoint Yc03 and Yc04 were difficult to predict due to the class imbalance.
- The endpoint Yco5 was also difficult to predict as for majority of models the targets were biased to class '0'.
- The endpoint Yco6 was easy 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:

- The challenges faced during developing the model are as following:
- The data interpretation was the foremost challenge we faced.
- Then while developing the models tuning the hyper parameters was challenging us. computationally. As it was consuming large time and memory.
- In many of the training for predictors we were given imbalanced data due to which the target variables predicted were biased towards a single class.
- One of Strategy we used to better our model was to check the MCC coefficient by varying the model of a single predictor.
- We also tried calculating the recall and precision and accuracy with different models. We split our train data into 30-70 and calculated the above.
- And choosing the base estimator in the adaboost to get better model was challenging.
- Multiple weak classifier were used for the adabosst like KNN, SVM with different kernels like RBF,poly, Decision stump. And based on metrics we finalised with SVM with RBF kernel.
- We tried to balance the data using imblearn package and then did under-sampling for Co4 predictor and SMOTE oversampling for Co5 predictor but both the attempts were not successful.