Statistical Data Mining Project

**Prediction of Cardiovascular Disease**

Exploring Predictors of CVD Risk

By: Yesha Desai

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**



**Index**

|  |  |  |
| --- | --- | --- |
| Sr no. | Topic | Page |
| 1. | Executive Summary | 3 |
| 2. | Problem Statement and significance | 4 |
| 3. | Prior Literature | 4 |
| 4. | Data Source and Preparation | 5 |
| 5. | Variable Choice | 6 |
| 6. | Exploratory Data Analysis and Visualizations | 7 |
| 7. | Models | 9 |
| 8. | Quality check | 11 |
| 9. | Recommendations | 11 |
| 10. | References | 12 |
| 11. | Appendix | 13 |

**Executive Summary**

The project aimed to predict cardiovascular disease, a leading cause of global mortality, using various medical indicators. Key findings reveal significant predictors such as age, weight, blood pressure, cholesterol, and glucose levels. Individuals aged 50 and above are particularly vulnerable to cardiovascular disease, as indicated by the data. Furthermore, lifestyle factors like smoking and alcohol intake contribute to increased disease risk.

Three models – Logistic Regression, Probit Regression, and Complementary Log-Log Regression were employed for prediction. While all models provided valuable insights, the Probit Regression model emerged as the most effective, exhibiting high accuracy and precision. Notably, age, weight, and blood pressure consistently emerged as significant predictors across all models, underlining their importance in assessing cardiovascular disease risk.

Quality checks, including residual analysis, linearity, homoscedasticity, and multicollinearity assessments, validated the robustness of the models. Additionally, performance metrics such as accuracy, precision, recall, F1 score, and area under the curve were calculated, further supporting the reliability of the Probit Regression model.

Recommendations based on the analysis emphasized the importance of maintaining healthy lifestyle habits, such as weight management, regular exercise, and monitoring blood pressure and cholesterol levels. These proactive measures can significantly reduce the risk of cardiovascular disease and promote overall heart health.

The findings provide valuable insights for healthcare providers, public health organizations, and medical researchers in devising preventive strategies and interventions to combat cardiovascular diseases effectively.

**Problem Definition and Significance**

Detecting and managing risk factors early is crucial to mitigate the impact of cardiovascular diseases, which rank among the leading causes of death worldwide. Accurately predicting the probability of cardiovascular disease using diverse factors can take quick actions and prevention measures.

Heart disease persists as the leading global cause of death, with approximately 17.9 million deaths associated to cardiovascular diseases in 2019, representing 32% of all global death. Out of which, 85% resulted from heart attacks and strokes.

The target clients for this project are Healthcare Providers, Public Health Organizations and Medical Researchers.

**Prior Literature**

|  |  |  |
| --- | --- | --- |
| Topic | Methodology | Insights |
| Efficient Heart Disease Prediction System | The dataset is preprocessed using KEEL tool (Knowledge Extraction based on Evolutionary Learning) and hill climbing algorithm. Then, the rules are classified into different sets, evaluating the system's performance through 10-fold cross-validation. | The outcome showed that the Efficient Heart Disease Prediction System achieved a higher accuracy rate (86.7%) compared to other well-known classifiers such as SVM, C4.5, NN, PART, MLP, RBF, and TSEAFS, as demonstrated by the evaluation on the heart disease dataset. |
| Heart Disease Prediction using Logistic Regression | This paper has used two methods: binary logistic regression and robust methods with BLR to predict heart disease. Robust methods: least quartile difference and median absolute deviation handled outliers and multicollinearity. | The comparison of percentage accuracy among all models was done and it showed that the binary logistic model with the applied MAD method achieved the highest accuracy at 86.6%. Hence, this model was selected as their best among the others. |

**Data Source**

The dataset utilized in this analysis originates from a cardiovascular examination, consisting a variety of medical indicators:

* **Age**: Measured in days, converted to years for better interpretability.
* **Height**: Recorded in centimeters.
* **Weight**: Captured in kilograms.
* **Gender**: Categorized as binary codes.
* **Systolic Blood Pressure (ap\_hi)**: Denoting the pressure in arteries during heartbeats.
* **Diastolic Blood Pressure (ap\_lo):** Representing the pressure in arteries between heartbeats.
* **Cholesterol**: Classified into three levels: normal, above normal, and well above normal.
* **Glucose**: Similar to cholesterol, classified into three levels.
* **Smoking**: A binary indicator denoting smoking habits.
* **Alcohol Intake:** Another binary indicator indicating alcohol consumption.
* **Physical Activity**: Yet another binary indicator representing engagement in physical activities.
* **Presence or Absence of Cardiovascular Disease**: The target variable indicating the existence or absence of cardiovascular disease.

**Data Preparation**

Data Preparation included several steps to ensure the dataset's suitability for analysis:

* Dropping 'id' Column: Since it merely served as an identifier, it was removed to avoid redundancy.
* Age Conversion: Transforming age from days to years enhances interpretability.
* Duplicate Removal: Eliminating duplicate entries ensures each record is unique, preventing redundancy in the analysis.
* Handling Missing Data: The dataset was complete, without any instances of missing values, ensuring the analysis's reliability and model training's accuracy.

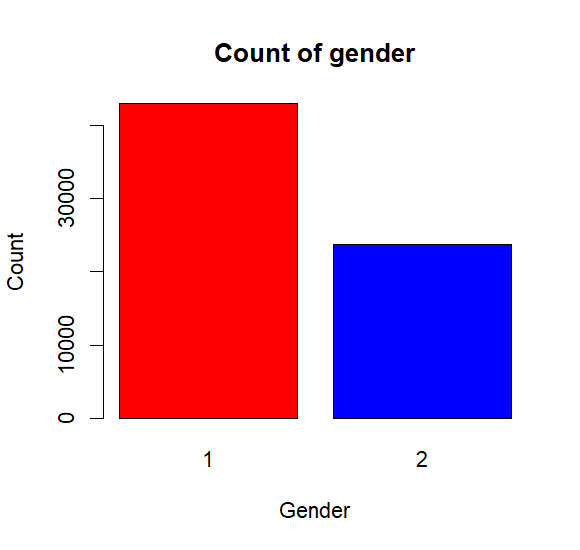
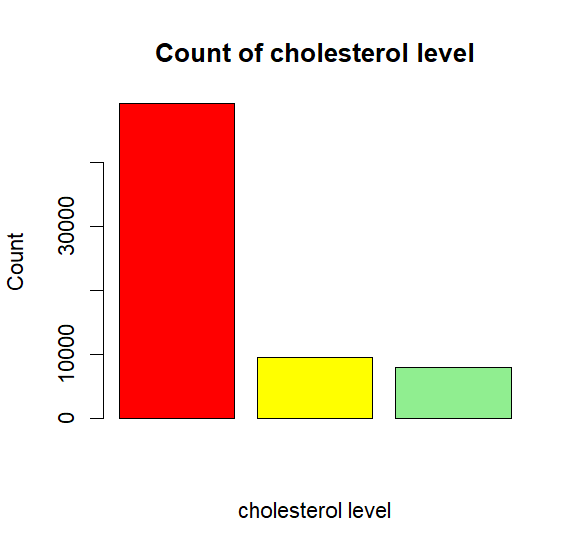
For analysis, all variables like: Age, Height, Weight, Gender, Systolic and Diastolic Blood Pressure, Cholesterol, Glucose, Smoking, Alcohol Intake, and Physical Activity were considered as Independent Variables and the presence or absence of cardiovascular disease was selected as the Dependent Variable. This approach allows for a thorough examination of how various factors contribute to the likelihood of cardiovascular disease occurrence.

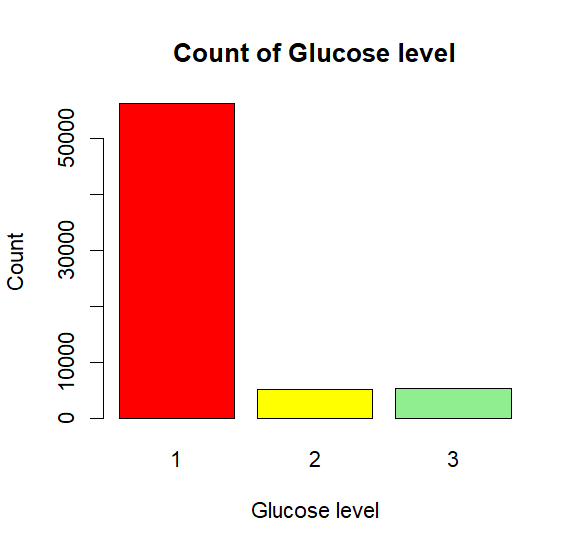
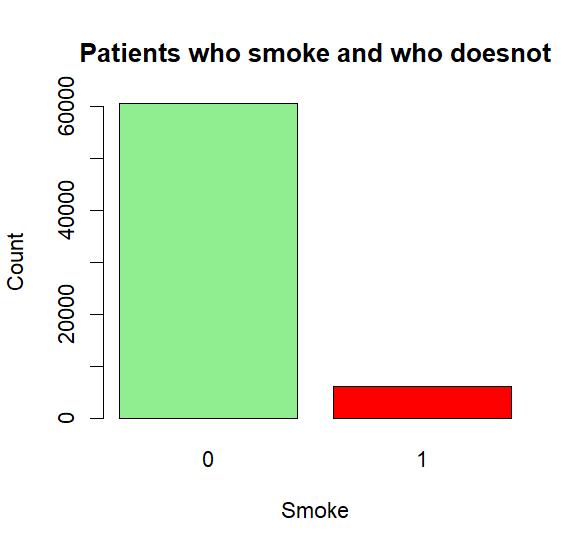
**Variable choice**

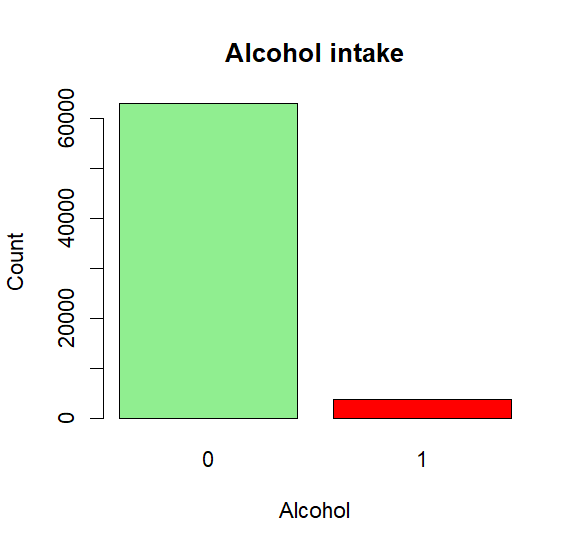
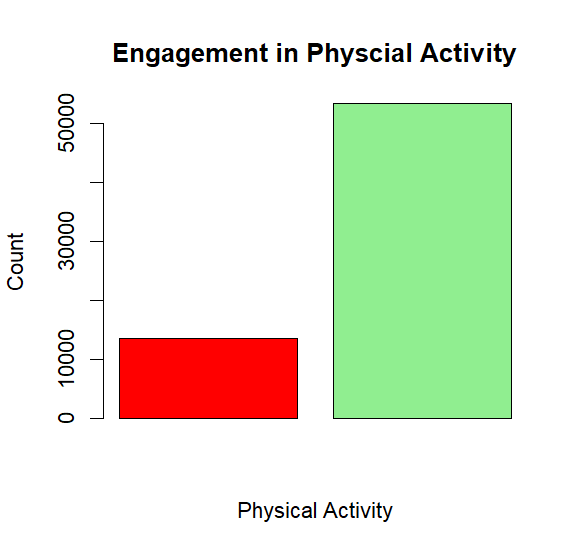
|  |  |  |
| --- | --- | --- |
| Predictor | Effect | Rationale |
| Age | + | Age is a well-established risk factor for cardiovascular disease. |
| Height | +/- | Height may correlate with cardiovascular health, as extremes may indicate underlying issues. |
| Weight | + | Weight is a crucial factor in cardiovascular health and risk assessment. |
| Gender | +/- | Gender differences may influence cardiovascular risk factors and disease prevalence. |
| Systolic blood pressure | + | Elevated systolic blood pressure is a significant risk factor for cardiovascular disease. |
| Diastolic blood pressure | + | Elevated diastolic blood pressure is also indicative of cardiovascular risk. |
| Cholesterol | + | High cholesterol levels are associated with increased cardiovascular risk. |
| Glucose | + | Elevated glucose levels can indicate diabetes, a significant risk factor for cardiovascular disease. |
| Alcohol intake | +/- | Excessive alcohol intake can contribute to cardiovascular risk. |
| Smoking | + | Smoking is a well-known risk factor for cardiovascular disease. |
| Physical Activity | - | Regular physical activity is associated with reduced cardiovascular risk. |

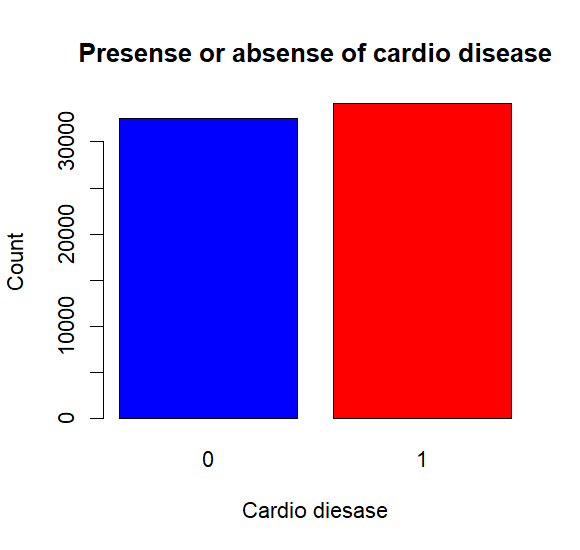
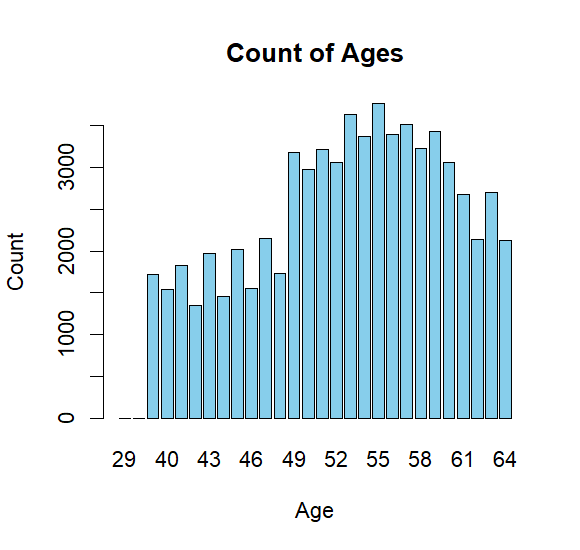
**Descriptive Analysis & Data Visualizations**

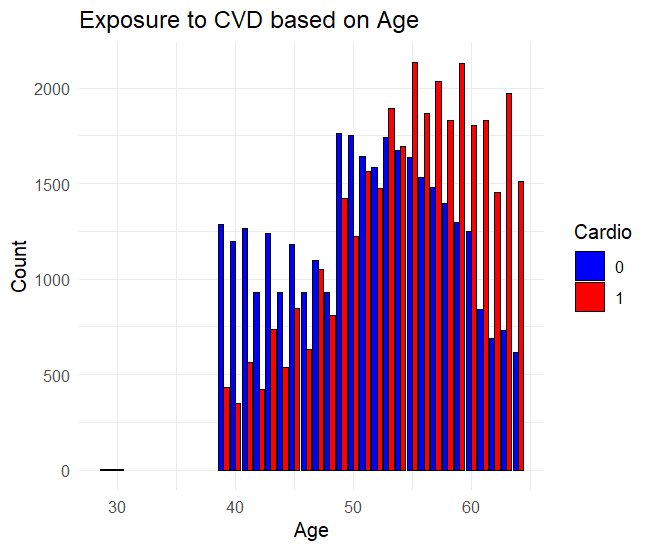
The below visualizations give a pictorial representation of independent variables of the datasets and its distribution count.



This visualization gives a descriptive analysis of, at what age patients are exposed to cardiovascular disease. And as it is seen that bar lines in red denotes presence of cardio vascular disease in patients age ranging from 50 to age 65 are having the disease, while, age 40 to 50 does not have the disease as the bar lines are in blue. This explains that patients of age 50 and above are more likely to get exposed with the cardiovascular disease.

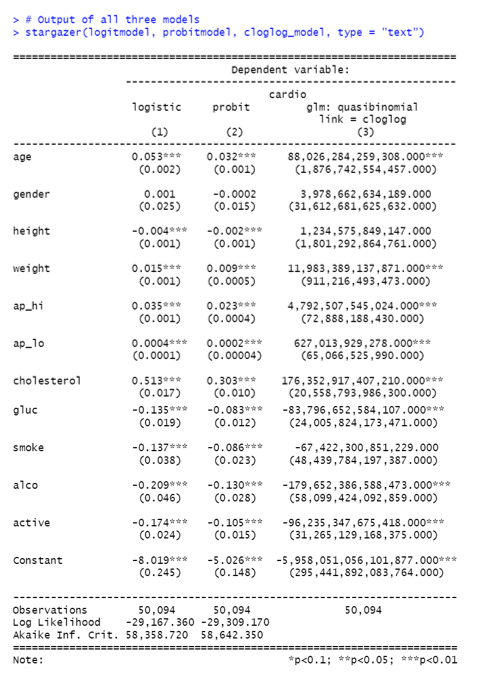
**Models**

* Logistic Regression Model
* Probit Regression Model
* Complementary Log Log Model

I have used these three models for the predicting the variables which would give result of predicting which variables are more likely cause of the cardiovascular disease.

* **Logistic Regression Model:** is a standard choice for binary classification problems like predicting cardiovascular disease presence or absence, and the dependent variable represents binary outcomes. Therefore, logistic regression is a good choice for the prediction.
* **Probit Regression Model:** is another commonly used method for binary classification, especially when the assumption of a logistic distribution is not met and it provides flexibility in modeling the relationship between the predictors.
* **Complementary Log Log Model:** is also suitable for binary outcomes and offers flexibility.

**Outcome**:



**Result interpretation:**

**Logistic Regression Model:**

* Positive coefficients for age, it is statistically significant that for each unit increase in age, there is a 0.053 unit increase in the odds of cardiovascular disease.
* Positive coefficients for weight, it is statistically significant that for each unit increase in weight, there is a 0.015 unit increase in the odds of cardiovascular disease.
* Positive coefficients for systolic blood pressure, it is statistically significant that for each unit increase in systolic blood pressure, there is a 0.035 unit increase in the odds of cardiovascular disease.
* Positive coefficients for cholesterol, it is statistically significant that for each unit increase in cholesterol, there is a 0.513 unit increase in the odds of cardiovascular disease.
* Conversely, statistically significant negative coefficients for height, diastolic blood pressure (ap\_lo), glucose levels, alcohol intake, and physical activity indicate that higher values of these variables are associated with decreased odds of cardiovascular disease.

**Probit Regression Model:**

* Positive coefficients for age, it is statistically significant that for each unit increase in age, there is a 0.032 unit increase in the odds of cardiovascular disease.
* Positive coefficients for weight, it is statistically significant that for each unit increase in weight, there is a 0.009 unit increase in the odds of cardiovascular disease.
* Positive coefficients for systolic blood pressure, it is statistically significant that for each unit increase in systolic blood pressure, there is a 0.023 unit increase in the odds of cardiovascular disease.
* Positive coefficients for diastolic blood pressure, it is statistically significant that for each unit increase in diastolic blood pressure, there is a 0.002 unit increase in the odds of cardiovascular disease.
* Positive coefficients for cholesterol, it is statistically significant that for each unit increase in cholesterol, there is a 0.303 unit increase in the odds of cardiovascular disease.
* Conversely, significant negative coefficients for height, diastolic blood pressure (ap\_lo), glucose levels, alcohol intak, and physical activity suggest a lower probability of cardiovascular disease with higher values of these variables.

**Complementary Log-Log Regression Model:**

* Positive coefficients statistically infer that an increase in age, weight, systolic blood pressure (ap\_hi), cholesterol, and smoking status results in a greater rate of change in the probability of cardiovascular disease occurrence.
* Conversely, significant negative coefficients for height, diastolic blood pressure (ap\_lo), glucose levels, alcohol intake, and physical activity suggest a decrease in the rate of change in the probability of cardiovascular disease with higher values of these variables.

**Quality Checks**

I performed assumption tests like residual check, linearity check, Homoscadascity and multicolinearity check on all of the three models which are Logistic Regression, Probit Model and Complementary log log Model.

Created Confusion Matrix and Checked Accuracy, Precision, Recall, F1 score, and Area under curve for all three models:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic Model | Probit Model | Complementary Log Log Model |
| Accuracy | **71.7%** | **71.9%** | **61%** |
| Precision | **73.8%** | **74%** | **58%** |
| Recall | **69%** | **69.3%** | **86%** |
| AUC | **0.711** | **0.71** | **0.60** |

* Based on the performance metrics and interpretations, the probit model appears to be the best model among the three, as it achieves the highest accuracy and precision while maintaining balanced recall.

**Recommendations**

* Factors such as weight, blood pressure (both systolic and diastolic), cholesterol level, and glucose level are significant predictors of cardiovascular disease.
* Individuals should strive to maintain a healthy weight, monitor their blood pressure regularly, and keep their cholesterol and glucose levels within recommended ranges.
* Being active is negatively associated with the likelihood of cardiovascular disease. Engaging in regular physical activity can help improve overall cardiovascular health and reduce the risk of developing heart-related conditions.

**References**

* https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data
* https://penerbit.uthm.edu.my/periodicals/index.php/ekst/article/view/2163
* https://www.sciencedirect.com/science/article/pii/S187705091630638X

**Appendix**

**R code:**

# Yesha Desai

rm(list=ls())

setwd("C:/USF BAIS/sem2/SDM")

library(rio)

library(moments)

library(dplyr)

library(ggplot2)

library(caret)

library(stargazer)

library(pROC)

cardio=import("cardio\_train.csv")

colnames(cardio)=tolower(make.names(colnames(cardio)))

attach(cardio)

names(cardio)

str(cardio)

# Data cleaning

# Drop the 'id' column

cardio <- cardio %>% select(-id)

# Convert Age from days to years

cardio$age <- cardio$age / 365

# Convert age from years to whole numbers

cardio$age <- floor(cardio$age)

# Check for and remove duplicate rows

cardio <- cardio[!duplicated(cardio),

]

# Check for missing values

if (sum(is.na(cardio)) > 0) {

print("The dataset contains missing values.")

} else {

print("The dataset is complete, with no missing values.")

}

cardio

# Data visualizations

# Count occurrences of each age

age\_counts <- table(cardio$age)

# Convert the age\_counts table to a data frame for plotting

age\_counts\_df <- as.data.frame(age\_counts)

names(age\_counts\_df) <- c("Age", "Count")

# Plot

barplot(age\_counts\_df$Count,

names.arg = age\_counts\_df$Age,

xlab = "Age",

ylab = "Count",

col= 'skyblue',

main = "Count of Ages")

# Count occurrences of each gender

gender\_counts <- table(cardio$gender)

# Convert the age\_counts table to a data frame for plotting

gender\_counts\_df <- as.data.frame(gender\_counts)

names(gender\_counts\_df) <- c("gender", "count")

# Plot

barplot(gender\_counts\_df$count,

names.arg = gender\_counts\_df$gender,

xlab = "Gender",

ylab = "Count",

col= c('red','blue'),

main = "Count of gender")

# Count occurrences of cholesterol

cholesterol\_counts <- table(cardio$cholesterol)

# Convert the age\_counts table to a data frame for plotting

cholesterol\_counts\_df <- as.data.frame(cholesterol\_counts)

names(cholesterol\_counts\_df) <- c("cholesterol", "count")

# Plot

barplot(cholesterol\_counts\_df$count,

names.arg = gender\_counts\_df$cholesterol,

xlab = "cholesterol level",

ylab = "Count",

col= c('red','yellow','lightgreen'),

main = "Count of cholesterol level")

# Count occurrences of glucose

glucose\_counts <- table(cardio$gluc)

# Convert the age\_counts table to a data frame for plotting

glucose\_counts\_df <- as.data.frame(glucose\_counts)

names(glucose\_counts\_df) <- c("glucose", "count")

# Plot

barplot(glucose\_counts\_df$count,

names.arg = glucose\_counts\_df$gluc,

xlab = "Glucose level",

ylab = "Count",

col= c('red','yellow','lightgreen'),

main = "Count of Glucose level")

# Count occurrences of smoke

smoke\_counts <- table(cardio$smoke)

# Convert the age\_counts table to a data frame for plotting

smoke\_counts\_df <- as.data.frame(smoke\_counts)

names(smoke\_counts\_df) <- c("smoke", "count")

# Plot

barplot(smoke\_counts\_df$count,

names.arg = smoke\_counts\_df$smoke,

xlab = "Smoke",

ylab = "Count",

col= c('lightgreen', 'red'),

main = "Patients who smoke and who doesnot")

# Count occurrences of alcohol

alcohol\_counts <- table(cardio$alco)

# Convert the age\_counts table to a data frame for plotting

alcohol\_counts\_df <- as.data.frame(alcohol\_counts)

names(alcohol\_counts\_df) <- c("alcohol", "count")

# Plot

barplot(alcohol\_counts\_df$count,

names.arg = alcohol\_counts\_df$alco,

xlab = "Alcohol",

ylab = "Count",

col= c('lightgreen', 'red'),

main = "Alcohol intake")

# Count occurrences of physical activity

pa\_counts <- table(cardio$active)

# Convert the age\_counts table to a data frame for plotting

pa\_counts\_df <- as.data.frame(pa\_counts)

names(pa\_counts\_df) <- c("pa", "count")

# Plot

barplot(pa\_counts\_df$count,

names.arg = pa\_counts\_df$active,

xlab = "Physical Activity",

ylab = "Count",

col= c('red', 'lightgreen'),

main = "Engagement in Physcial Activity")

# Count occurrences of cardio

cardio\_counts <- table(cardio$cardio)

# Convert the age\_counts table to a data frame for plotting

cardio\_counts\_df <- as.data.frame(cardio\_counts)

names(cardio\_counts\_df) <- c("cardio", "count")

# Plot

barplot(cardio\_counts\_df$count,

names.arg = cardio\_counts\_df$cardio,

xlab = "Cardio diesase",

ylab = "Count",

col= c( 'blue', 'red'),

main = "Presense or absense of cardio disease")

# Set the figure size

options(repr.plot.width = 11, repr.plot.height = 8)

# Create the count plot

ggplot(cardio, aes(x = age, fill = factor(cardio))) +

geom\_bar(position = "dodge", color = "black") +

labs(x = "Age", y = "Count", fill = "Cardio") +

scale\_fill\_manual(values = c("blue", "red")) +

theme\_minimal() +

ggtitle("Exposure to CVD based on Age")

# Model 1- Logistic regression with train and test data

set.seed(1234)

# Splitting data into training and test sets

train\_cardio <- createDataPartition(cardio$cardio, p = 0.75, list = FALSE)

traind\_cardio <- cardio[train\_cardio, ]

testd\_cardio <- cardio[-train\_cardio, ]

# logistic regression model

logitmodel <- glm(cardio ~ age+ gender+ height+ weight+ ap\_hi+ ap\_lo+

cholesterol+ gluc+ smoke+ alco+ active, data = traind\_cardio, family = "binomial")

summary(logitmodel)

# Evaluation on the test dataset

cardio\_pred <- predict(logitmodel, newdata = testd\_cardio, type = "response")

cardio\_pred <- ifelse(cardio\_pred>0.5,1,0)

cardio\_pred

# Model 2- Probit Model

set.seed(1234)

# Splitting data into training and test sets

train\_cardio <- createDataPartition(cardio$cardio, p = 0.75, list = FALSE)

traind\_cardio <- cardio[train\_cardio, ]

testd\_cardio <- cardio[-train\_cardio, ]

# Probit regression model

probitmodel <- glm(cardio ~ age + gender + height + weight + ap\_hi + ap\_lo +

cholesterol + gluc + smoke + alco + active,

data = traind\_cardio, family = binomial(link = "probit"))

summary(probitmodel)

# Evaluation on the test dataset

cardio\_pred\_probit <- predict(probitmodel, newdata = testd\_cardio, type = "response")

cardio\_pred\_probit <- ifelse(cardio\_pred\_probit > 0.5, 1, 0)

cardio\_pred\_probit

# Model 3- Complementary log log reg Model

set.seed(1234)

# Splitting data into training and test sets

train\_cardio <- createDataPartition(cardio$cardio, p = 0.75, list = FALSE)

traind\_cardio <- cardio[train\_cardio, ]

testd\_cardio <- cardio[-train\_cardio, ]

# Complementary log-log regression model

cloglog\_model <- glm(cardio ~ age + gender + height + weight + ap\_hi + ap\_lo +

cholesterol + gluc + smoke + alco + active,

data = traind\_cardio, family = quasibinomial(link = "cloglog"))

summary(cloglog\_model)

# Evaluation on the test dataset

cardio\_pred\_cloglog <- predict(cloglog\_model, newdata = testd\_cardio, type = "response")

cardio\_pred\_cloglog <- ifelse(cardio\_pred\_cloglog > 0.5, 1, 0)

cardio\_pred\_cloglog

# Output of all three models

stargazer(logitmodel, probitmodel, cloglog\_model, type = "text")

# quality check

# Residual analysis for Logistic Regression

logit\_res <- residuals(logitmodel, type = "deviance")

hist(logit\_res, main = "Residuals Distribution - Logistic Regression", xlab = "Residuals")

# Linearity check for Logistic Regression

plot(logitmodel$fitted.values ~ traind\_cardio$age, main = "Linearity Check - Logistic Regression")

# Homoscedasticity check for Logistic Regression

plot(logit\_res ~ logitmodel$fitted.values, main = "Homoscedasticity Check - Logistic Regression")

# Multicollinearity check for Logistic Regression

library(car)

vif(logitmodel)

# Residual analysis for Probit Model

probit\_res <- residuals(probitmodel, type = "deviance")

hist(probit\_res, main = "Residuals Distribution - Probit Model", xlab = "Residuals")

# Linearity check for Probit Model

plot(probitmodel$fitted.values ~ traind\_cardio$age, main = "Linearity Check - Probit Model")

# Homoscedasticity check for Probit Model

plot(probit\_res ~ probitmodel$fitted.values, main = "Homoscedasticity Check - Probit Model")

# Multicollinearity check for Probit Model

vif(probitmodel)

# Residual analysis for Complementary Log-Log Model

cloglog\_res <- residuals(cloglog\_model, type = "deviance")

hist(cloglog\_res, main = "Residuals Distribution - Cloglog Model", xlab = "Residuals")

# Linearity check for Complementary Log-Log Model

plot(cloglog\_model$fitted.values ~ traind\_cardio$age, main = "Linearity Check - Cloglog Model")

# Homoscedasticity check for Complementary Log-Log Model

plot(cloglog\_res ~ cloglog\_model$fitted.values, main = "Homoscedasticity Check - Cloglog Model")

# Multicollinearity check for Complementary Log-Log Model

vif(cloglog\_model)

# Confusion metrics

#logit model

table1 <- table(testd\_cardio$cardio, cardio\_pred)

table1

# Calculating Recall, Precision, Accuracy, F1score, AUC

recall <- table1[2,2]/(table1[2,2]+table1[2,1])

recall

precision <- table1[2,2]/(table1[2,2]+table1[1,2])

precision

accuracy <- sum(diag(table1)) / sum(table1)

accuracy

f1\_score <- (2 \* precision \* recall) / (precision + recall)

f1\_score

roc\_obj <- roc(response = testd\_cardio$cardio, predictor = cardio\_pred)

auc <- auc(roc\_obj)

auc

# Probit model

# Calculating Recall, Precision, Accuracy, F1score, AUC

table2 <- table(testd\_cardio$cardio, cardio\_pred\_probit)

table2

recall <- table2[2,2]/(table2[2,2]+table2[2,1])

recall

precision <- table2[2,2]/(table2[2,2]+table2[1,2])

precision

accuracy <- sum(diag(table2)) / sum(table2)

accuracy

f1\_score <- (2 \* precision \* recall) / (precision + recall)

f1\_score

roc\_obj <- roc(response = testd\_cardio$cardio, predictor = cardio\_pred\_probit)

auc <- auc(roc\_obj)

auc

# Complementary log- log model

# Calculating Recall, Precision, Accuracy, F1score, AUC

table3 <- table(testd\_cardio$cardio, cardio\_pred\_cloglog)

table3

recall <- table3[2,2]/(table3[2,2]+table3[2,1])

recall

precision <- table3[2,2]/(table3[2,2]+table3[1,2])

precision

accuracy <- sum(diag(table3)) / sum(table3)

accuracy

f1\_score <- (2 \* precision \* recall) / (precision + recall)

f1\_score

roc\_obj <- roc(response = testd\_cardio$cardio, predictor = cardio\_pred\_cloglog)

auc <- auc(roc\_obj)

auc