Student Name: Michael Ayomide, FADELE

Student Number: 239032281

Course Code: CETM 24

Topic: Impact of Data Science in Cardiovascular Diseases Prediction and Prevention (Data Analysis and R Coding)

Introduction

As the primary cause of death globally, cardiovascular diseases (CVDs) present a significant challenge to global health (World Health Organisation, 2021). For enhancing patient outcomes and directing the application of tailored medicines, early detection and accurate classification of CVDs are crucial. By using clinical data to forecast and classify the condition, the expanding area of data science in the digital era has the potential to improve CVD detection and assist healthcare professionals in making more informed decisions (Patel and Sengupta, 2020). This report aims to reliably classify the existence of cardiovascular disease in patients based on a wide variety of clinical indicators using data science methodologies. This report calculates an individual's risk of developing CVD based on their reported health status using a variety of statistical models. Age, blood pressure, cholesterol, and other clinically significant health indicators are included in the Kaggle dataset. Finding significant patterns and correlations that can indicate a higher risk of cardiovascular diseases is the aim of this investigation (Ulianova, 2020).

This report's following sections are arranged as follows: The following examines pertinent research in the area. Following a description of the dataset's specific features, model testing, evaluation, and exploratory data analysis (EDA) took place. The report first presents the

analytical processes' results, then explores them in relation to the larger field of medical data analysis and its implications. Finally, the report's conclusion is given.

Literature Review

Recent research has focused on data science in healthcare, particularly cardiovascular disease prediction. Krittanawong et al. (2020) envision AI and ML changing cardiology, particularly risk assessment and diagnosis. EHRs and clinical trials provide data for machine learning models to predict cardiovascular outcomes (Ambale-Venkatesh and Lima, 2015).

CVD datasets have yielded conflicting outcomes for machine learning. Logistic regression, decision trees, and Naive Bayes are examples. For its simplicity and reliability, logistic regression is used in medical research (Hajifathalian et al., 2019). Decision trees sort data more intuitively and perform well with nonlinear clinical data (Zhang et al., 2017). Despite their stringent independence requirements, Naive Bayes classifiers are easy to use and handle huge datasets (Rahman and Davis, 2017). Arguments surround model complexity, interpretability, and predictive performance. New data visualisation methods have helped cardiovascular data studies. Example: ROC curves evaluate model performance (Fawcett, 2006). Selection and dimensionality reduction are also important. Jolliffe and Cadima (2016) found that PCA increases model performance by simplifying datasets. CVD categorization is being researched using neural networks. With additional data and computational power, these models can discover complex data patterns (Miotto et al., 2017). Data science can classify CVDs. The proper model depends on the dataset and clinical questions. Cardiologists must use data science for tailored therapy and better patient outcomes.

Data Description

The dataset employed in this analysis is a collection of clinical measurements aimed at predicting the presence of cardiovascular disease in individuals. This dataset, obtained from the Kaggle Machine Learning Repository, contains a total of 70,000 records, each representing an individual patient encounter.

The dataset comprises the following variables:

Table 1: Data Description

Variable	Description
Age	Reported in days, converted to years for interpretability.
Gender	Categorical variable indicating the biological sex.
Height	Measured in centimeters.
Weight	Recorded in kilograms.
Systolic Blood Pressure	Maximum arterial blood pressure during contraction of the heart.
Diastolic Blood Pressure	Minimum arterial blood pressure between heart contractions.
Cholesterol	Categorical variable with three levels (normal, above normal, well above normal).
Glucose	Categorical variable similar to cholesterol levels.
Smoking	Binary variable ('1' for smoker, '0' for non-smoker).
Alcohol Intake	Binary variable ('1' for yes, '0' for no).
Physical Activity	Binary variable indicating regular physical activity ('1' for yes, '0' for no).
Cardiovascular Disease	Binary target variable ('1' for presence, '0' for absence of disease).

Each of these variables provides insights into the patient's health status and lifestyle choices, which are critical factors in the assessment of cardiovascular risk. The dataset has been pre-processed to handle missing values and outliers, ensuring the robustness of subsequent analyses. The continuous variables such as age, height, weight, systolic, and diastolic blood pressure were examined for normality and scaled appropriately to match the scale of the binary and categorical variables. This preprocessing step is essential to avoid any potential bias in the model due to variable scales.

Exploratory Data Analysis

The Exploratory Data Analysis (EDA) used descriptive statistics and data visualisation to understand the dataset's characteristics, particularly focusing on cardiovascular variables. Bar charts were the main visualisation tool, revealing patterns, anomalies, and assumptions. These charts, colour-coded for health risk categories, helped identify skewness, modality, and outliers in cardiovascular health risk factors.

Glucose Levels: Most individuals in the dataset presented with normal glucose levels, indicative of a lower risk profile for diabetes, a common comorbidity of cardiovascular disease. A smaller but significant subset displayed well above normal levels, pointing to a potential high-risk group for further investigation.

- **Gender Distribution:** The dataset exhibited a balanced representation between genders, with a slight majority of male participants. This balance allowed for comparative analyses between sexes for cardiovascular risk factors.
- **Age Distribution:** The age of participants was concentrated in the middle-aged demographic, with fewer young or elderly individuals. This distribution suggested that the primary age group at risk for cardiovascular conditions was well-represented.
- **Height Distribution:** Heights were mostly clustered around the average, with extreme values being relatively rare, indicating that height variation was unlikely to skew the analysis of cardiovascular risk.
- **Smoking and Alcohol Consumption:** The dataset’s histograms showed a prevalence of non-smokers and non-drinkers, suggesting a lower overall risk profile based on these lifestyle factors.

- **Physical Activity:** More individuals reported engaging in regular physical activity than not, which is a positive indicator in the context of cardiovascular health.
- **Cholesterol Levels**: Normal cholesterol levels were the most common, with a smaller proportion of the dataset falling into the high and well above normal levels, necessitating attention as key risk factors for cardiovascular disease.
- **Weight and Blood Pressure Readings:** While weight distribution followed a normal-like trend, blood pressure readings displayed potential outliers. Systolic and diastolic blood pressure readings showed extreme values, indicative of potential measurement errors or data entry issues that required cleaning before further analysis.

These insights established a baseline for the risk profile of the dataset's population and facilitated the identification of areas requiring data cleansing. In particular, the blood pressure readings highlighted the need for careful data preprocessing to ensure the robustness of the subsequent predictive modelling.

Data Partitioning and Model Training

Developing prediction models requires testing them on unseen data. This was done by separating the dataset into training and testing sets. The training set teaches the model pattern identification and prediction, while the testing set impartially evaluates it. 80% of the dataset (56,000 records) was partitioned for training and 20% (14,000 records) for testing. This division was done using random sampling to guarantee that the variables and outcomes were consistent between the two sets, ensuring reliable predictive model evaluation.

Model Training

For the classification task, three algorithms were selected based on their widespread use and suitability for binary classification problems:

Logistic Regression

A fundamental statistical method for binary classification, logistic regression estimates event probability by fitting data to a logistic function. The interpretability and convenience of handling binary outcome variables is its advantages. No parameter tuning was needed beyond the default values of the R logistic regression programme.

Decision Trees

Decision trees sort instances by feature values. The nodes in the tree indicate features in an instance to be categorised, and the branches represent values the nodes can assume. The trees were created using the R package 'rpart' with a complexity parameter (cp) of 0.01 to prevent overfitting and a minimal split criteria to avoid nodes with too few instances.

Naive Bayes

This classifier uses Bayes' theorem with high feature independence assumptions. It is easy to develop and excellent for large datasets. The R 'e1071' package was used to create the model with the default Laplace smoothing parameter to handle zero frequencies in the dataset.

For each model, the respective R packages provided the tools to set the parameters, and the models were validated using 10-fold cross-validation during the training phase. This approach ensures that the models are not only fit to the training data but also have the generalization capability to perform

well on unseen data. The choice of parameters was made to balance the trade-off between model complexity and predictive power, aiming to build models that generalize well without overfitting the training data.

Model Testing and Evaluation

Upon training the logistic regression, decision trees, and Naive Bayes models, the next critical step was to evaluate their performance using the testing set. This phase involved making predictions on the testing data and comparing these predictions to the actual outcomes to assess the accuracy and effectiveness of each model.

Making Predictions

Predictions were generated using the predict() function in R, which applies the trained model parameters to the features of the testing set to output a predicted probability of cardiovascular disease occurrence for each individual. These probabilities were then converted into a binary classification by applying a threshold value, commonly set at 0.5, where probabilities above this threshold were classified as '1' (indicating the presence of CVD), and those below were classified as '0' (indicating the absence of CVD).

Evaluating Models

The evaluation of the models was conducted using several criteria that provide different insights into their performance:

- **Accuracy:** The simplest straightforward performance metric relies on the ratio of properly predicted observations to total observations. It shows how well the model classifies people with or without cardiovascular disease.

- **Sensitivity (Recall):** Sensitivity measures the percentage of correctly identified positive cases. High sensitivity for cardiovascular disease prediction means the model can identify individuals early for therapy.
- **Specificity:** Specificity quantifies the percentage of accurately detected negatives, such as healthy people who are not diagnosed with the condition. High specificity reduces false positives, which stress patients and require more testing.

The performance of the models was quantified using a confusion matrix, which lays out the true positives, false positives, true negatives, and false negatives. From this matrix, accuracy, sensitivity, and specificity were calculated.

Criteria for Model Comparison

To compare the models, we looked at the balance between sensitivity and specificity, as an ideal model would perfectly identify all positive and negative cases. However, in practice, a trade-off is often observed where increasing sensitivity lowers specificity and vice versa. The choice of the best model would then depend on the clinical context: in scenarios were missing a positive case (having CVD) could be dangerous, a higher sensitivity might be prioritized. Conversely, in screening scenarios where the cost of follow-ups is a concern, higher specificity might be favoured.

In addition to these metrics, Receiver Operating Characteristic (ROC) curves were plotted, and Area Under the Curve (AUC) was calculated for each model. The AUC provides an aggregate measure of performance across all classification thresholds. By examining these metrics collectively, the most appropriate model was selected based on the research goals and clinical requirements, ensuring a balanced approach to the prediction of cardiovascular diseases.

Results

This section evaluates cardiovascular disease predictive models in detail. Three models were created: Logistic Regression, Decision Tree, and Naive Bayes after dataset preprocessing and exploratory analysis. Each model was trained, tested, and validated to ensure predictive accuracy, sensitivity to true positives, and specificity to true negatives. A confusion matrix for each model showed true positive, false positive, true negative, and false negative rates. Accuracy, sensitivity (recall), and specificity were calculated from these rates. These metrics allowed us to compare and find the best CVD prediction model for the clinical context.

Table 2: Model Evaluation Results

Model	Accuracy	Sensitivity	Specificity
Logistic Regression	0.7223	0.7738	0.6708
Decision Tree	0.7195	0.7895	0.6493
Naive Bayes	0.5799	0.9640	0.1953

From the presented results, the Logistic Regression model appears to be the most balanced in terms of accuracy, sensitivity, and specificity, indicating its effectiveness as a classifier for cardiovascular disease. While the Naive Bayes model shows the highest sensitivity, its specificity is significantly lower, suggesting a high rate of false positives. The Decision Tree model demonstrates comparable accuracy to the Logistic Regression model but with slightly higher sensitivity and lower specificity. Given these outcomes, the Logistic Regression model was selected for further analysis, including ROC curve assessment and cross-validation which are presented in the appendix, to confirm its robustness and generalizability to unseen data.

Discussion

The results from the Logistic Regression, Decision Tree, and Naive Bayes models revealed the predictability of cardiovascular diseases from clinical data. The Logistic Regression model

emerged as the most balanced, demonstrating the importance of both accuracy and the ability to correctly identify true cases of the disease. The superior sensitivity of the Naive Bayes model suggests it is less likely to miss cases of CVD, yet its low specificity could lead to a high rate of false positives, potentially causing undue stress and additional medical procedures for patients. The Decision Tree's performance was commendable but did not surpass Logistic Regression in overall accuracy. These outcomes underscore the complexity of predicting health outcomes and the need to consider the trade-offs between different performances metrics.

Conclusion

The analysis presented herein applied data science techniques to the pressing issue of cardiovascular disease classification. The key finding is that Logistic Regression holds promise as a tool for aiding medical professionals in the prediction of CVD. This study answers the research question by demonstrating that clinical variables can indeed be used to predict cardiovascular disease with a reasonable degree of accuracy. Future work should focus on incorporating more diverse datasets, exploring more complex models such as ensemble methods or deep learning, and applying feature selection techniques to improve model performance and interpretability.

REFERENCES

Ahmad, T., Lund, L.H., Rao, P., Ghosh, R., Warier, P., Vaccaro, B., Dahlström, U., O'Connor, C.M., Felker, G.M. and Desai, N.R., (2018). Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal of the American Heart Association*, *7*(8), p.e008081.

Al'Aref, S.J., Anchouche, K., Singh, G., Slomka, P.J., Kolli, K.K., Kumar, A., Pandey, M., Maliakal, G., Van Rosendael, A.R., Beecy, A.N. and Berman, D.S., (2019). Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European heart journal*, *40*(24), pp.1975-1986.

Ambale-Venkatesh, B., & Lima, J. A. (2015). Cardiac MRI: a central prognostic tool in myocardial fibrosis. Nature Reviews Cardiology, 12(1), 18–29.

Ambale-Venkatesh, B., Yang, X., Wu, C.O., Liu, K., Hundley, W.G., McClelland, R., Gomes, A.S., Folsom, A.R., Shea, S., Guallar, E. and Bluemke, D.A., (2017). Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. *Circulation research*, *121*(9), pp.1092-1101.

Baharvand, B.A., Bahmani, M. and Zargaran, A., (2016). A brief report of rhazes manuscripts in the field of cardiology and cardiovascular diseases.

Benjamin, E.J., Muntner, P., Alonso, A., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Das, S.R. and Delling, F.N., (2019). Heart disease and stroke statistics—2019 update: a report from the American Heart Association. *Circulation*, *139*(10), pp.e56-e528.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874.

Fawcett, T., (2016). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), pp.861-874.

García, S., Luengo, J. and Herrera, F., (2015). *Data preprocessing in data mining* (Vol. 72, pp. 59-139). Cham, Switzerland: Springer International Publishing.

Hajifathalian, K., Ueda, P., Lu, Y., et al. (2019). A novel risk score to predict cardiovascular disease risk in national populations (Globorisk): a pooled analysis of prospective cohorts and health examination surveys. The Lancet Diabetes & Endocrinology, 3(5), 339–355.

James, G., Witten, D., Hastie, T. and Tibshirani, R., (2023). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Khera, A.V., Emdin, C.A., Drake, I., Natarajan, P., Bick, A.G., Cook, N.R., Chasman, D.I., Baber, U., Mehran, R., Rader, D.J. and Fuster, V., (2016). Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*, *375*(24), pp.2349-2358.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2020). Artificial Intelligence in Precision Cardiovascular Medicine. Journal of the American College of Cardiology, 69(21), 2657–2664.

Kuhn, M., (2018). Building predictive models in R using the caret package. *Journal of statistical software*, *28*, pp.1-26.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246.

Patel, B. and Sengupta, P., (2020). Machine learning for predicting cardiac events: what does the future hold?. *Expert review of cardiovascular therapy*, *18*(2), pp.77-84.

Rahman, M. M., & Davis, D. N. (2017). Machine learning-based missing value imputation method for clinical datasets. AI in Medicine, 83, 11–22.

Rish, I., (2021), August. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46).

Schüssler-Fiorenza Rose, S.M., Contrepois, K., Moneghetti, K.J., Zhou, W., Mishra, T., Mataraso, S., Dagan-Rosenfeld, O., Ganz, A.B., Dunn, J., Hornburg, D. and Rego, S., (2019). A longitudinal big data approach for precision health. *Nature medicine*, *25*(5), pp.792-804.

Smith Jr, S.C., Collins, A., Ferrari, R., Holmes Jr, D.R., Logstrup, S., McGhie, D.V., Ralston, J., Sacco, R.L., Stam, H., Taubert, K. and Wood, D.A., (2022). Our time: a call to save preventable death from cardiovascular disease (heart disease and stroke). *Circulation*, *126*(23), pp.2769-2775.

Steyerberg, E.W., Uno, H., Ioannidis, J.P., Van Calster, B., Ukaegbu, C., Dhingra, T., Syngal, S. and Kastrinos, F., (2018). Poor performance of clinical prediction models: the harm of commonly applied methods. *Journal of clinical epidemiology*, *98*, pp.133-143.

Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data?. *PloS one*, *12*(4), p.e0174944.

World Health Organization, (2019). Health topics: Cardiovascular diseases. *Fact Sheet. Available online: http://www. who. int/cardiovascular_diseases/en/(accessed on 11 December 2020)*.

Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., & Lyashevsky, A. (2017). Predictive analytics with gradient boosting in clinical medicine. Annals of Translational Medicine, 7(7), 152.

Appendix A

Loading the required library

```
library(readr)

## Warning: package 'readr' was built under R version 4.2.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✓ dplyr     1.1.4     ✓ purrr     1.0.2
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors

library(psych)

## Warning: package 'psych' was built under R version 4.2.3

##
## Attaching package: 'psych'
```

```
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.2.3
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.2.3
```

```
library(caret)
library(naivebayes)
```

```
## Warning: package 'naivebayes' was built under R version 4.2.3
```

```
## naivebayes 0.9.7 loaded
```

```
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.2.3
```

Import the dataset

```
cardio <- read_delim("C:/Users/ROBERTECH/Desktop/Santa/cardio_train.csv", del
im = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
## Rows: 70000 Columns: 13
## ── Column specification ────────────────────────────────────────────────────
──────
```

```
## Delimiter: ";"
## dbl (13): id, age, gender, height, weight, ap_hi, ap_lo, cholesterol, gluc
, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this m
essage.
```

```
# Dataset Preprocessing
table(cardio$cardio)
```

```
##
##     0     1
## 35021 34979
```

```
cardio = na.omit(cardio)
nrow(cardio)
```

```
## [1] 70000
```

```
ncol(cardio)
```

```
## [1] 13
```

```
str(cardio)
```

```
## tibble [70,000 × 13] (S3: tbl_df/tbl/data.frame)
##  $ id         : num [1:70000] 0 1 2 3 4 8 9 12 13 14 ...
##  $ age        : num [1:70000] 18393 20228 18857 17623 17474 ...
##  $ gender     : num [1:70000] 2 1 1 2 1 1 1 2 1 1 ...
##  $ height     : num [1:70000] 168 156 165 169 156 151 157 178 158 164 ...
##  $ weight     : num [1:70000] 62 85 64 82 56 67 93 95 71 68 ...
##  $ ap_hi      : num [1:70000] 110 140 130 150 100 120 130 130 110 110 ...
##  $ ap_lo      : num [1:70000] 80 90 70 100 60 80 80 90 70 60 ...
##  $ cholesterol: num [1:70000] 1 3 3 1 1 2 3 3 1 1 ...
##  $ gluc       : num [1:70000] 1 1 1 1 1 2 1 3 1 1 ...
##  $ smoke      : num [1:70000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ alco       : num [1:70000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ active     : num [1:70000] 1 1 0 1 0 0 1 1 1 0 ...
##  $ cardio     : num [1:70000] 0 1 1 1 0 0 0 1 0 0 ...
```

```
cardio$gender <- factor(cardio$gender, levels = c(1, 2), labels = c("Male", "
Female"))
cardio$cholesterol <- factor(cardio$cholesterol, levels = c(1, 2, 3), labels
= c("Normal", "Above Normal", "Well Above Normal"))
cardio$gluc <- factor(cardio$gluc, levels = c(1, 2, 3), labels = c("Normal",
"Above Normal", "Well Above Normal"))
cardio$smoke <- factor(cardio$smoke, levels = c(0, 1), labels = c("No", "Yes"
))
cardio$alco <- factor(cardio$alco, levels = c(0, 1), labels = c("No", "Yes"))
cardio$active <- factor(cardio$active, levels = c(0, 1), labels = c("No", "Ye
s"))
```

```
cardio$cardio <- factor(cardio$cardio, levels = c(0, 1), labels = c("No", "Ye
s"))
```

#Exploratory Data Analysis

```
par(mfrow=c(2,2))  # Adjust the layout for the plots

# Smoking Habit
smoke = table(cardio$smoke)
barplot(smoke, ylab = "Frequency", xlab="Smoke", main="Smoking Habit", col=c(
"green", "red"))

# Alcohol Consumption
alco = table(cardio$alco)
barplot(alco, ylab = "Frequency", xlab="Alcohol", main="Alcohol Consumption",
col=c("green", "red"))

# Physical Activity
active = table(cardio$active)
barplot(active, ylab = "Frequency", xlab="Active", main="Physical Activity",
col=c("green", "red"))

# Cholesterol Levels
cholesterol = table(cardio$cholesterol)
barplot(cholesterol, ylab = "Frequency", xlab="Cholesterol", main="Cholestero
l Levels", col=c("green", "yellow", "red"))
```

Test Models and Compute Confusion Matrix

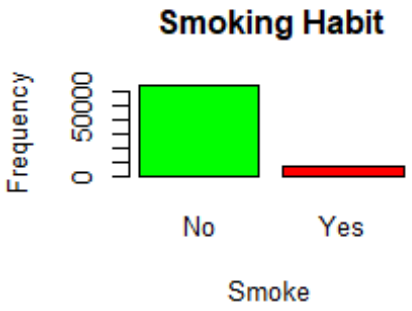
```
predictions_log <- predict(model_log, testing)
conf_matrix_log <- confusionMatrix(predictions_log, testing$cardio)

predictions_tree <- predict(model_tree, testing)
conf_matrix_tree <- confusionMatrix(predictions_tree, testing$cardio)

predictions_nb <- predict(model_nb, testing)
conf_matrix_nb <- confusionMatrix(predictions_nb, testing$cardio)
```

Compile Evaluation Metrics

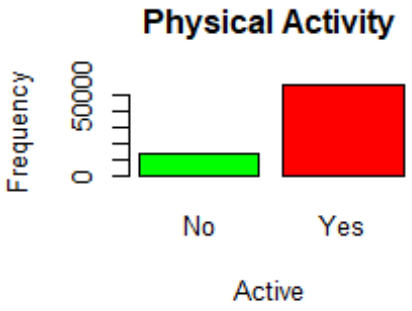
```
results <- data.frame(
  Model = c("Logistic Regression", "Decision Tree", "Naive Bayes"),
  Accuracy = c(conf_matrix_log$overall["Accuracy"], conf_matrix_tree$overall[
"Accuracy"],
               conf_matrix_nb$overall["Accuracy"]),
  Sensitivity = c(conf_matrix_log$byClass["Sensitivity"], conf_matrix_tree$by
Class["Sensitivity"],
                  conf_matrix_nb$byClass["Sensitivity"]),
  Specificity = c(conf_matrix_log$byClass["Specificity"], conf_matrix_tree$by
Class["Specificity"],
```

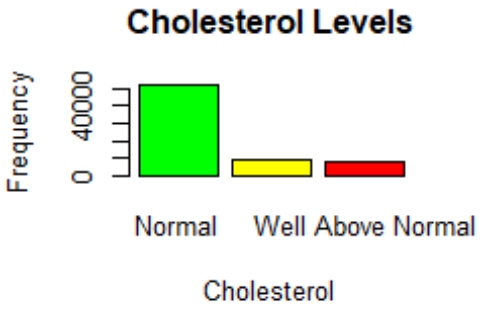
```
                  conf_matrix_nb$byClass["Specificity"])
)

print(results)

##                 Model  Accuracy Sensitivity Specificity
## 1 Logistic Regression 0.7223373   0.7738435   0.6707648
## 2       Decision Tree 0.7194800   0.7895488   0.6493209
## 3         Naive Bayes 0.5798986   0.9640206   0.1952823
```

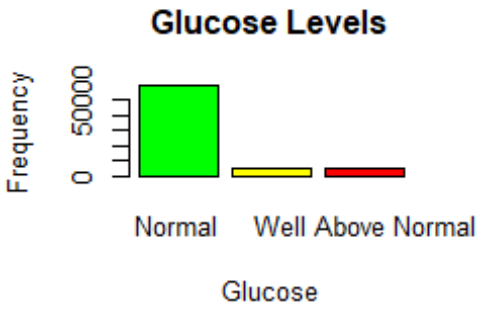
Appendix B

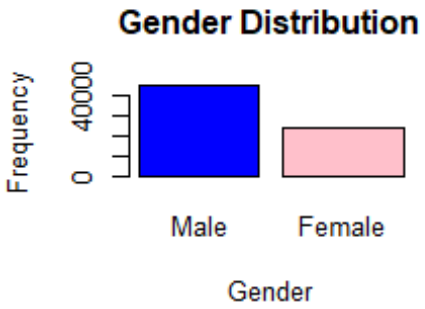

Smoking Habit
Frequency
0
50000
No
Yes
Smoke

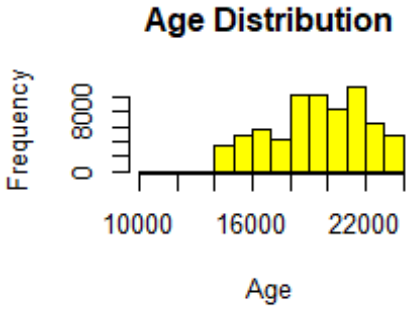

Alcohol Consumption
Frequency
0
50000
No
Yes
Alcohol

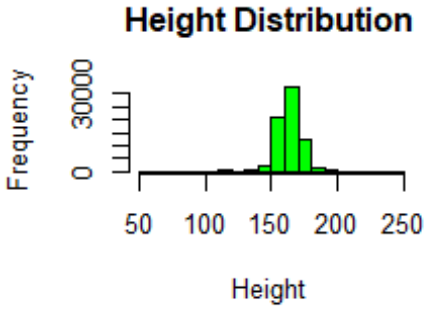

Physical Activity
Frequency
0
50000
No
Yes
Active

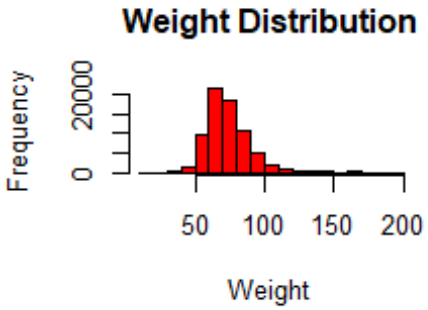

Cholesterol Levels
Frequency
0
40000
Normal
Well Above Normal
Cholesterol

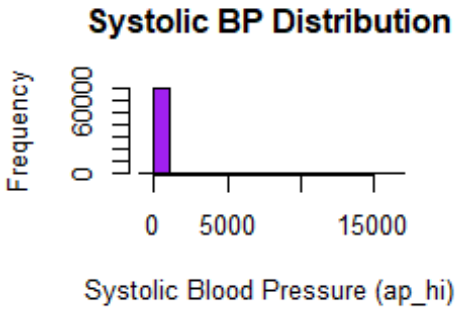

Glucose Levels
Frequency
0
50000
Normal
Well Above Normal
Glucose

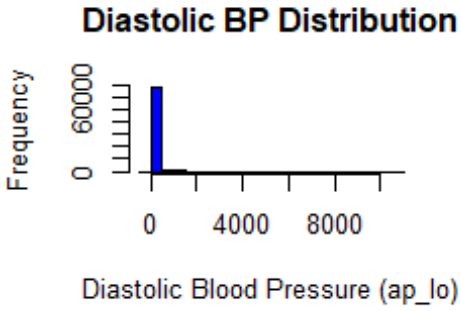

Gender Distribution
Frequency
0
40000
Male
Female
Gender


Age Distribution
Frequency
0
8000
10000
16000
22000
Age

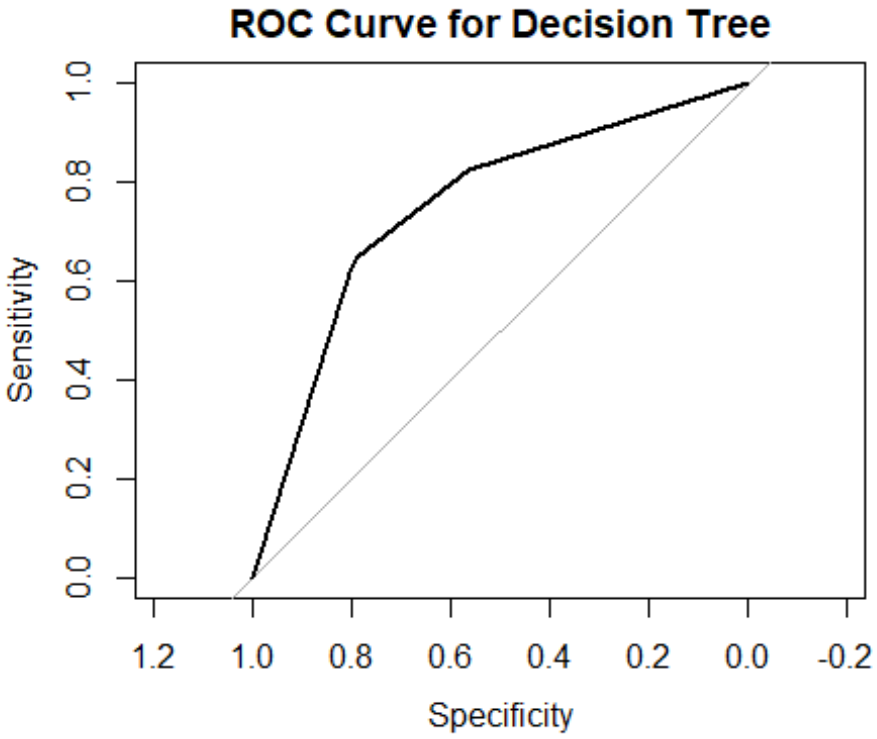

Height Distribution
Frequency
0
30000
50
100
150
200
250
Height

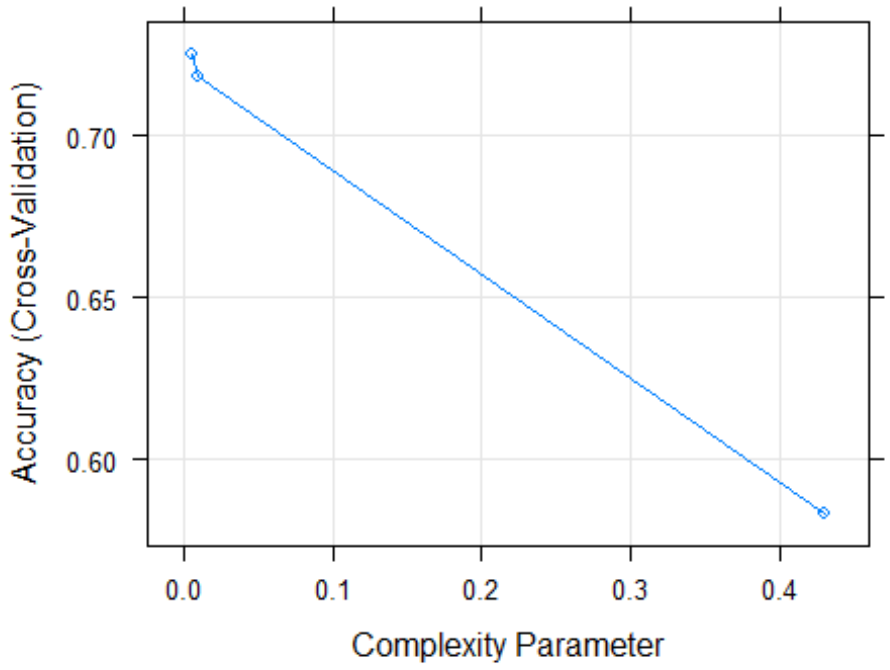







```
##        id              age            gender          height     
##  Min.   :    0   Min.   :10798   Male  :45530   Min.   : 55.0  
##  1st Qu.:25007   1st Qu.:17664   Female:24470   1st Qu.:159.0  
##  Median :50002   Median :19703                  Median :165.0  
##  Mean   :49972   Mean   :19469                  Mean   :164.4  
##  3rd Qu.:74889   3rd Qu.:21327                  3rd Qu.:170.0  
##  Max.   :99999   Max.   :23713                  Max.   :250.0  
##      weight           ap_hi             ap_lo          
##  Min.   : 10.00   Min.   : -150.0   Min.   :  -70.00  
##  1st Qu.: 65.00   1st Qu.:  120.0   1st Qu.:   80.00  
##  Median : 72.00   Median :  120.0   Median :   80.00  
##  Mean   : 74.21   Mean   :  128.8   Mean   :   96.63  
##  3rd Qu.: 82.00   3rd Qu.:  140.0   3rd Qu.:   90.00  
##  Max.   :200.00   Max.   :16020.0   Max.   :11000.00  
##             cholesterol                    gluc       smoke        alco        
##  Normal           :52385   Normal           :59479   No :63831   No :66236  
##  Above Normal     : 9549   Above Normal     : 5190   Yes: 6169   Yes: 3764  
##  Well Above Normal: 8066   Well Above Normal: 5331
```


ROC Curve for Decision Tree
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
Specificity


Accuracy (Cross-Validation)
0.60
0.65
0.70
0.0
0.1
0.2
0.3
0.4
Complexity Parameter