Machine Learning for Cardiac Health: Predicting Coronary Artery Disease

By

George Baffuor Awuah & Nii Adjetey Adjei-Annan

May 3, 2024.

Abstract

This study evaluates the efficacy of multiple machine learning models in predicting coronary artery disease (CAD) using clinical data from the Sani dataset. We employed five different classification techniques: Logistic Regression, K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), and Random Forest. Our analysis included a thorough assessment of each model based on accuracy, recall, F1 score, Area Under the Curve (AUC), and mean sensitivity. The results indicated that the Logistic Regression model outperformed others with an accuracy of 83.52%, an F1 score of 72.73%, and an AUC of 88.52%, demonstrating high effectiveness in distinguishing between patients with and without CAD. The Random Forest model also showed strong performance, especially in accuracy and AUC. Our findings highlight the potential of advanced machine learning techniques in enhancing diagnostic accuracy in medical settings, thereby supporting healthcare professionals in making more informed decisions regarding CAD management and guiding future research and clinical applications in cardiovascular disease diagnosis.

1. Introduction

The World Health Organization (WHO) lists cardiovascular diseases (CVDs) as the leading cause of death globally (WHO, 2018; as cited in Alizadehsani et al., 2019). Coronary artery disease (CAD), the most common type of (CVDs) contributes to about 32% of the total deaths globally (WHO, 2022; as cited in Kolukisa & Bakir-Gungor, 2023), and it is projected to cause the death of approximately 23.6 million people in 2030 according to the WHO estimates. The potential of early diagnosis and timely medical care in preventing adverse outcomes and reducing the cost of treatment and the high cost and invasive nature of clinical diagnosis has made the use of machine learning (ML) techniques for early detection of CAD popular. (Sayadi et al., 2022; Alizadehsani et al., 2019; Abdar et al., 2019).

In recent years, several studies have applied various ML algorithms to different datasets in order to detect CAD. According to a recent review, as of 2019, there have been 149 published ML-based CAD detection studies, with demonstrated promising results (See Alizadehsani et al., 2019). Even then, it can be challenging to determine which ML model works best as performance results vary significantly depending, among other factors, "the sample size, the number and type of features and the ethnicity of the patients in the data set" Kolukisa and Bakir-Gungor, 2023, p. 1). In the present study, we implement and compare the classification performance of five ML algorithms using the publicly available Z-Alizadeh Sani dataset.

2. Data Description and Pre-processing

Data description

The Z-Alizadeh Sani dataset is publicly available and obtainable from the UCL Machine Learning Repository. It contains 54 features and 303 records of patients "who visited Shaheed Rajaei Cardiovascular, Medicine, and Research Centre of Tehran, Iran" (Dahal & Gautam, 2020, p. 696). Of the 303 patients, 216 are diagnosed with CAD, and the remaining 87 are normal. Key features of patients recorded are of four broad categories: (1) demographic, (2) symptoms and examinations, (3) electrocardiogram (ECG), and (4) laboratory and echo features (Abdar et al., 2019; Dahal & Gautam, 2020) The target variable "Cath" is binary variable with labels "Cad" and "Normal" where the former indicates the presence of CAD and the later stands for normal patients.

Feature selection

To reduce noise in the dataset and enhance computation accuracy, we dropped irrelevant features using Pearson correlation algorithms. In Figure 1, we present the correlation matrix plot of chosen variables with respect to the target variable. In Table 1, we present summary statistics of the selected variables.

Figure 1: Correlation matrix of parameters in the dataset

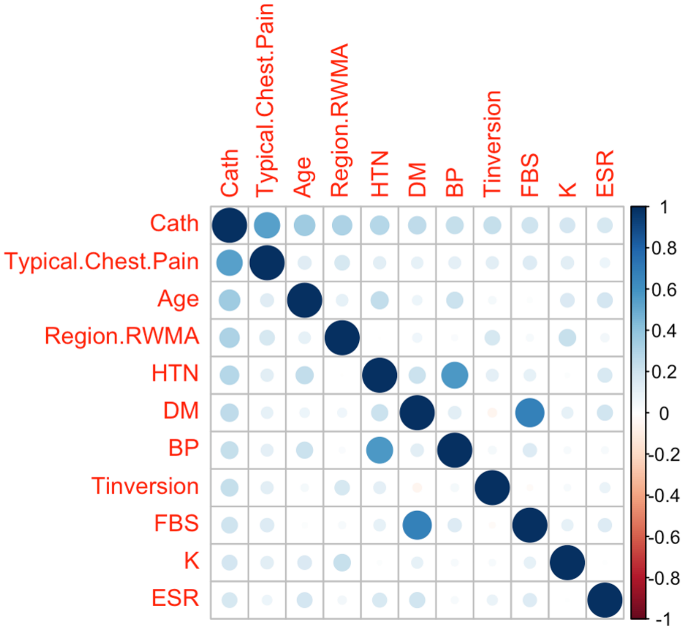


Table 1. Variable names, role, type, level and range

Variable name	Role	Type	Level	Range
Cath	Target	Characteristics	Nominal	CAD, Normal
Typical Chest Pain	Input	Characteristic	Nominal	Yes, No
Age	Input	Numerical	Interval	30 - 86
(Region.RWMA)*	Input	Numerical	Discrete	0, 1, 2, 3, 4
Hypertension (HTN)	Input	Characteristic	Nominal	Yes, No
Diabetes Mellitus (DM)	Input	Characteristic	Nominal	Yes, No
Blood Pressure(BP)	Input	Numerical	Interval	90-190
Tinversion	Input	Characteristic	Nominal	Yes, No
Fasting Blood Sugar(FBS)	Input	Numerical	Interval	62-400
Potassium (K)	Input	Numerical	Interval	3-6.6
Ejection fraction (EF.TTE)	Input	Numerical	Interval	15-60

* Regional Wall Motion Abnormality

Data Partition

We split the data into two in the ratio 70:30 training vs testing data to avoid overfitting. We proceed to use the train data (with 212 observations) to find the relationship between target and predictor variables, and then we use the test data with 91 observations to assess the performance of the model.

3. Machine Learning Algorithms

We implement five ML algorithms, namely, Logistic Regression, K-Nearest Neighbors (KNN), Naïve Bayes (NB), Random Forest, and Support Vector Machine.

The logistic Regression (LR) model is useful for predicting binary outcomes. It is one of the simplest and most popular methods to solve classification problems, particularly where the target variable is binary, as it is in the current study (Dahal & Gautam, 2020). Here, we fitted a logistic regression model on the selected variables using the glm command of R package.

K-Nearest Neighbors (KNN) modelling involves a completely nonparametric clustering process that is based on proximity relations between objects(Yuvalı et al., 2022). We used the KNN command of the R package to fit the KNN model.

Support Vector Machine (SVM) is a popular and powerful machine learning model that is useful for solving both regression and classification problems. The technique is “equipped with various kernels, such as linear, polynomial, radial, and sigmoid(Dahal & Gautam, 2020, p. 699). We use the svm command of the R package to fit this model.

Random Forest (RF) is one of the most widely used machine learning algorithms. It is obtained “by improving the Bagging algorithm since it selects trees that are not correlated” (Dahal & Gautam, 2020, p. 699). Random Forest is able to “handle large datasets with automatic variable selection and many estimators… and is reported to provide unbiased estimates” (Yuvalı et al., 2022, p. 4).

Naïve Bayes (NB) is a supervised learning algorithm and classification technique based on Bayes’ Theorem with an assumption of conditional independence among predictors. This assumption makes the algorithm make predictions quickly and accurately.

4. Model Comparison

To ascertain which model had better performance, we trained the models on the training data, fit them to the test data and retrieved Accuracy, Recall, F1 score and area under the receiver operating characteristic curve (AUC).

Recall, all called true positive rate (TPR) or sensitivity, is the proportion of the actual positive cases that are correctly predicted as positive, while F1 score, all called true negative rate (TNR) or specificity, is the proportion of the actual negative cases that is correctly predicted as negative. Type 1 error, also false positive rate (FPR), is the proportion of the actual negative cases that is incorrectly predicted as positive, while type II error, also called false negative rate (FNR), is the proportion of the actual positive cases that is incorrectly predicted as negative. Finally, the proportion of the cases that is predicted accurately is called the Accuracy (Dahal & Gautam, 2020). The equations below define these measures.

$$Recall/Sensitivity = True\ positive\ rate(TPR) = \frac{True\ postive\ (TP)}{True\ positive(TP) + False\ negative(FN)}$$

$$F1\ score/Specificity = True\ negative\ rate(TNR) = \frac{True\ negative(TN)}{True\ negative(TN) + False\ positive(FP)}$$

$$Type\ I\ Error = False\ positive\ rate(FPR) = \frac{False\ postive(FP)}{True\ negative(TN) + False\ positive(FP)}$$

$$Type\ II\ Error = False\ negative\ rate(FNR) = \frac{False\ negative(FN)}{True\ positive(TP) + False\ negative(FN)}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

For a good model, we want lower type I and II errors and higher sensitivity and specificity. The model with the highest statistics, which are sensitivity, specificity, accuracy, and AUC, is considered the best model. The receiver operating characteristic (ROC) curve characterizes sensitivity/specificity tradeoffs for a binary classifier. We plot the ROC curve with false positive rate (1-specificity) on x -axis and sensitivity on y-axis at various threshold settings. This yields area under the ROC curve (AUC), an important measure of model performance with values ranging between 0 and 1 - AUC is close to 1, suggesting excellent performance. Here, we used the roc command of the R package to compute the AUC of ROC curve of each model.

5. Results

We present the summary of the performance statistics from the five models in Table 2. and highlight the highest value of the performance matrices. Based on these performance metrics, the Logistic Regression model outperformed all other models in all the metrics except AUC, where KNN outperforms the Logistic Regression model - the KNN model has the highest AUC score of 0.8969.

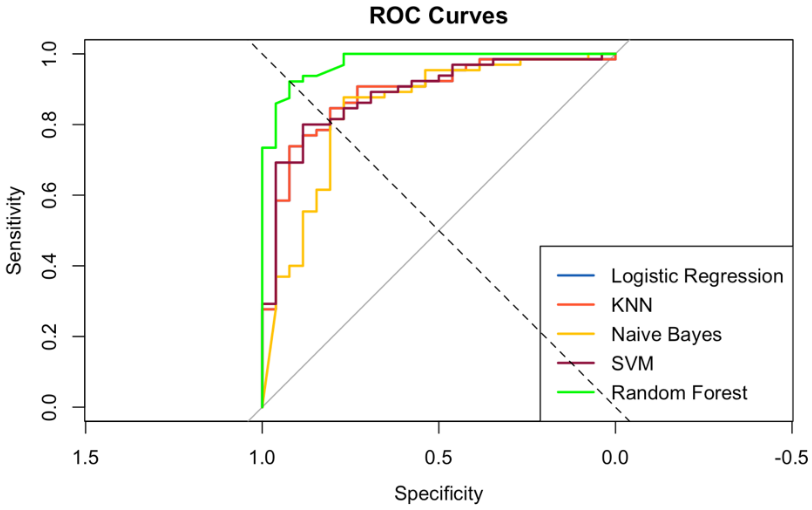
The performance of the LR model is outstanding because it has an accuracy of 0.8667, recall of 0.7692, F1 score of 0.7273 and AUC of 0.8852, which yields the highest mean score of 0.79 across the three of the four metrics. KNN model has the lowest score on F1 score, Naïve Bayes has the lowest value on recall and RF has the lowest value on AUC. Thus, while RF is the best performer on AUC, because of poor performance on other metrics, it records the lowest mean score - making it the worst performer of the five models.

Table 2. Model performance metrics obtained using test dataset.

Model	Accuracy	Recall	F1 score	AUC	Mean
Logistic Regression	**0.8352**	**0.7692**	**0.7273**	0.8852	**0.79**
KNN	0.7363	0.5714	*0.4000*	**0.8969**	0.76
Naïve Bayes	0.76923	*0.56756*	0.6667	0.8952	0.77
Support Vector Machine	0.8132	0.6923	0.6792	0.8940	0.75
Random Forest	*0.8022*	0.6538	0.6538	*0.8720*	*0.61*

Note: Highest value is boldened while the lowest is italicized in yellow

Figure 2. ROC curve for NB, LR, RF, SVM, and KNN.



6. Conclusion

In this study, we used Logistic Regression, K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), and Random Forest and the publicly available Z-Alizadeh Sani dataset to evaluate the performance of these models in predicting coronary artery disease (CAD). Utilizing the following performance matrices: sensitivity, specificity, accuracy, and area under the ROC curve (AUC) of the testing data, we find that that the Logistic regression model is able to predict the presence of CAD more effectively and accurately than other models with an accuracy of 83.52%, an F1 score of 72.73%, and an AUC of 88.52%. We propose further research to include other machine learning algorithms, such as artificial neural network, using more data or exploring other ways of extracting important features before feeding to the machine learning algorithm.

References

Sayadi, M., Varadarajan, V., Sadoughi, F., Chopannejad, S., & Langarizadeh, M. (2022). A machine learning model for detection of coronary artery disease using noninvasive clinical parameters. *Life*, *12*(11), 1933.

Abdar, M., Książek, W., Acharya, UR, Tan, RS, Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. *Computer methods and programs in biomedicine* , *179* , 104992.

Dahal, K. R., & Gautam, Y. (2020). Argumentative comparative analysis of machine learning on coronary artery disease. *Open Journal of Statistics*, *10*(4), 694-705.

Yuvalı, M., Yaman, B., & Tosun, Ö. (2022). Classification comparison of machine learning algorithms using two independent CAD datasets. *Mathematics*, *10*(3), 311.

Kolukisa, B., & Bakir-Gungor, B. (2023). Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis. *Computer Standards & Interfaces*, *84*, 103706.

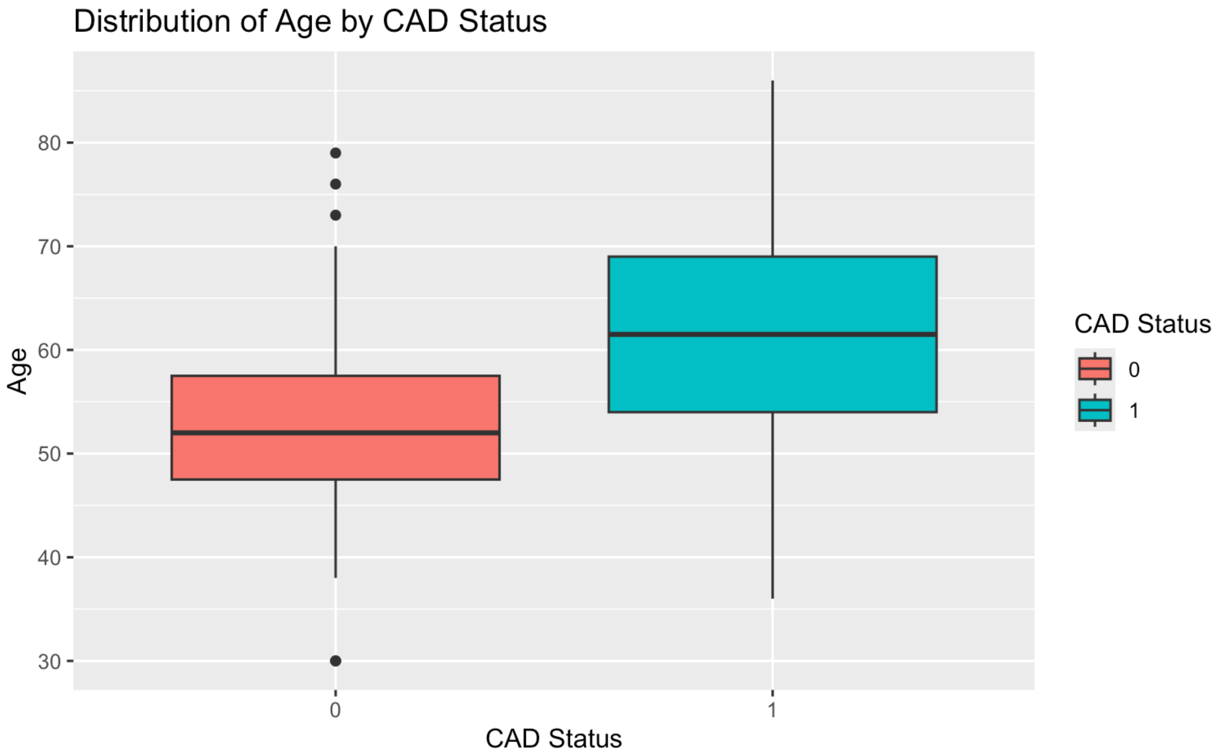
Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., ... & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in biology and medicine*, *111*, 103346.

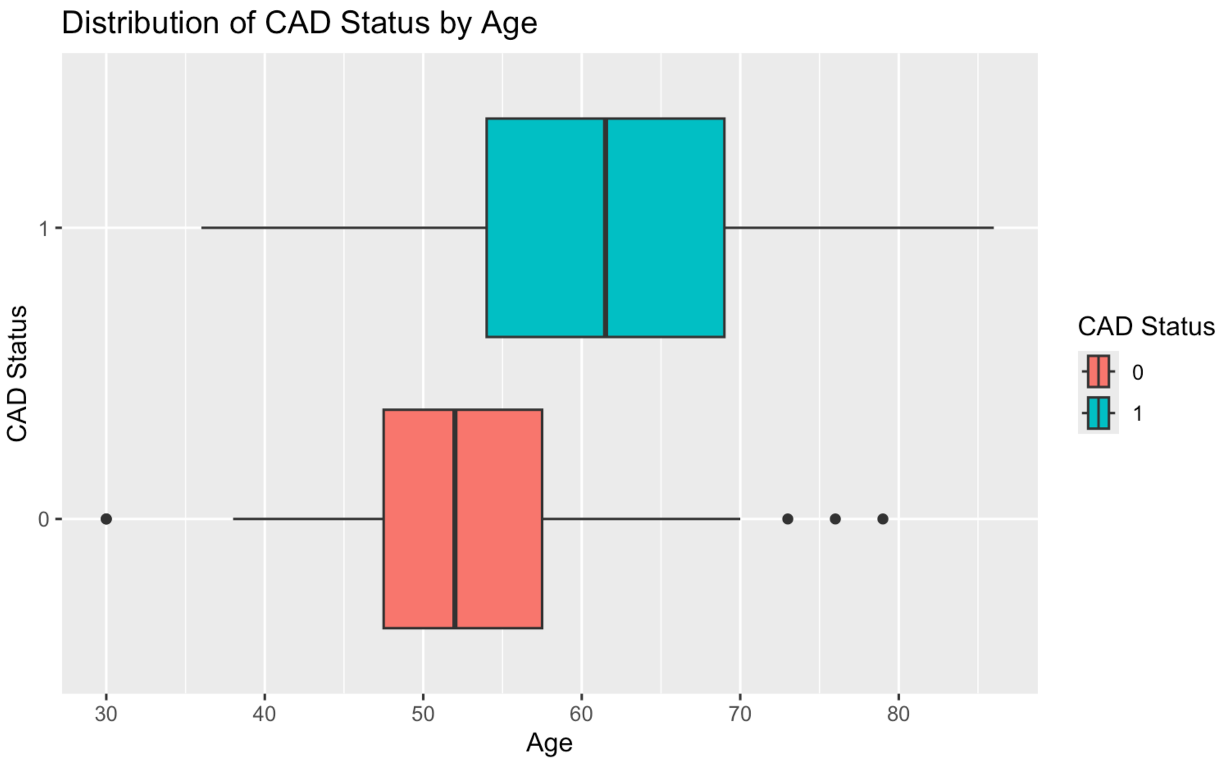
Terrada, O., Hamida, S., Cherradi, B., Raihani, A., & Bouattane, O. (2020). Supervised machine learning based medical diagnosis support system for prediction of patients with heart disease. *Advances in Science, Technology and Engineering Systems Journal*, *5*(5), 269-277.

Appendix: Table & Figures from descriptive analysis

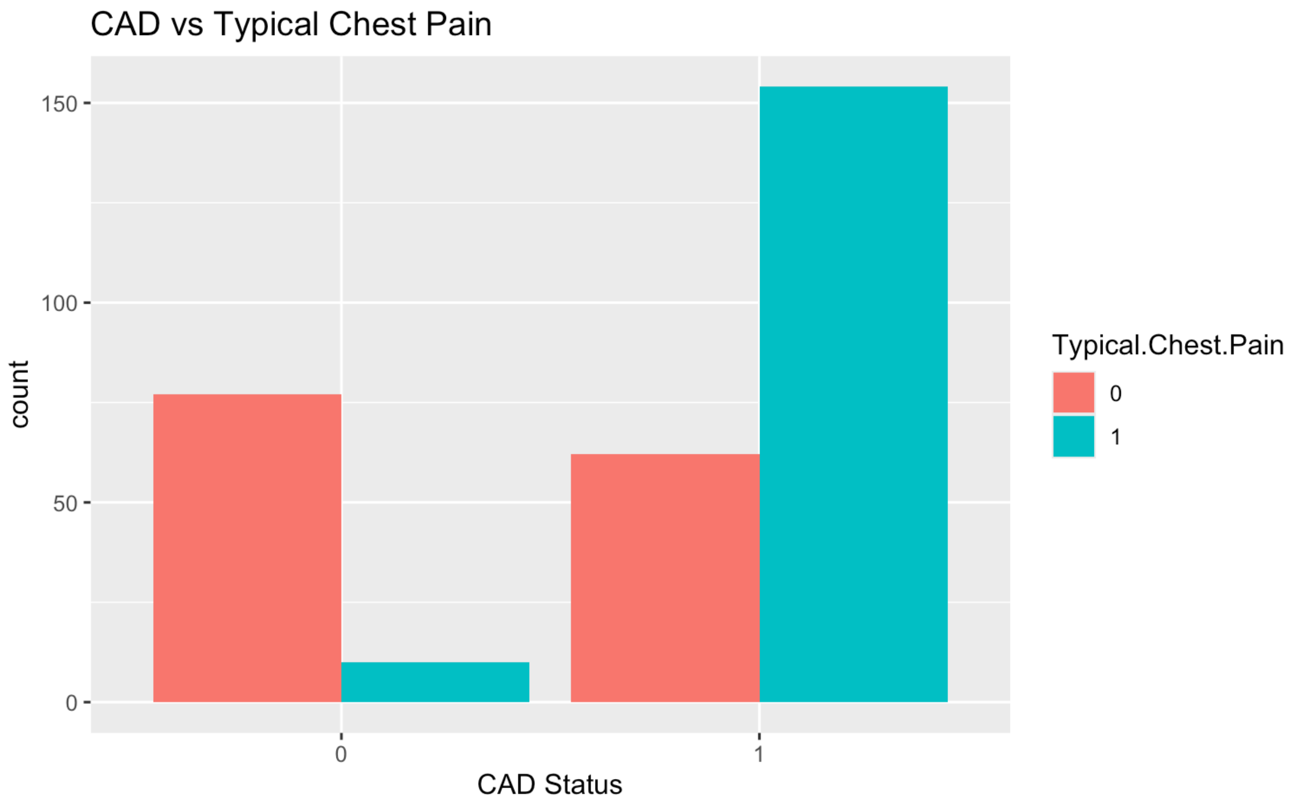
Summary of the name, type and description of chosen variables

Variable	Mean	Median	SD	Min	Max
Cath	0.71	1.0	0.45	0	1.0
Typical.Chest.Pain	0.54	1.0	0.50	0	1.0
Age	58.90	58.0	10.39	30	86.0
Region.RWMA	0.62	0.0	1.13	0	4.0
HTN	0.59	1.0	0.49	0	1.0
DM	0.30	0.0	0.46	0	1.0
BP	129.55	130.0	18.94	90	190.0
Tinversion	0.30	0.0	0.46	0	1.0
FBS	119.18	98.0	52.08	62	400.0
K	4.23	4.2	0.46	3	6.6
ESR	19.46	15.0	15.94	1	90.0

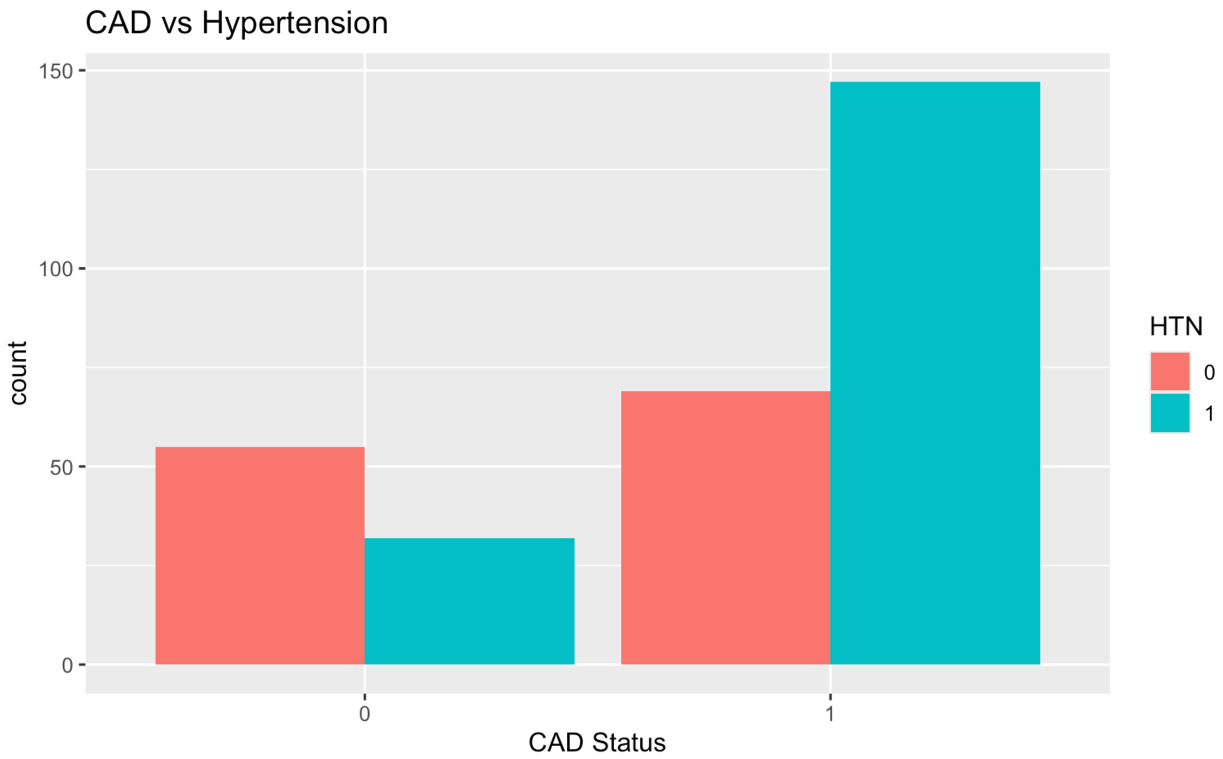


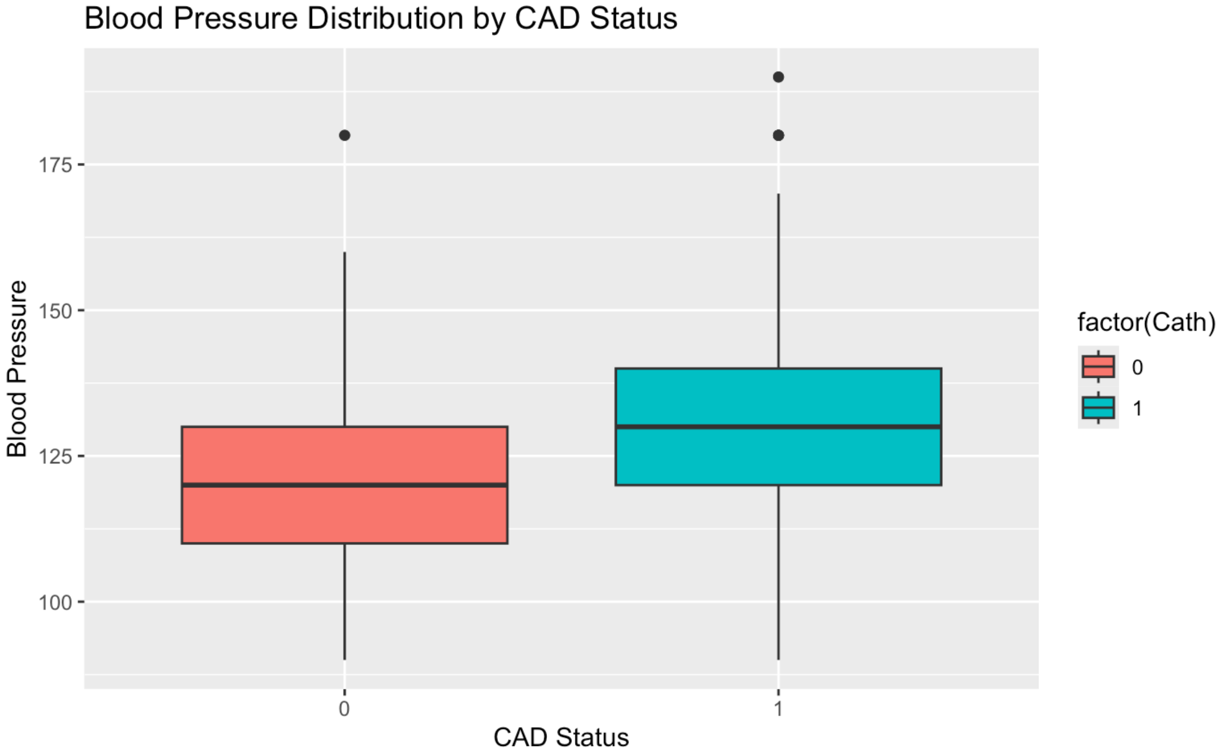


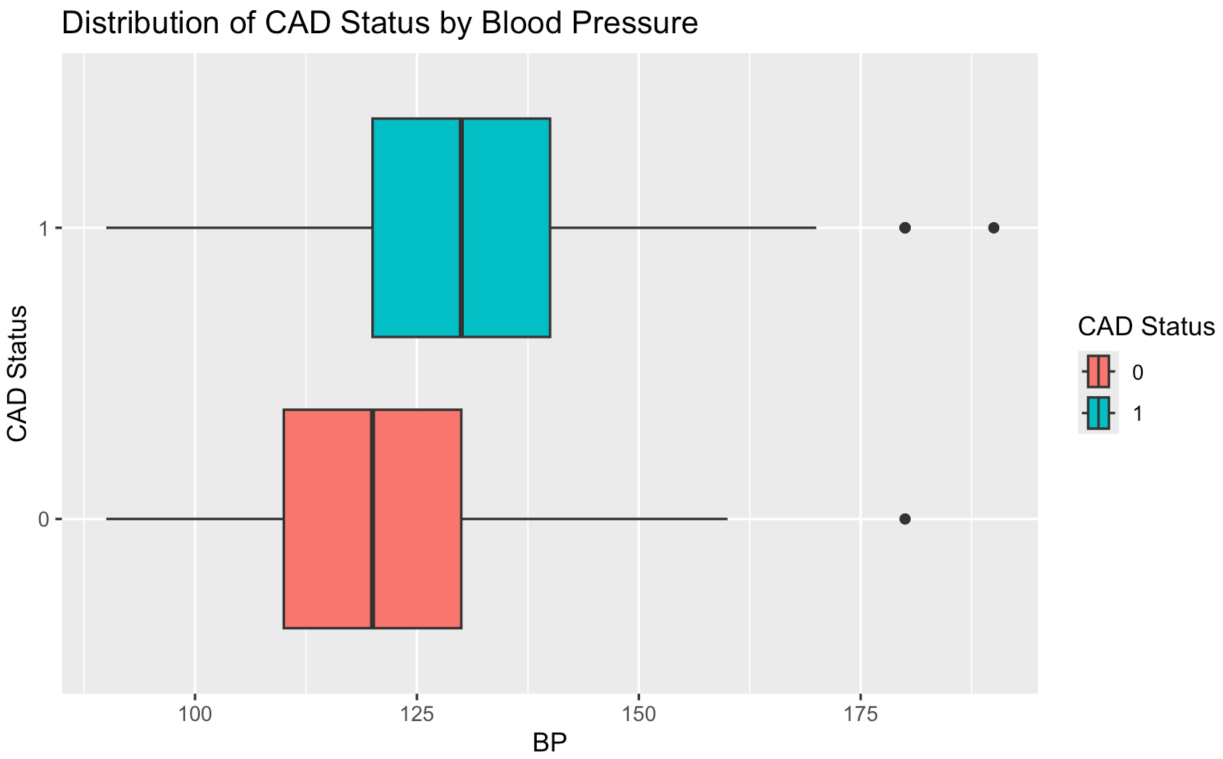
Distribution of CAD Status by Age



CAD vs Typical Chest Pain


CAD vs Hypertension
count
150
100
50
0
0
1
CAD Status
HTN
🟥 0
🟦 1


Blood Pressure Distribution by CAD Status
Blood Pressure
175
150
125
100
0
1
CAD Status
factor(Cath)
🟥 0
🟦 1


Distribution of CAD Status by Blood Pressure
CAD Status
1
0
100
125
150
175
BP
CAD Status
0
1