ECG Heartbeat Analysis
BIOST 546 Final Project Report
Nayan Kaushal

1. **Overview:** I will be analyzing the dataset containing 12552 Electrocardiogram (ECG) signals of single heartbeats and predicting whether the readings are normal or abnormal. The dataset has been derived from The PTB Diagnostic ECG Database. The ECG signals have been decomposed into 187 vectors. Each of the vectors provides measurements at consecutive time points. Preprocessing of the signals has been performed by cropping, down sampling to sampling frequency of 125Hz, and padding with zeroes if necessary. The response variable is a categorical variable indicating whether the heartbeat is normal or abnormal (0: normal, 1 abnormal).

2. **Problem Statement:** To train a classification model that can predict whether the ECG signal of a heartbeat is normal or abnormal.

3. **Preliminary Data Analysis:**

   (i) Following is a snippet of the dataset to be analyzed:

Description: df [6 × 188]

	X1 <dbl>	X2 <dbl>	X3 <dbl>	X4 <dbl>	X5 <dbl>	X6 <dbl>	X7 <dbl>
1	1.0000000	0.9003242	0.358589947	0.051458672	0.04659643	0.126823336	0.13330632
2	1.0000000	0.7946815	0.375386506	0.116883114	0.00000000	0.171923310	0.28385898
3	0.9090289	0.7914821	0.423168659	0.186712101	0.00000000	0.007836456	0.06303237
5	1.0000000	0.8672383	0.201360136	0.099349499	0.14133649	0.120934360	0.10851567
6	0.9489833	0.5052651	0.004175744	0.022512708	0.05954975	0.107298478	0.11038490
8	1.0000000	0.4603807	0.122177958	0.009296149	0.12571934	0.220008850	0.26737493

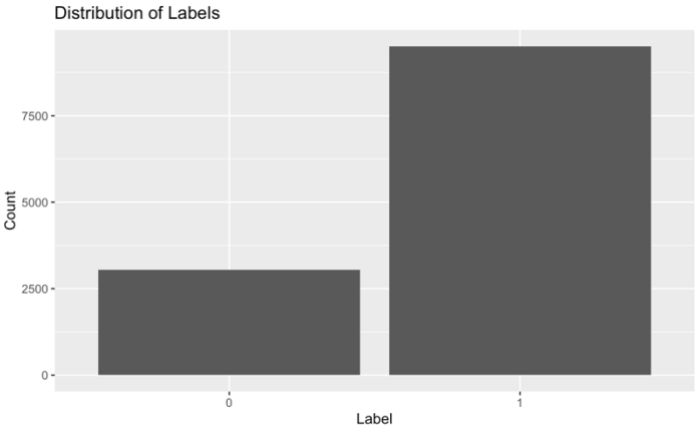
6 rows | 1-8 of 188 columns

   (ii) Size of the dataset:

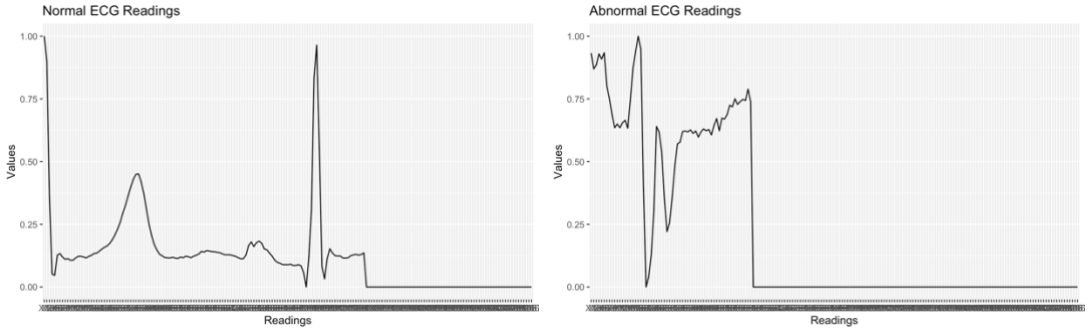
Sample size of the data, n	Number of predictor variables, p
12552	187

   (iii) Number of observations in each class:

Normal (class 0) = 3046
Abnormal (class 1) = 9506



4. **Exploratory Data Analysis:**
   To understand how normal ECG readings differed from abnormal ones, I visualized and compared the observations of two ECG signals in the dataset – one, which was labeled normal, and second labeled abnormal.



5. **Modeling:** The training dataset (train.csv) is further split into train and test data to check and understand the accuracy of the model before it can be used for predicting the unlabeled data (test.csv).

   Generalized Linear Model: I began my analysis by building a simple logistic regression model. The reason why I selected a logistic regression model was to assess how well the model fit on the high dimensional training data. If the model could accurately predict the outcome, it would have the advantage of being more interpretable than other complex models. However, I believe that because of high dimensions of the data, the model would likely have high variance and will overfit.

   After fitting the logistic regression model on the training data, the following accuracy scores were obtained:

   Training Accuracy = 84.80537%
   Test Accuracy = 83.4838%

   Although this seemed like a decent test accuracy, on predicting the labels of the test data, I obtained an accuracy of only 75.15%. The model could not accurately predict unlabeled data and most likely overfit on the training data.

   Ridge Regression: To overcome overfitting, I decided to reduce the dimensions of the data. Although there are various techniques for variable selection – best subset selection, forward or backward stepwise regression, and Lasso & Ridge regularization, I decided to use Ridge regression method because it penalizes the predictor variables based on their association with the response variable. It reduces the coefficient values of the predictor variables which do not have a significant association with the response variables.

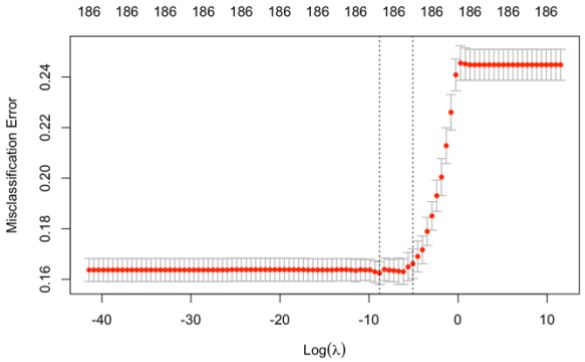
After performing ridge regression and cross validation on the training data, I obtained the following accuracy scores:

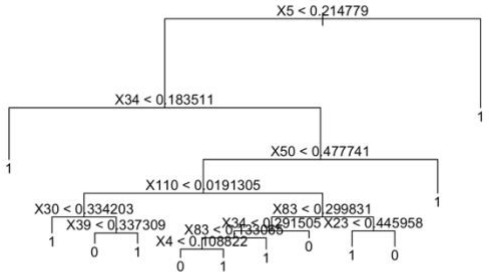
Training Accuracy = 84.86228%
Test Accuracy = 83.66968%

There wasn't a huge change in the accuracy scores as compared to a logistic regression model. Therefore, I concluded that it is likely that a linear model would not be able to capture the relationships between predictor and response variables well. Therefore, I decided to use non-linear tree-based models for the analysis.

Decision Tree Classifier: I selected a decision tree classifier for non-linear modeling. By fitting the overgrown tree, performing cross-validation, and pruning it, I obtained the following accuracy scores:

Training Accuracy = 86.61507%
Test Accuracy = 85.71429%





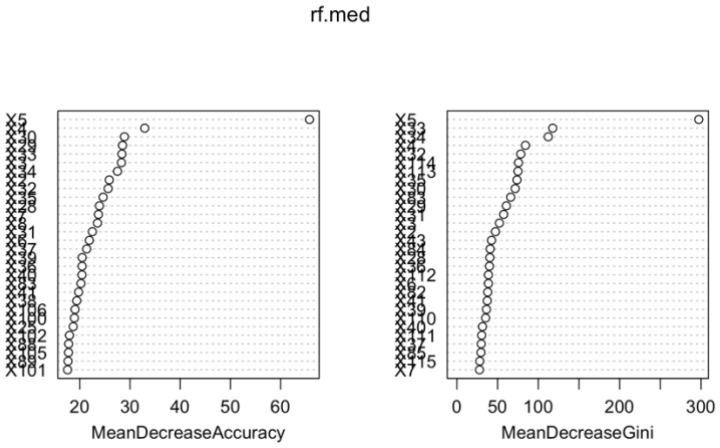
By analyzing the pruned decision tree, it could be seen that X5 variable had the highest impact on the accuracy and gini impurity which is why it has been chosen as the root node by the decision tree classifier. Although the accuracy scores were in a similar range as those obtained with linear models, it is possible that the model didn't overfit on the training data like linear models did. To obtain a better accuracy, I decided to fit a random forest classifier because it develops multiple decision trees with different variable combinations and generally performs well on non-linear data as compared to decision trees.

Random Forest Classifier: Because random forest classifiers perform well on non-linear data, I fit a random forest classifier on the training data. One of the limitations of the model compared to decision trees is its low interpretability. However, by analyzing the mean decrease in accuracy and mean decrease in Gini impurity, the most significant variables can be identified. After modeling the

classifier by setting mtry = p/3 which is approximately equal to 60, following was the test accuracy score obtained with a Random Forest Classifier:

Test Accuracy = 96.99947%

With a test accuracy of ~97%, the random forest classifier outperformed all the other models. Following is the feature importance graph:



Mean decrease accuracy gives the measure by which the accuracy of the model would decrease if the variable was not selected for modeling. Mean decrease gini accuracy gives the measure by which the gini impurity would reduce if the variable was not selected for modeling. When gini impurity reduces, the homogeneity of the data reduces, and the model cannot accurately predict the response. From the plot, it can be observed that the variable X5 most likely has a high association with the response variable. The rest of the variables which have high prediction power seem to be in the range from X4 to X115.

6. **Conclusion:** The random forest classifier model gives the best accuracy score for predicting whether an ECG signal is normal or abnormal. Moreover, analyzing the mean decrease accuracy and mean decrease gini impurity, a strong association of X5 variable with the response label can be observed. The model captures non-linear relationships between the variables, is interpretable, and provides a high test prediction accuracy.