Predicting Hospital Patients' Duration of Stay, Possibility of Having Heart Failure, Glucose Levels, Hemoglobin Levels, and Coronary Artery Disease

Bashar Jirjees, Bipasa Agrawal, Dylan Kemp
University of Victoria
Computer Science Department

Abstract

This project consists of the ability to predict hospital patient residence periods, the possibility of having heart failure or coronary artery disease, and the ability to predict glucose and hemoglobin levels.
Different features were used in each prediction like blood tests of patients like platelets and creatinine levels, activities like smoking and alcohol, and diseases/symptoms like anemia and chest pain. This project tackles the possibility of predicting crucial health data new hospital patients might need to progress in their treatments through the use of various machine learning algorithms for accuracy check-ups and what features are most responsible for predicting labels.

1. Introduction

1.1 Data Description

The data consists of a CSV file containing patients' info when admitted to Hero DMC Heart Institute in India. This info consists of patients' blood tests, a few habits/activities, the original community, and diseases/conditions.
The sample consists of 15757 patients whose data was recorded across a period of 2 years (from 1st of April 2017 to 31st March 2019); furthermore, the total number of admission contains a majority of new/distinct patients of about 12238 minimally.
The patients' blood tests and age consist of positive integers, diseases consist of 0s and 1s which only indicates if a patient has the specific illness without illustrating its stage, patients' date of admission, date of discharge, gender, type of admission, and outcomes are all string based. Empty cells within all the features in the dataset are all illustrated with an "EMPTY" string.

	date of admission	diabetes mellitus	hypertension	coronary artery disease	prior metabolic panel	chronic kidney disease	ejection fraction	leucocyte count	brain natriuretic peptide	myocardial infarction	...	kidney injury	cerebrovascular accident	cerebrovascular accident
0	4/3/2017	1	0	0	0	0	35	16.1	1880	0	...	0	0	0
1	4/5/2017	0	1	1	0	0	42	9	0	0	...	0	0	0
2	4/3/2017	1	0	1	0	0	0	14.7	210	0	...	1	0	0
3	4/8/2017	0	1	1	0	0	42	9.9	0	0	...	0	0	0
4	4/23/2017	0	1	0	1	0	16	9.1	1840	0	...	0	0	0

	date of admission	age	gender	rural	type of admission-emergency/opd	duration of stay	duration of intensive unit stay	outcome	smoking	alcohol	...	paroxysmal tachycardia	congenital	urinary tract infections	neuro cardiogenic syncope
0	4/3/2017	81	M	R	E	3	2	DISCHARGE	0	0	...	0	0	0	0
1	4/5/2017	65	M	R	E	5	2	DISCHARGE	0	1	...	0	0	0	0
2	4/3/2017	53	M	U	E	3	3	DISCHARGE	0	0	...	0	0	0	0
3	4/8/2017	67	F	U	E	8	6	DISCHARGE	0	0	...	0	0	0	0
4	4/23/2017	60	F	U	E	23	9	DISCHARGE	0	0	...	0	0	0	0

As we can see from the snippets of the datasets above how blood tests like brain natriuretic peptide, leucocyte count, and age are represented via positive integers; furthermore, diseases, gender, rural, and diseases (congenital, kidney injury, etc..) categorical features are represented via strings and 0s and 1s.

1.2 Previous and Current Work

The data was harnessed by the doctors at Hero DMC Heart Institute to distinguish whether patients have cardiogenic shocks based on blood pressure and current heart diseases they are suffering from [1][2].

Majorly, this data was also used to efficiently predict the mortality rate of patients with severe heart problems, classify patients to predict if they have pulmonary embolisms, and predict if patients could have a heart attack due to a blocked artery based on their current health status[1][2]. The research took place at Massachusetts General Hospital and Institute of Technology, Indian Institute of Technology Ropar and Indian Hero DMC Heart Institute[1][2].

Our current work will harness the data to see how efficient it is in predicting the patients' duration of hospital stay so the hospital staff can measure the needed resources for each patient, predict whether patients can be diagnosed with heart failure, see how efficiently we can predict glucose and hemoglobin levels of patients based on their current overall health conditions and provide insight on what factors could mostly boost the prediction process.

We have also checked if it is possible to predict the type of admission each patient needs and what diseases and symptoms could help mostly in predicting it, this could significantly eradicate misdiagnosis of severely ill patients and for them to receive the correct medical help as soon as in case of emergencies.

Finally, we checked if it is possible to predict if a patient could have coronary artery disease based on their current health conditions to further help in the diagnosis process at admission and for the patient to receive the most efficient treatment possible.

2. Approach

Most numerical/continuous features of hemoglobin levels, leucocyte count, platelets, glucose, urea, creatinine, brain natriuretic peptide, heart ejection fraction, and duration of intensive unit stay were normalized using min and max values to remove data impurities as much as possible and to create one range of the data from 0 to 1 (uniformed). The algorithm below was used which is also known as MinMaxScaler:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

The only continuous features that weren't normalized are age and total duration of stay to achieve due to concerns that the research could grow and we need the values predicted in their true/original form without manipulation.

String categorical features like gender were converted to numerical/continuous features using Label Encoding where each string feature is represented by a number that is discreetly linked to it and is between 0 and the total number of classes in that feature. The example below illustrates this process:

Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

Label Encoded Data

Team	Points
0	25
0	12
1	15
1	14
1	19
1	23
2	25
2	29

All cells with no data or with the "EMPTY" keyword were replaced by a zero value.

3. Analysis and Results

3.1 Duration of Stay

We Chose:

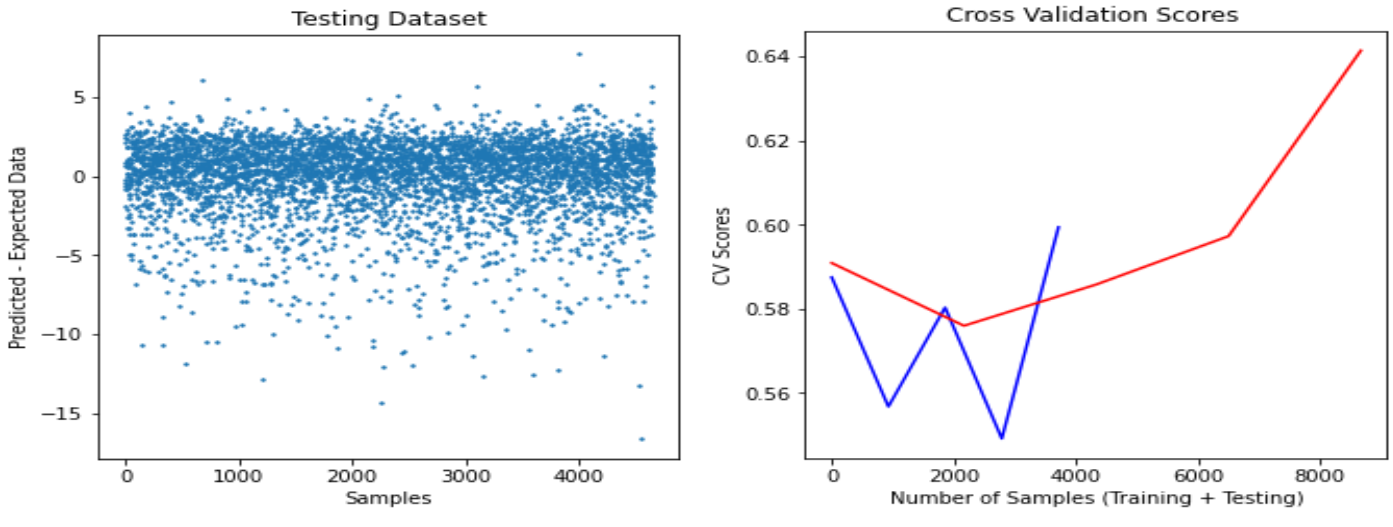
1. Dummy Regressor as we are predicting continuous data. The regressor will predict labels using the mean and median of the labeled dataset without taking into consideration the features used. This provides insight into how far the predictions are from the mean/median and asserts the imbalanced variations between the label points.

2. Linear Regression to fit a line through the values of the features used and determines the dependencies between the values used to provide an efficient prediction based on the final best fit-line equation.

3. Random Forest Regressor to Random Forest Regressor to classify the data based on votes done by multiple decision trees where it classifies the most appropriate prediction by taking the average of all predictions by the generated stumps based on the given example/feature.

After removing the label data outliers and providing an efficient number of features to minimize overfitting. The error margins were reduced significantly but they are still unacceptable. The graphs and data below illustrate the results achieved by Random Forest Regressor which is the best-tested model:

Training Error Margin	Testing Error Margin	Training CV core	Testing CV Score
3.899	5.707	0.598	0.574



Note: For the CV score the Red line is training data and the Blue line is testing data.

As illustrated in the code, the final features used to consist of blood tests and severe disorders, chronic diseases, severe heart conditions, duration of EU stay, age, and gender as providing worse health conditions will help the model in predicting a worst-case scenario for patients. It is worth noting that the “duration of intensive unit stay” was mostly the detriment factor of the prediction.

3.2 Heart Failure

We Chose:

1. Random Forest Classifier as mainly the predicted label is categorical and conversion to 0s and 1s with a threshold is not needed; furthermore, it is also vote-based between multiple checked subtrees which emphasizes the accuracy of predictions and minimizes weak stumps/predictors.

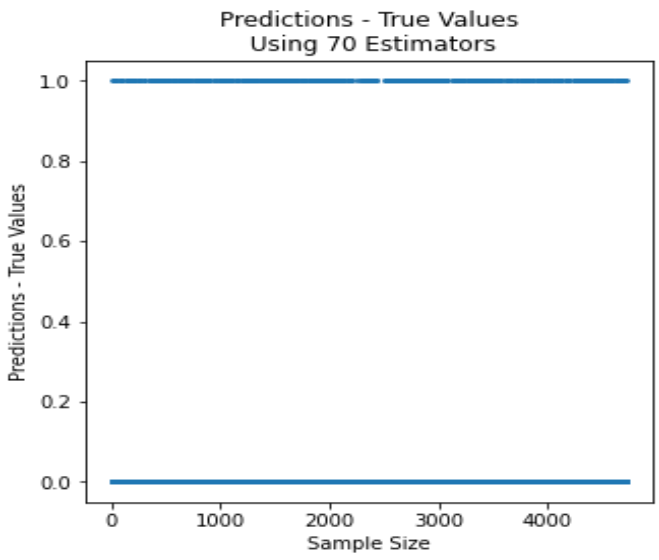
2. AdaBoost Classifier is an improved model of Random Forest Classifier where unneeded features or weak learners are eradicated/pruned from the forest using low weighted values which leave the algorithm to decide using only strong learners based on voting too; however, overfitting can occur if excessive pruning occurs and strong learners are eradicated.

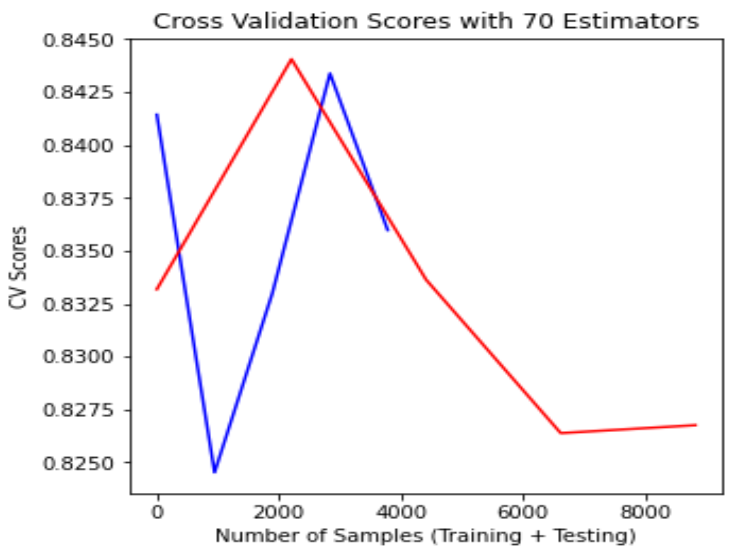
3. Desne Neural networks as it eradicates the weak outputs based on the given example through a total of 4 hidden layers with a total of 400 nodes that choose a random weighted values where in the end, the decided output is the one with the highest weight.

Dropout and regularizations are added to have a clearer range of data and drop unnecessary low values at the output from each hidden layer.

Random forest classifier performed the best regarding computational time and CV scores which are illustrated below:

Training Error Margin	Testing Error Margin	Training CV core	Testing CV Score
0.132	0.162	0.832	0.835





As illustrated in the code, the final features used consist of provided blood tests and severe disorders, chronic diseases, severe heart conditions, blood clots, lung failure, age, stroke, generated substances and enzymes, patients' activities, and gender to help the model in predicting with minimal overfitting, and also considering a worst case scenario regarding patients' health. It is worth noting that substances generated by the heart called "brain natriuretic peptide" was the strongest predictor among all the other used features.

3.3 Glucose

We Chose:

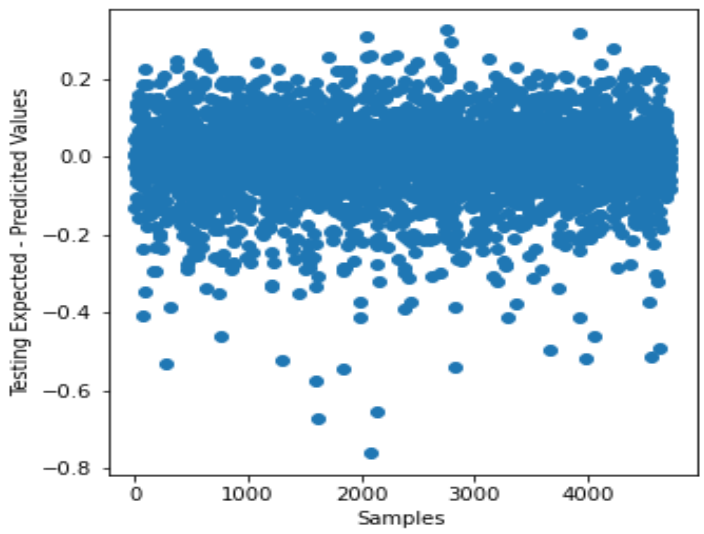
1. Random Forest Regressor to classify the data based on votes done by multiple decision trees where it classifies the most appropriate prediction by taking the average of all predictions by the generated stumps based on the given example/features.

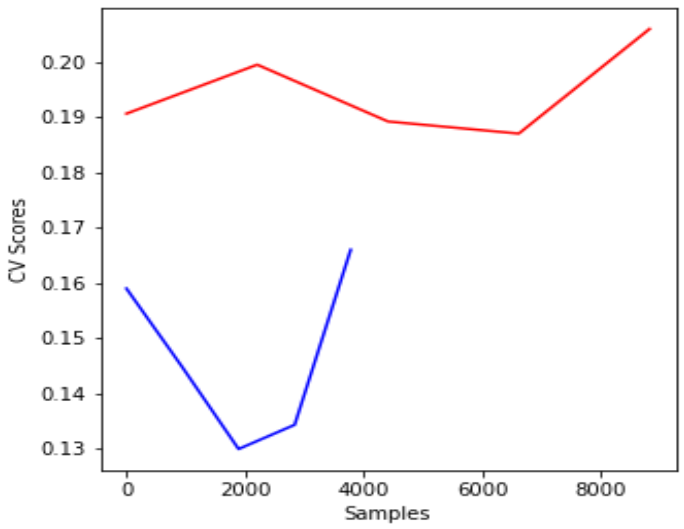
2. Linear, RBF, and polynomial Support Vector Regression as it separates similar in-range values from each other using scalable gamma values, a coefficient that guarantees maximal results. For linear and Polynomial SVR a line/curve passes between the points or examples values that generate minimal error margins. For RBF SVR the points are separated by projecting them into a higher hyperplane for clearer separation of points and labels.

3. Logistic XGBoost where it minimizes the error made by each prior prediction based on the provided example by deleting the previous margin of error from the current prediction and using the result as input for the next recursive call. The built-in Logistic regression is used by measuring the relationship between the features and calculating the probabilities of each label to occur which eases the process when predicting approximate values after fitting the model.

Unfortunately, we weren't able to efficiently predict patients' glucose levels without severe overfitting based on the given data and further research is required; however, XGBoost still provided the most efficient results regarding CV scores and error margin. XGBoost results are illustrated below:

Training Error Margin	Testing Error Margin	Training CV core	Testing CV Score
0.004	0.008	0.194 (Severe Overfitting)	0.146 (Severe Overfitting)





As illustrated in the code, the final features used consist of provided blood tests and severe disorders, chronic diseases, severe heart conditions, blood clots, lung failure age, stroke, generated substances and enzymes, patients' activities, blood flow to the brain, hypertension, blood infections, body use of sugar and gender to help the model in predicting with minimal overfitting, and also considering a worst case scenario regarding patients' health. It is worth noting that "diabetes mellitus" disease was the strongest predictor among all the other used features. All these features are related to high glucose levels that damage the body's arteries, increase inflammation, decrease blood flow, and are a main cause of strokes.

3.4 Hemoglobin

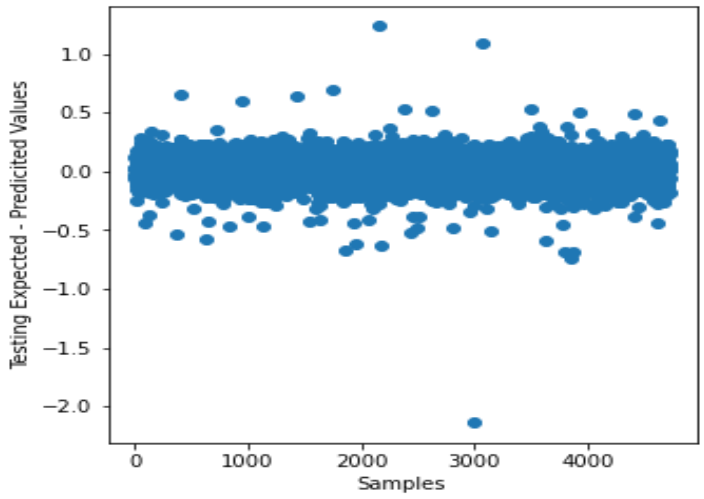
We Chose:

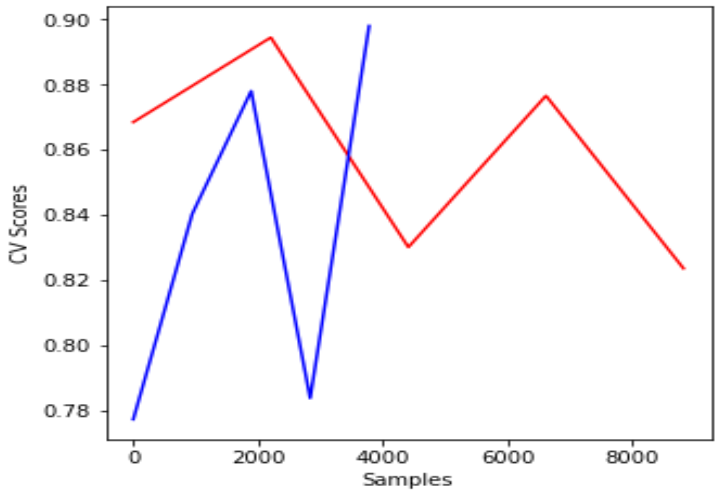
1. Random Forest Regressor to classify the data based on votes done by multiple decision trees where it classifies the most appropriate prediction by taking the average of all predictions by the generated stumps based on the given example/features.

2. Logistic XGBoost Where it minimizes the error made by each prior prediction based on the provided example by deleting the previous margin of error from the current prediction and using the result as input for the next recursive call. The built-in Logistic regression is used by measuring the relationship between the features and calculating the probabilities of each label to occur which eases the process when predicting approximate values after fitting the model.

Logistic XGBoost performed the best with an average CV score of approximately 0.86 for training data and 0.84 for the testing data after using log value transformation of blood tests so further check-ups weren't necessary. The results graphs and data are illustrated below:

Training Error Margin	Testing Error Margin	Training CV Score	Testing CV Score
0.0148	0.006	0.858	0.835





As illustrated in the code, the final features used consist of provided blood tests and severe disorders, chronic diseases, severe heart conditions, blood clots, lung failure age, stroke, generated substances and enzymes, patients' activities, blood flow to the brain, hypertension, blood infections, body use of sugar, age and gender to help the model in predicting with minimal overfitting, and also considering a worst case scenario regarding patients' health. It is worth noting that "anemia" disease was the strongest predictor among all the other used features. All these features are related to blood flow and quality within the body in which hemoglobin levels can affect blood fluidity, oxygen reaching the heart and the brain, organ damage, and heart failure.

3.5 Patients Type of Admission

We Chose:

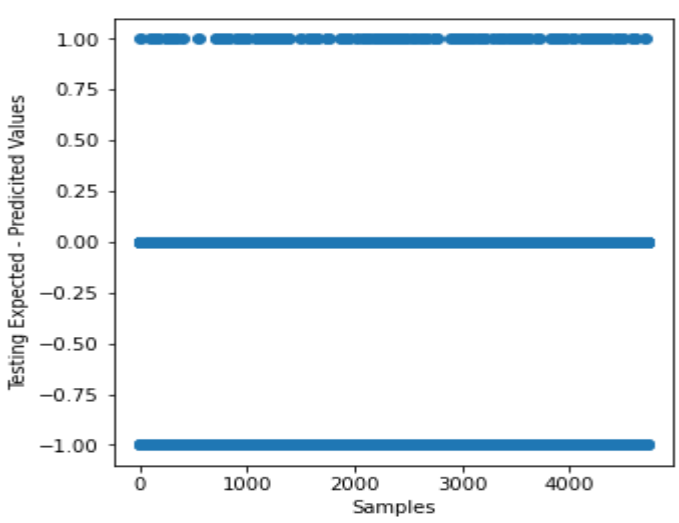
1. Support Vector Classifier as we are dealing with only 2 classes of either an emergency patient or an outpatient, the SVC algorithm will be able to separate the data based on the given features into a higher dimension and pass a plane between points that are likely linked to different labels. The data are separated using a specific coefficient and scalable gamma values that adjust the hyperplane size.

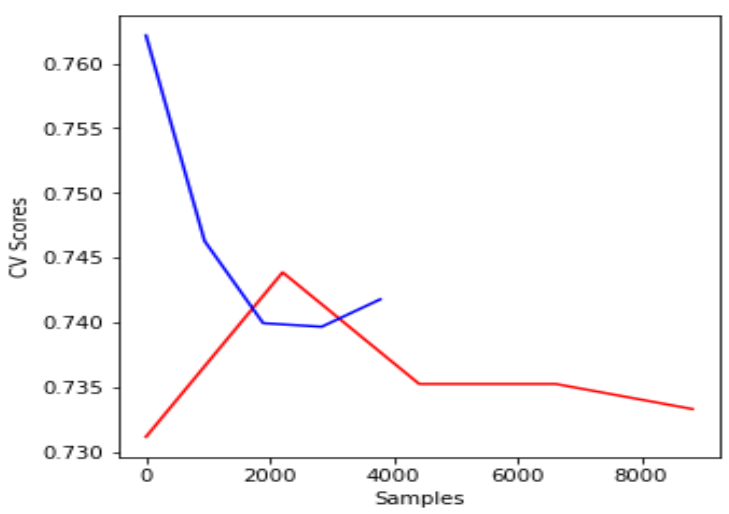
2. Binary Hinge XGBoost as our label data is only 0s and 1s, the algorithm can build future sub-models based on the error margin produced from the output of previous models and using the difference between this error margin and the current inputted value till the model reach the lowest achievable error margin. The hinge loss maximizes the accuracy of predictions by minimizing the variations between the probability of predicted outputs through the use of adjusted values that affirms the correctness of the prediction if both the value and the prediction have the same sign. This check-up adjusts the values depending on the expected data and examples inputted [3].

3. Gaussian Naive Bayes as we are dealing with only 2 classifications of 0s and 1s and we can link discrete values within the given features to each label so every tested future example can be linked to one class based on its provided values and how close they are to the data stored for each label during training.

Both SVC and Gaussian Naive Bayes performed the best here with almost identical error margins and CV scores; however, Gaussian Naive Bayes was more efficient due to significantly less computational time. Gaussian Naive Bayes results are illustrated below:

Training Error Margin	Testing Error Margin	Training CV core	Testing CV Score
0.264	0.254	0.735	0.745





As illustrated in the code, the final features used consist of only "stable angina", and "atypical chest pain" as they were the only features to provide any noticeable boost in CV scores and error margins. Both of these features link to less oxygen flowing to the heart which generates chest pain and it is also an indicator of underlying heart disease. This indicates that people with chest pain and low blood oxygen can be easily predicted to be outpatient or emergency patients.
It is worth noting that "stable angina" disease was the strongest predictor by being able to boost the CV score by approximately 0.7 which indicates that blood oxygen level is a very good indicator of what type of admission should the patient receive.

3.6 Coronary Artery Disease

We Chose:

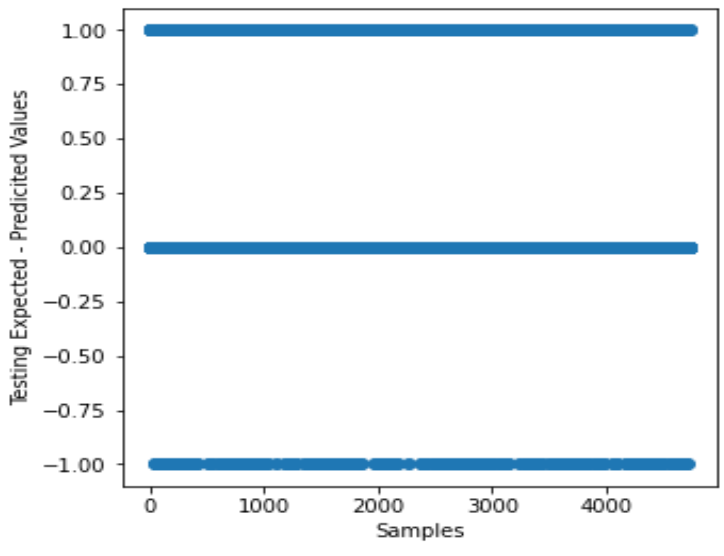
1. Dense Neural Network with 4 layers and a total of 400 inner nodes to minimize weak features and data using weights and considering the most dominant examples that could generate a trustable output. As we are predicting binary outputs too this makes it easier for the network to make guesses as the range of values is only 0 and 1. Regulation and dropout parameters are added to further minimize unneeded values and variations between examples' values.

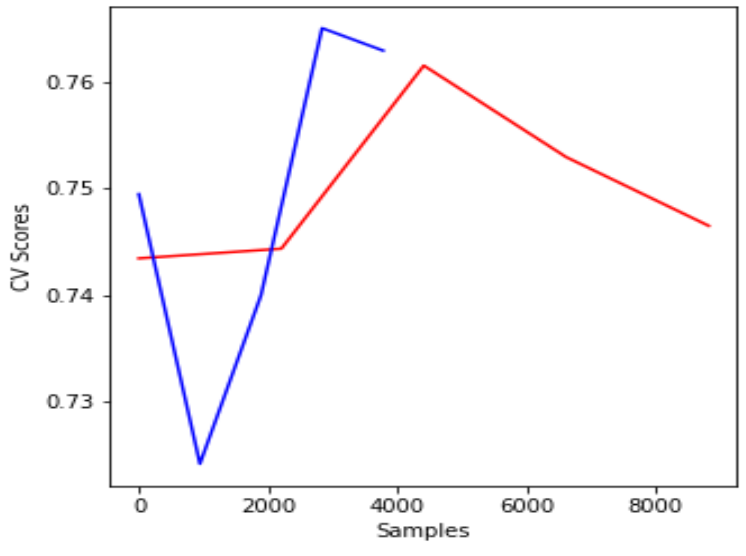
2. Random Forest Classifier as mainly the predicted label is categorical and conversion to 0s and 1s is not needed; furthermore, it is also vote-based between multiple checked stumps/subtrees which emphasizes the accuracy of predictions and minimizes weak predictors.

3. Binary Hinge XGBoost as our label data is only 0s and 1s, the algorithm can build future sub-models based on the error margin produced from the output of previous models and using the difference between this error margin and the current inputted value till the model reach the lowest achievable error margin. The hinge loss maximizes the accuracy of predictions by minimizing the variations between the probability of predicted outputs through the use of adjusted values that affirms the correctness of the prediction if both the value and the prediction have the same sign. This check-up adjusts the values depending on the expected data and examples inputted [3].

Random Forest Classifier is the best model for predicting if the patients have coronary artery disease or not by achieving the highest CV and accuracy values among all other tested models. The results are illustrated below:

Training Error Margin	**Testing Error Margin**	**Training CV core**	**Testing CV Score**
0.172	0.243	0.749	0.748





As illustrated in the code, the final features used consist of provided blood tests, severe disorders, chronic diseases, severe heart conditions, blood clots, lung failure age, stroke, heart-produced substances and enzymes, patients' activities, blood flow to the brain, hypertension, blood infections, body use of sugar, age, and gender. It is worth noting that "brain natriuretic peptide" disease was the strongest predictor among all the other used features separately. All these features are related to heart conditions and are a symptom or cause of heart problems as coronary artery diseases stop the blood from reaching the organs efficiently, causing them to fail and forcing the heart to pump more blood which also increases blood pressure and risk of having heart failure/conditions.

References

[1] A. Sahani, "Hospital Admissions Data," Kaggle, 21-Jan-2022. [Online]. Available: https://www.kaggle.com/datasets/ashishsahani/hospital-admissions-data. [Accessed: 13-Dec-2022].

[2] S. C. Bollepalli, A. K. Sahani, N. Aslam, B. Mohan, K. Kulkarni, A. Goyal, B. Singh, G. Singh, A. Mittal, R. Tandon, S. T. Chhabra, G. S. Wander, and A. A. Armoundas, "An optimized machine learning model accurately predicts in-hospital outcomes at admission to a Cardiac unit," Diagnostics, vol. 12, no. 2, p. 241, Jan. 2022.

[3] "Hinge loss," Wikipedia, 18-Nov-2022. [Online]. Available: https://en.wikipedia.org/wiki/Hinge_loss#:~:text=In%20machine%20learning%2C%20the%20hinge,support%20vector%20machines%20(SVMs). [Accessed: 13-Dec-2022].