# **Cardiovascular Disease Prediction using Machine learning**

Gloria Gu, Jingya Cheng, Yvonne Feng

**Abstract:**

The WHO states that CVDs are the leading cause of mortality worldwide. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. In this study, we used ML models and compared the accuracy of models for making predictions on having cardiovascular disease. In results, KNN had slightly higher accuracy than others. Further adjustments need to be made to achieve better performance.

**Problem Description:**

Cardiovascular diseases (CVDs) refers to a group of diseases that cause problems in heart and blood vessel conditions. The World Health Organization states that CVDs are the leading cause of mortality worldwide, killing more people each year than any other cause combined. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths(3). From the information above, it is clear that cardiovascular disease (CVD) is the leading cause of increased hospitalization, mortality, medical expenses, and productivity losses in the United States(3).

Premature deaths can be avoided by identifying those who are most at risk for CVDs and ensuring they receive the proper care. The previous study indicated the limitations with moderate discrimination and poorly calibrated utilizing the predicted. Therefore, further analysis on the predicted model is required.

In order to ensure that more people may live healthy lives, we will apply effective data-driven methods for predicting cardiac problems that can enhance the overall prevention process. The unsupervised learning algorithm K-means will be conducted before the supervised algorithm. Then, the variety of Machine learning algorithms will be implemented with two types of predicted models – regression models and classification models, including logistic regression, K-Nearest Neighbors, Linear discriminant analysis, decision tree. The logistic regression will be conducted to check the effect of variables on having a CVD. Our goal is to find the most accurate algorithms to predict the presence of cardiovascular diseases.

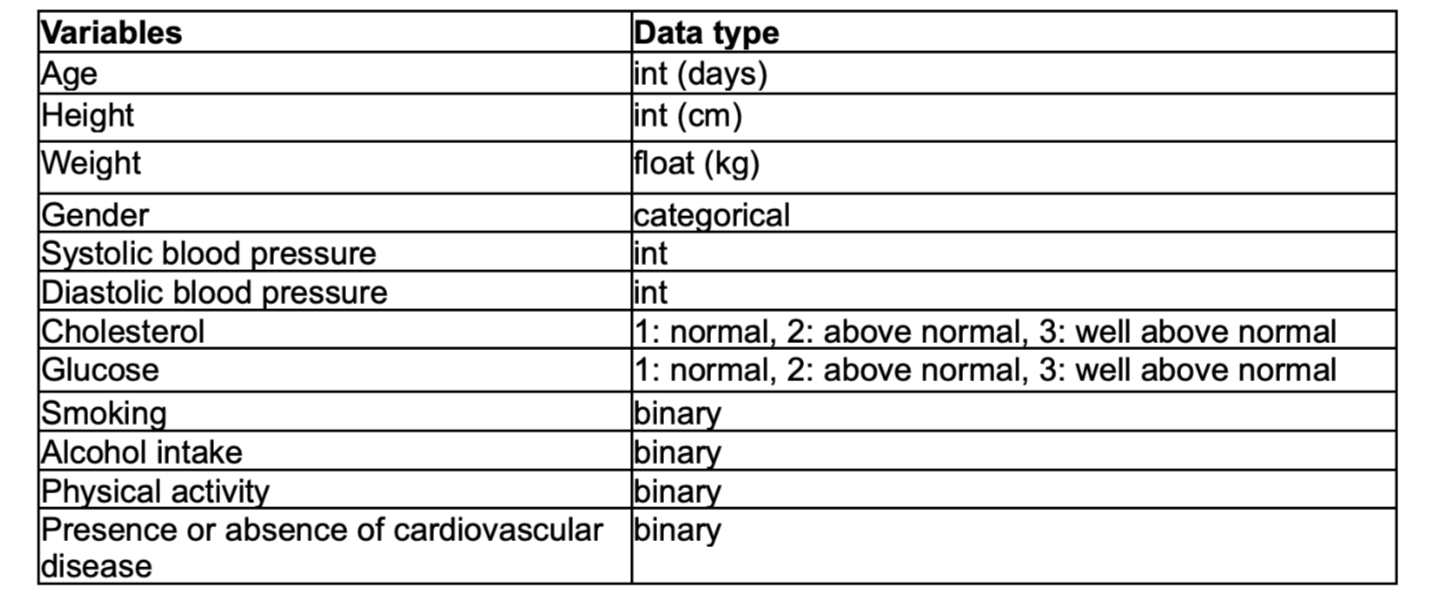
According to previous studies, demographic variables and health related variables are all associated with diagnosis of cardiovascular disease. Following the selection of the variables, we will suggest a model for predicting whether a particular individual is susceptible to cardiovascular disease based on a classification and prediction of the most significant variables. Then, the machine learning models will be used to predict the accuracy of findings.

**Aim1**: Find the significant features related to cardiovascular diseases

**Aim2**: Compare accuracy of models for making predictions on having cardiovascular disease

**Data Description and Preprocessing:**

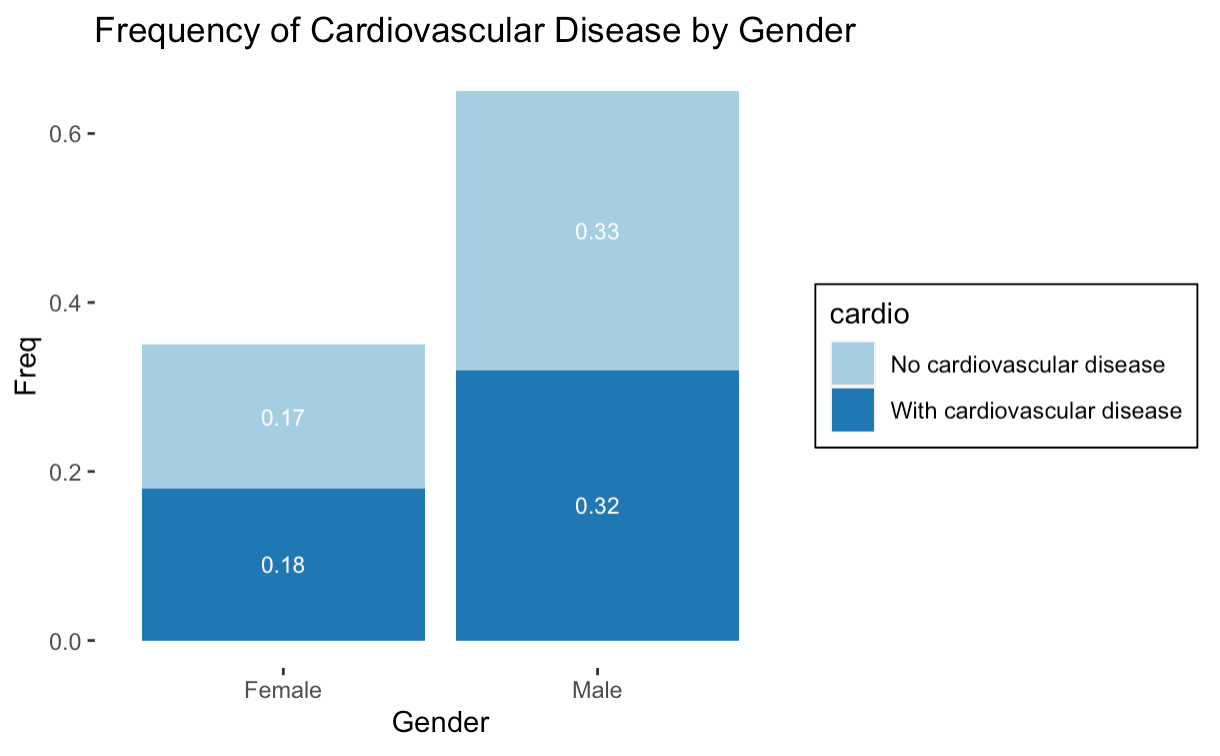
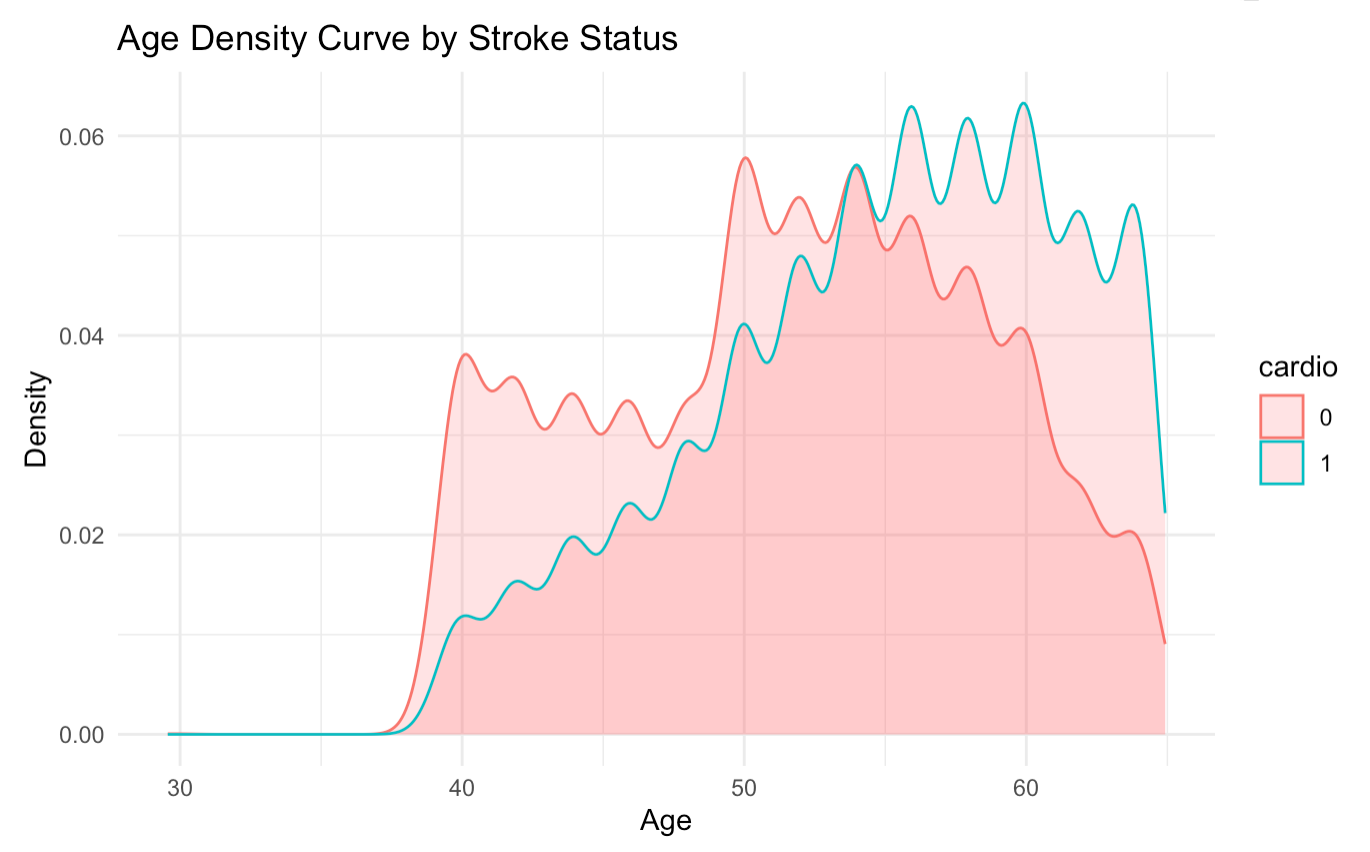
The dataset for this project is cardiovascular disease dataset collected by Svetlana Ulianova, which can be downloaded from kaggle. The datasets are collected in medical examinations and contain information about patients including demographics, encounters, lab Results and vital Signs, which can be used to predict whether the disease exists or not. Those variables are important features for model training and validation. There are 70,000 observations in the datasets and 12 variables.

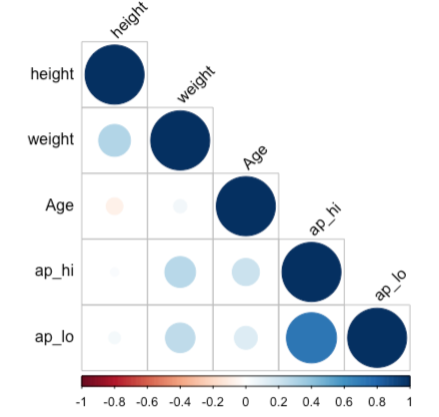
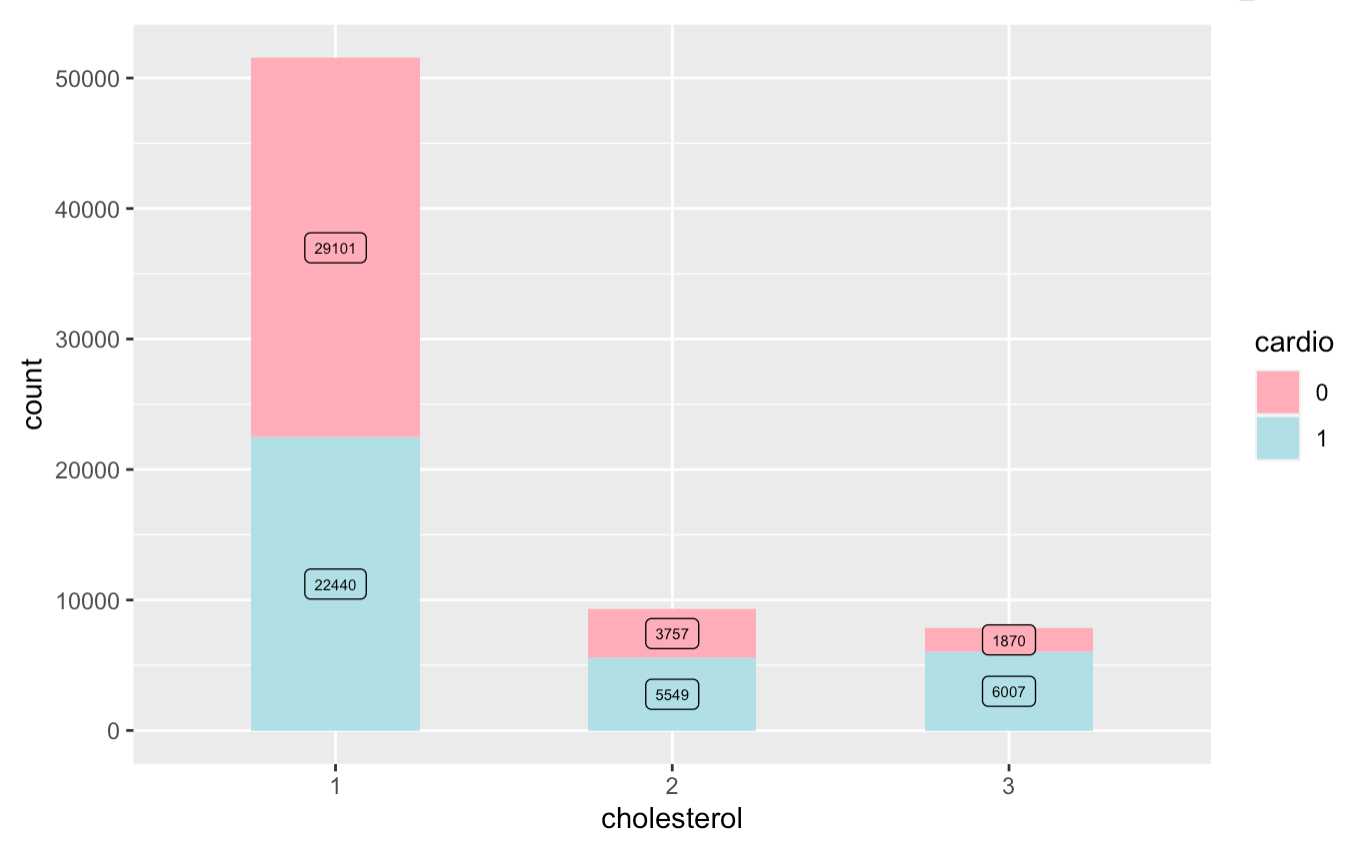


*Table1: Description of variables of cardiovascular disease dataset*

No missing value was reported from original datasets. However, there are some incorrect values which would not be presented in reality. In the data cleaning process, there are negative records and extreme high values in blood pressure. We excluded the **ap\_hi** that is smaller than **ap\_low** and > 370 or <0 for both variables. The dimension of the cleaned dataset contains 68724 observations with 12 variables.

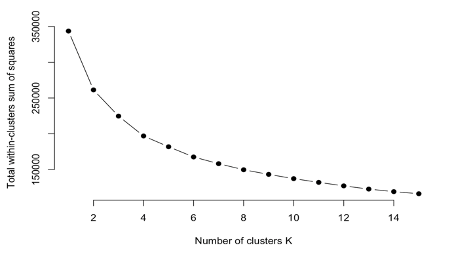
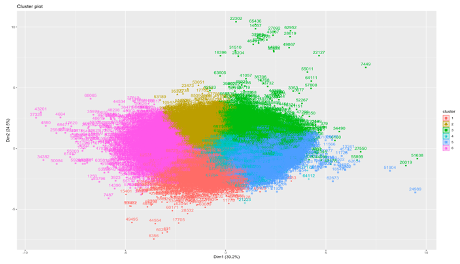
Exploratory analysis is performed to give an overall description on the whole dataset. Figure 1 shows that more male had cardiovascular disease than females and approximately half of the individuals in each gender were diagnosed with it. By Figure 2 the density curve of age indicates individuals diagnosed with cardiovascular disease tend to have higher age than those without. Also, more individuals with cardiovascular disease have abnormal cholesterol levels (either above or well above normal level) indicated in Figure 3. The correlation plot of all continuous facility and resident level covariates is shown in Figure 4, it is clear that variables ap\_hi and ap\_low are highly correlated, therefore, a variable selection procedure is recommended for further analysis to improve the model accuracy.

*Figure 1: Frequency of Cardiovascular Disease by Gender**Figure 2: Density Curve of Age by Cardiovascular Disease*

*Figure 3: Number of cholesterol level by cardiovascular disease Figure 4: Correlation Plot of all continuous variables*

We divided the cleaned dataset into training and test sets based on an 80:20 partition to ensure feasibility and accuracy of analysis.In both cases, the training set serves as a basis for fitting the model, while the test set is used to determine the model's performance in terms of various indicators, such as accuracy and precision. Also, data was scaled before applying further algorithms, the reason is that a feature with a larger range will overshadow or diminish a smaller range feature, thereby negatively impacting the performance of all distance-based models as higher magnitudes will be given more weight.

Since the number of variables is small, PCA is not required to reduce dimensionality. To obtain an understanding of the cluster structures and number of features, we use k-means to cluster the continuous variables. K-means is an unsupervised machine learning algorithm that takes unlabeled data points as input and clusters them into similar groups according to their distance to the cluster center. Firstly, the cluster centers are randomly chosen and the number of centers, which are the number of clusters, are determined before implementing the algorithm. Then, the distances of each data point from the centers are calculated. By changing the centers to obtain the sum of minimized distances, we will have the cluster structure. Therefore, we calculated the Within-Cluster-Sum of Squared Errors and plotted the elbow plot for optimal K value for k-means for our dataset. From Figure 5, we can observe that the elbow method indicates there should be six clusters. Therefore, we used the “kmeans” function in R to cluster the five continuous variables. Figure 6 is the visualization of the six clusters.

*Figure 5: Elbow Plot Figure 6: Cluster plot for k-means*

**Core methods:**

***Stepwise regression:***

After fitting in the generalized linear model, we performed stepwise regression to select variables using AIC. There are two different criteria to choose variables:1) AIC 2) Individual P-value. AIC is a goodness of fit measure that favors smaller residual error in the model but penalizes for including further predictors and helps avoiding overfitting. In addition, it can be misleading by just Looking at individual P-values, especially in the case when variables are highly correlated, it will give big p-values, which does not mean the variables are useless.

***Logistic regression:***

Then, we fit in a logit model that is used for classification and prediction. Logistic regression estimates the probability of an event occurring. In logistic regression, a logit transformation is applied on the odds— the probability of success divided by the probability of failure. The model estimates maximum likelihood estimation (MLE). The logistic function is represented by the following formulas:

Logit(pi) = 1/(1+ exp(-pi)) lng(pi/(1-pi)) = Beta\_0 + Beta\_1\*X\_1 + … + B\_k\*K\_k

The inference of statistics can be used to check if the log odds and continuous covariates are linear related.

***KNN and LDA:***

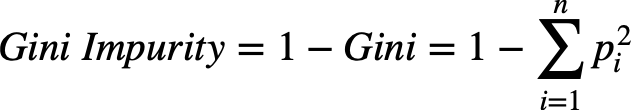
Starting KNN as the first non-parametric classification method, given a test observation , the KNN classifier identifies the K points in the training data that are closest to represented by , a tuning parameter K determines the way model is trained. The k-nearest neighbors model using K neighbors estimates this probability for as the fraction of training data points in whose response values equal ,

An alternative classification algorithm is linear discriminant analysis (LDA), which assumes predictors to follow multivariate normal distribution with a common covariance matrix for producing linear decision boundaries similar to logistic regression. LDA calculates the linear discriminants using the estimated mean and variance from a normal distribution.

ROC (Receiver Operating Characteristic) Curve helps to visualize the tradeoffs between sensitivity and specificity of a prediction based on selected models for KNN and LDA which tells how well a model has fitted the data. Computing the area under the curve is one way to summarize it in a single value to determine the specificity and sensitivity of the model. The closer the AUC value is to the 1, the better the given model fits the data.

***Decision Tree:***

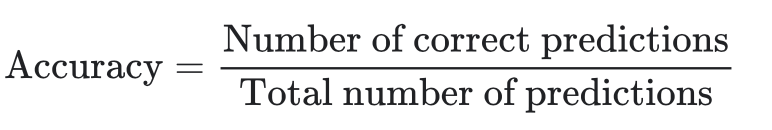
Decision Tree is a flowchart-like tree structure that is used for classification and regression with no assumption of the distribution. In this case, it can be used when outcome and covariates are not linearly related. It splits a data set into smaller subsets. The tree starts from a root node. This root node is then divided into groups of decision nodes based on conditional findings and observations. A node is referred to as a leaf node if it doesn't divide into additional nodes. A branch or sub-tree is a division of a decision tree. Gini impurity is calculated to identify the best split.



It is required to determine the weighted total of Gini Impurity for both child nodes in order to determine the optimal split. This is done for all potential divides, and the split with the lowest Gini Impurity is chosen as the best split.

***Accuracy:***

For all the models that we use, a confusion matrix is calculated. Then we calculate the accuracy of the model according to the test prediction accuracy. Accuracy is a metric used to evaluate classification models, it is the percentage of predictions that the model is correct by using the following formula:

****

**Results:**

***Logistic regression:***

We selected the best model with the smallest AIC which excluded the “GENDER” variable. The final model

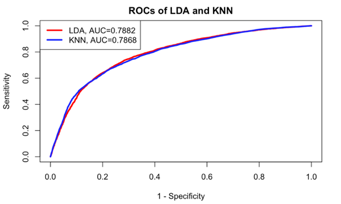
Logit ()..

Considering the hypothesis with H0…….. . The P-values are all smaller than 0.05. Therefore, we fail to reject the null hypothesis and conclude that the variables are all associated with the diagnosis of cardiovascular disease. For a continuous variable (e.g. height), beta1: A one unit increase in height is associated with a multiplicative change of 0.99 in the odds of a patient diagnosed with cardiovascular disease. For a categorical variable (e.g. smoke), beta9: for a person who had higher level of cholesterol has1.46 times the odds of a person who had lower level of cholesterol of having cardiovascular disease.

***KNN and LDA:***

In our model, an initial sight was given by using the starting choice for K was the square root of the number of observations in the training set, then optimizing it with computing distance between each observation in the test set and each observation in the training set, k-fold cross validation was used to decide the K neighbors. It is highlighted that the result for the tuning parameter k had the best overall accuracy which is 232 (since we have 10 folds in cross validation), the model with 232 neighbors performed the best. The estimated performance is evaluated in our training dataset, the confusion matrix gave 1502 false negatives and 2343 false positives.

Figure 1 indicated the AUC values and showed the ROC curves for KNN and LDA algorithms respectively, in our case AUC for both are quite similar with LDA has a slightly higher AUC, however it is fair to conclude that KNN and LDA provided an equivalent performance in making predictions.

**

*Figure 1. ROC curves and AUC values for LDA and KNN algorithms*

***Decision Tree:***

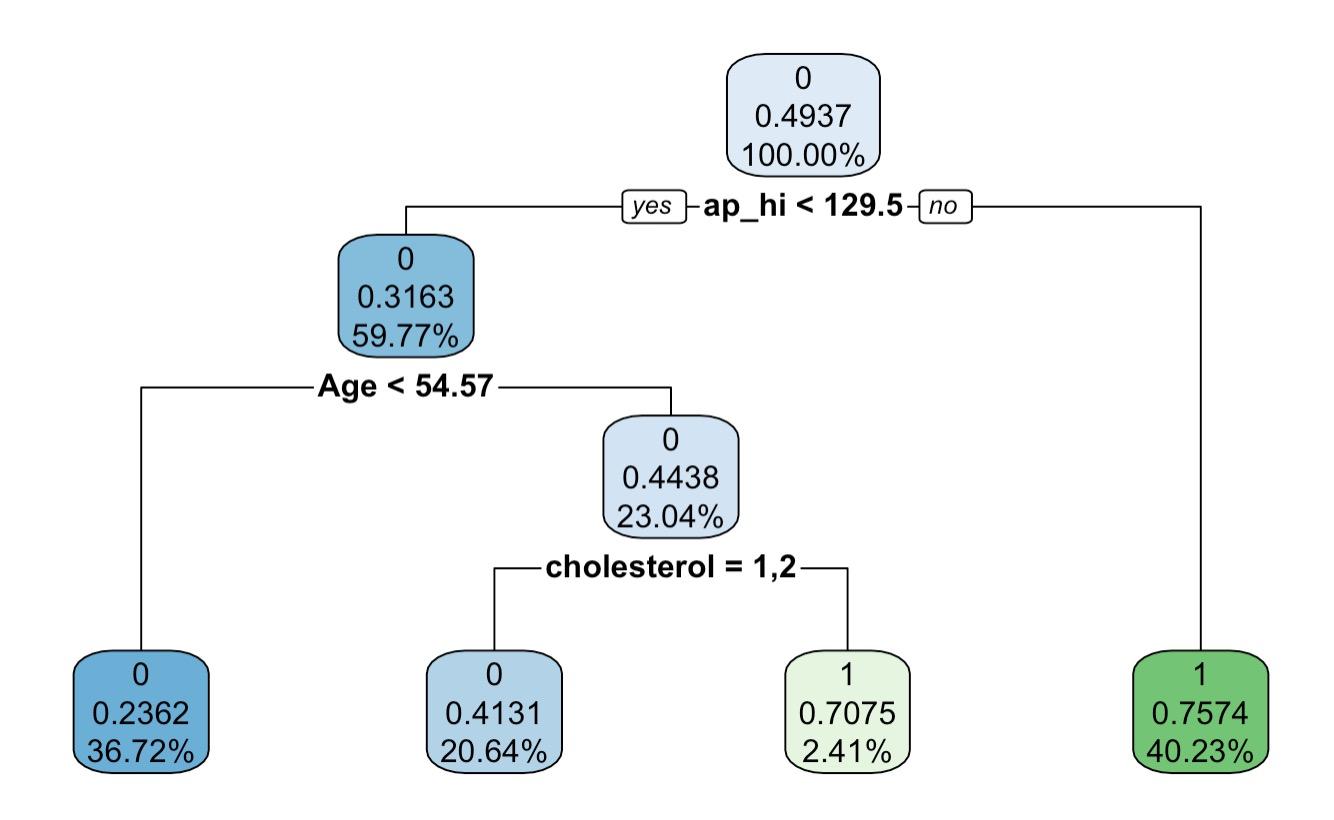


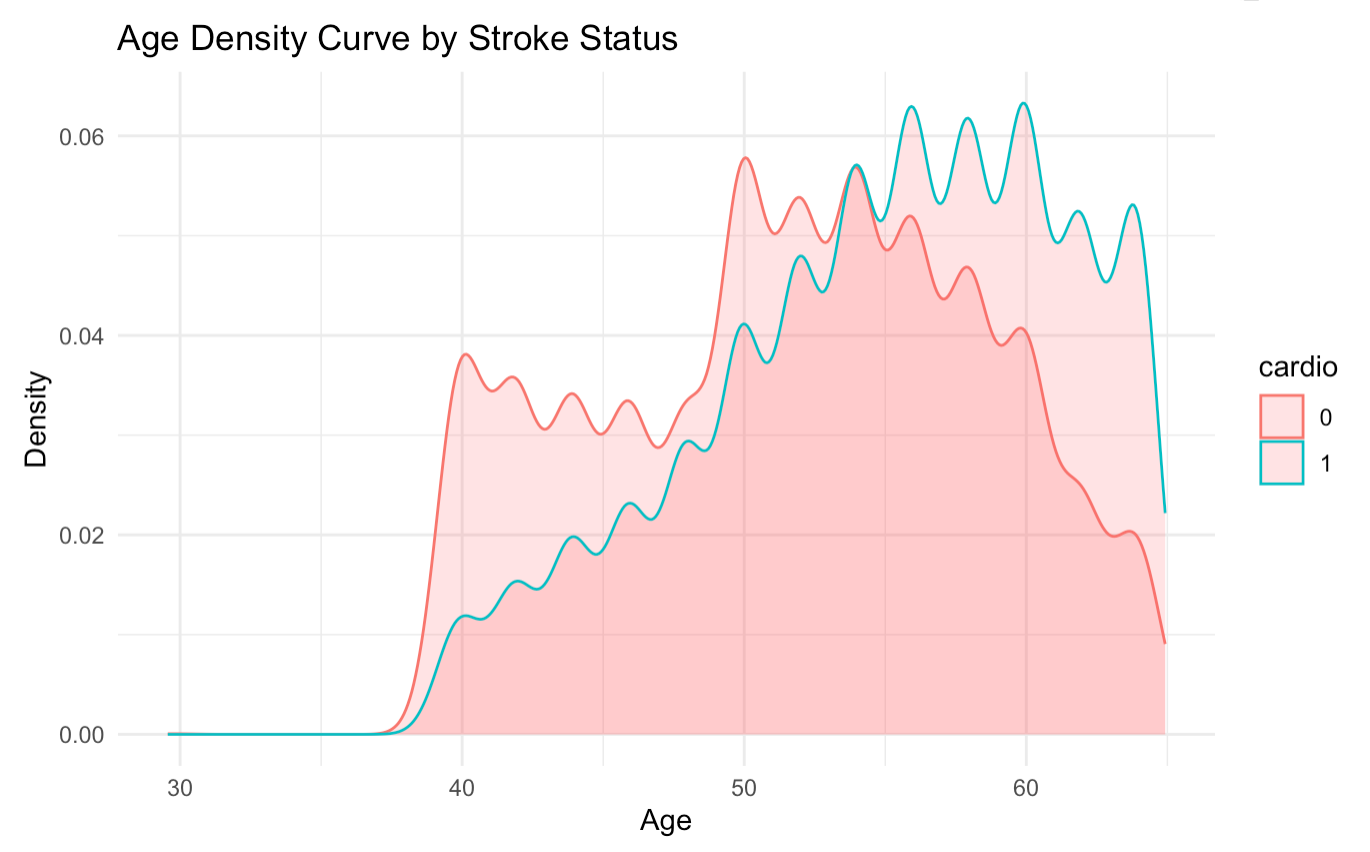
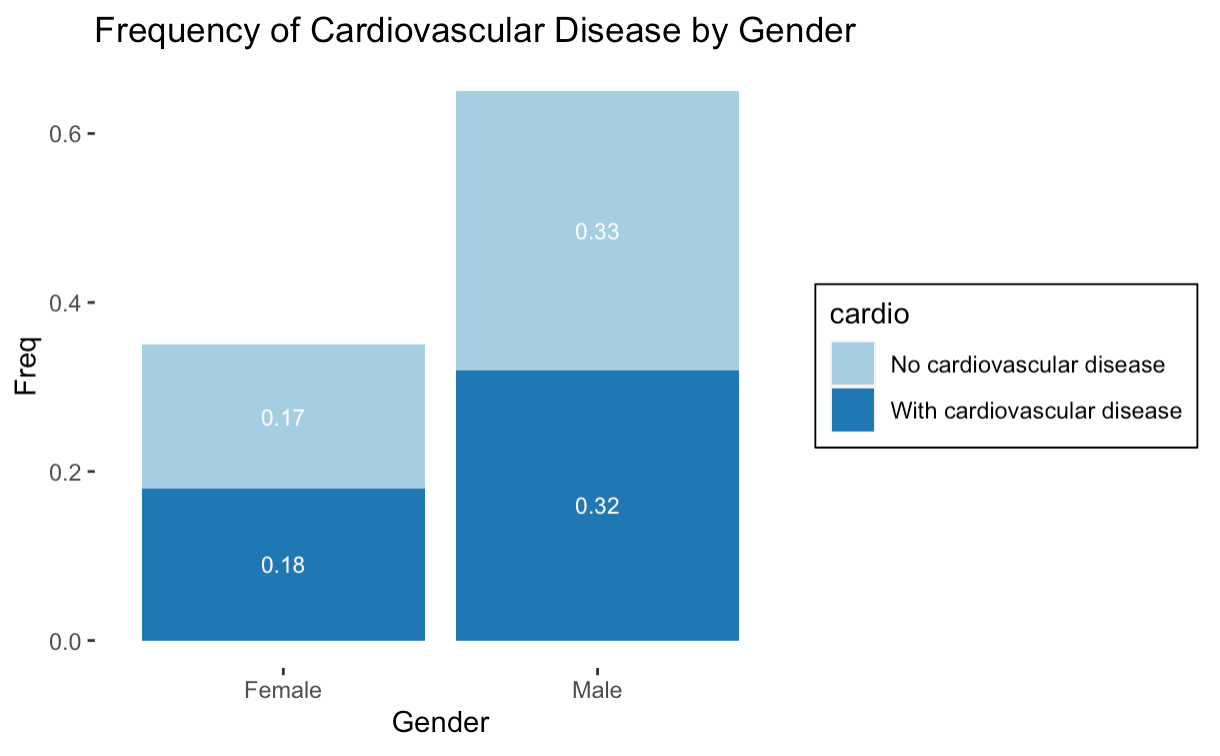
Figure: Decision Tree Chart of Training set

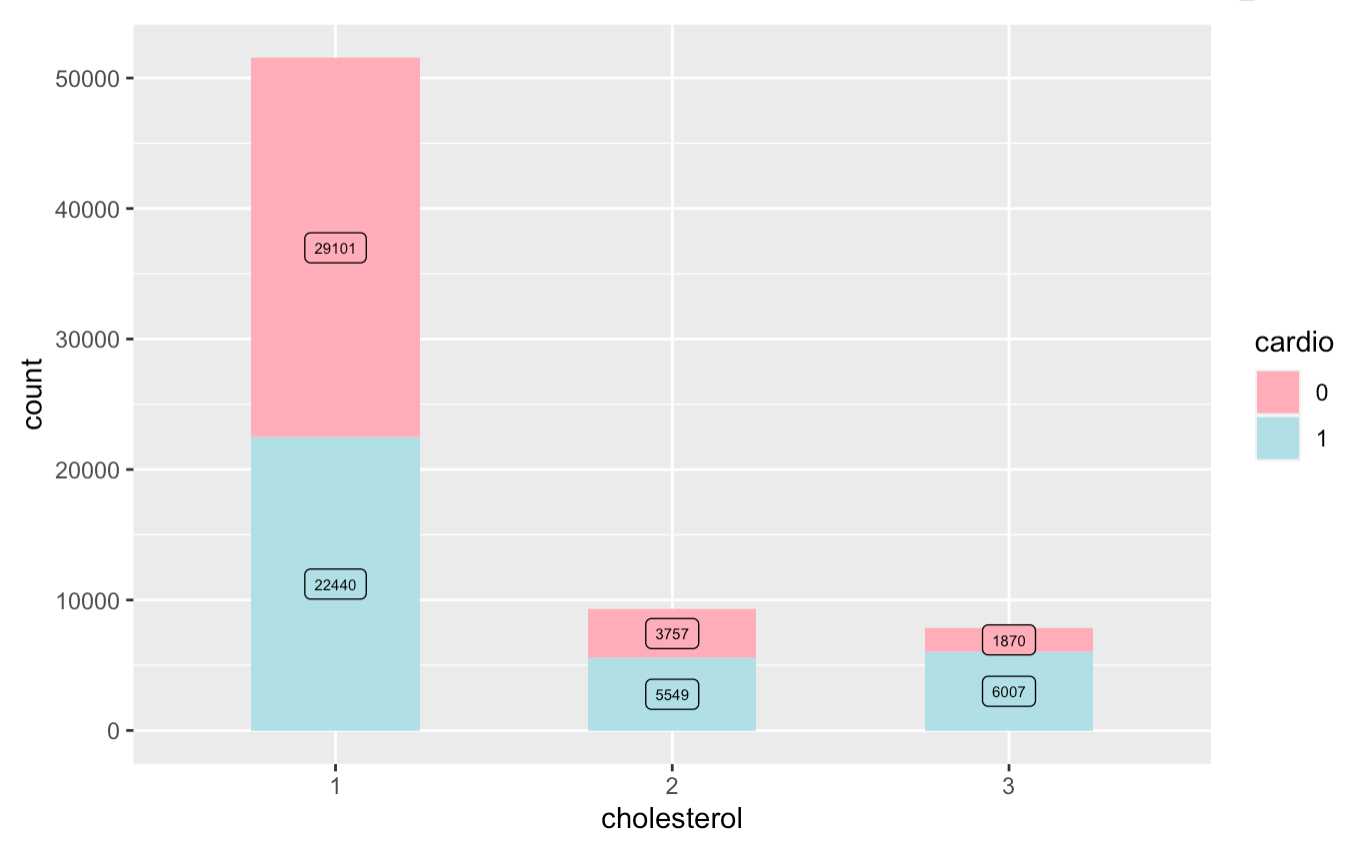
From the decision tree chart, the train set was classified into 4 groups according to three classifiers: ap\_hi, Age and cholesterol. The Gini impurity was calculated for each node. For interpretation, an individual who has systolic blood pressure below 129.6 and age below 54.57 years old is less likely to have cardiovascular diseases. While a person with systolic blood pressure above 129.6 is more likely to have cardiovascular diseases. The confusion matrix presented the performance of the decision tree on the test dataset. The accuracy is around 72%, recall is , precision is , and F-measure is .

**Conclusion:**

In table 1, the results for the accuracy of each model are presented. We can observe that there is no significant difference among models. All the models that we use are all around 72% accurate in prediction. The accuracy values indicate the performance of the models are relatively good. The ROC curves show that the LDA and KNN model have similar performance in prediction. We can observe that KNN is slightly more accurate than other models. However, we cannot conclude that KNN is the “best” model, because the accuracy only represents the prediction performance on the test set, but we can use these values to see whether we can achieve better performance by changing or adjusting models.





1. Logistic regression:

* What type of problem are you addressing? E.g. classification, regression, forecasting, recommendation system...
* What characteristics were key when choosing the approach?
* How does the chosen approach address these characteristics?
* How did you choose model parameters?
* **How did you evaluate the performance and perform model selection? (e.g. cross-  
  validation and metrics used as evaluation criteria)**

**References**

1. Jiang, Y., Zhang, X., Ma, R., Wang, X., Liu, J., Keerman, M., ... & Guo, H. (2021). Cardiovascular disease prediction by machine learning algorithms based on cytokines in Kazakhs of China. *Clinical epidemiology*, *13*, 417
2. .Narloch, J. A., & Brandstater, M. E. (1995). Influence of breathing technique on arterial blood pressure during heavy weight lifting. *Archives of physical medicine and rehabilitation*, *76*(5), 457-462.
3. Sanz, M., D'Aiuto, F., Deanfield, J., & Fernandez-Avilés, F. (2010). European workshop in periodontal health and cardiovascular disease—scientific evidence on the association between periodontal and cardiovascular diseases: a review of the literature. *European heart journal supplements*, *12*(suppl\_B), B3-B12.