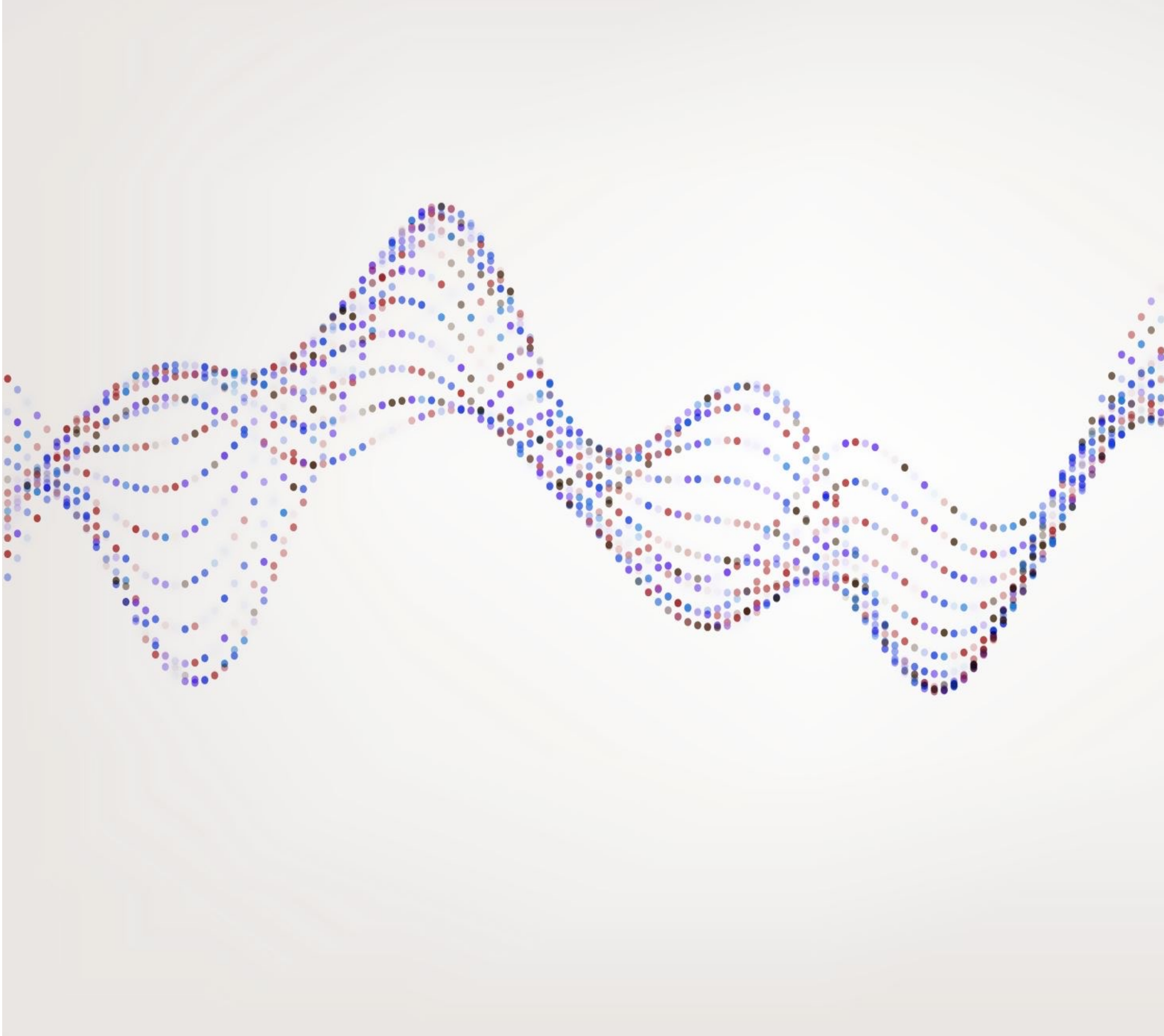
# CORONARY HEART DISEASE PREDICTION

TEAM 5:

BURUGU SAI SHARAN (G01383197)

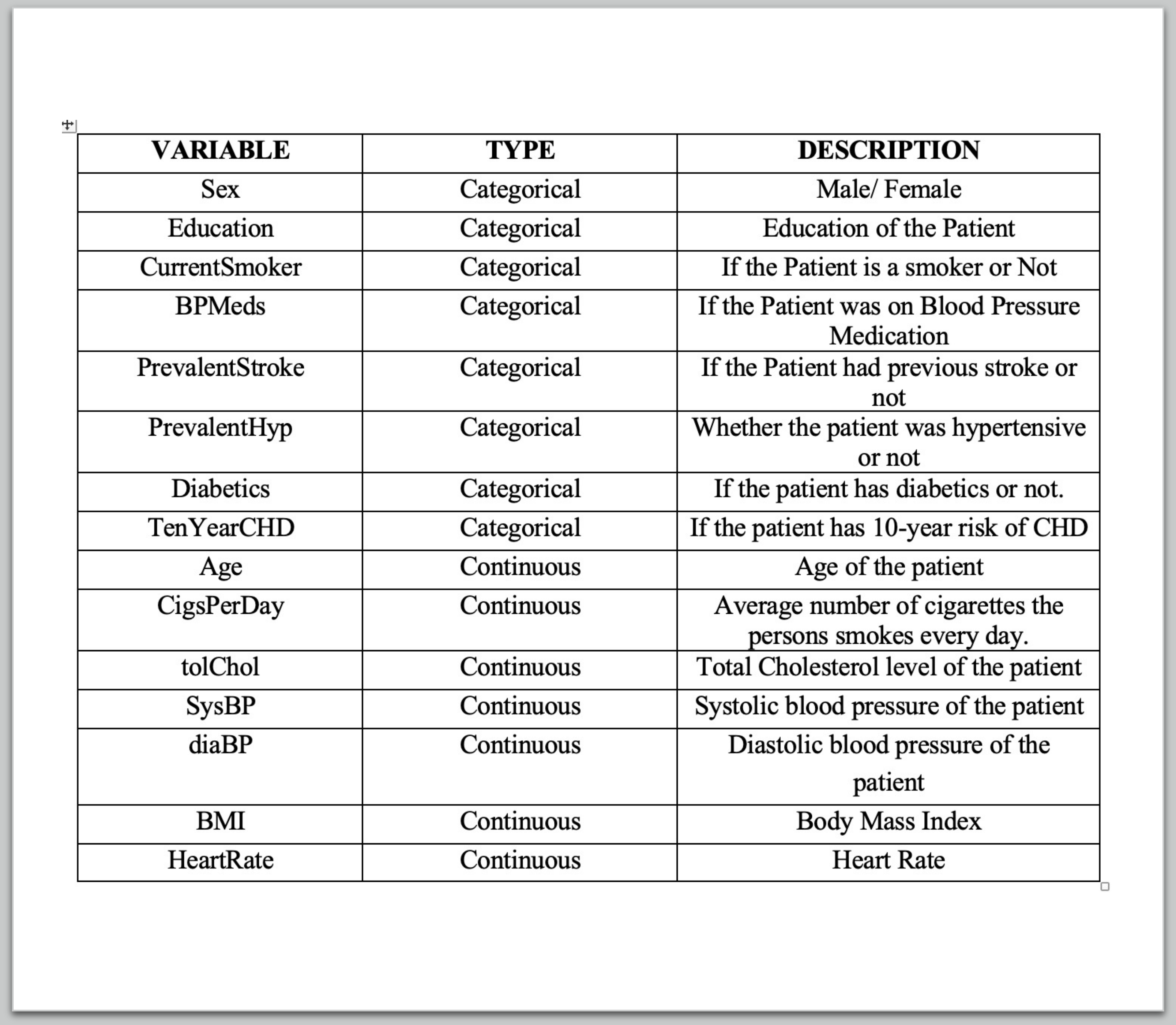
AKHIL RACHURE (G01390507)

SOWMYA CHAKRAVARTHY (G01380299)



# Dataset Description:

- This data set corresponds to the ongoing cardiovascular study of people living in the town of Framingham, Massachusetts and is available online from Framingham heart study.
- The objective is to determine if the patient has a 10-year risk of developing future Coronary Heart Disease (CHD).



| VARIABLE | TYPE | DESCRIPTION |
|---|---|---|
| Sex | Categorical | Male/ Female |
| Education | Categorical | Education of the Patient |
| CurrentSmoker | Categorical | If the Patient is a smoker or Not |
| BPMeds | Categorical | If the Patient was on Blood Pressure Medication |
| PrevalentStroke | Categorical | If the Patient had previous stroke or not |
| PrevalentHyp | Categorical | Whether the patient was hypertensive or not |
| Diabetics | Categorical | If the patient has diabetics or not. |
| TenYearCHD | Categorical | If the patient has 10-year risk of CHD |
| Age | Continuous | Age of the patient |
| CigsPerDay | Continuous | Average number of cigarettes the persons smokes every day. |
| tolChol | Continuous | Total Cholesterol level of the patient |
| SysBP | Continuous | Systolic blood pressure of the patient |
| diaBP | Continuous | Diastolic blood pressure of the patient |
| BMI | Continuous | Body Mass Index |
| HeartRate | Continuous | Heart Rate |

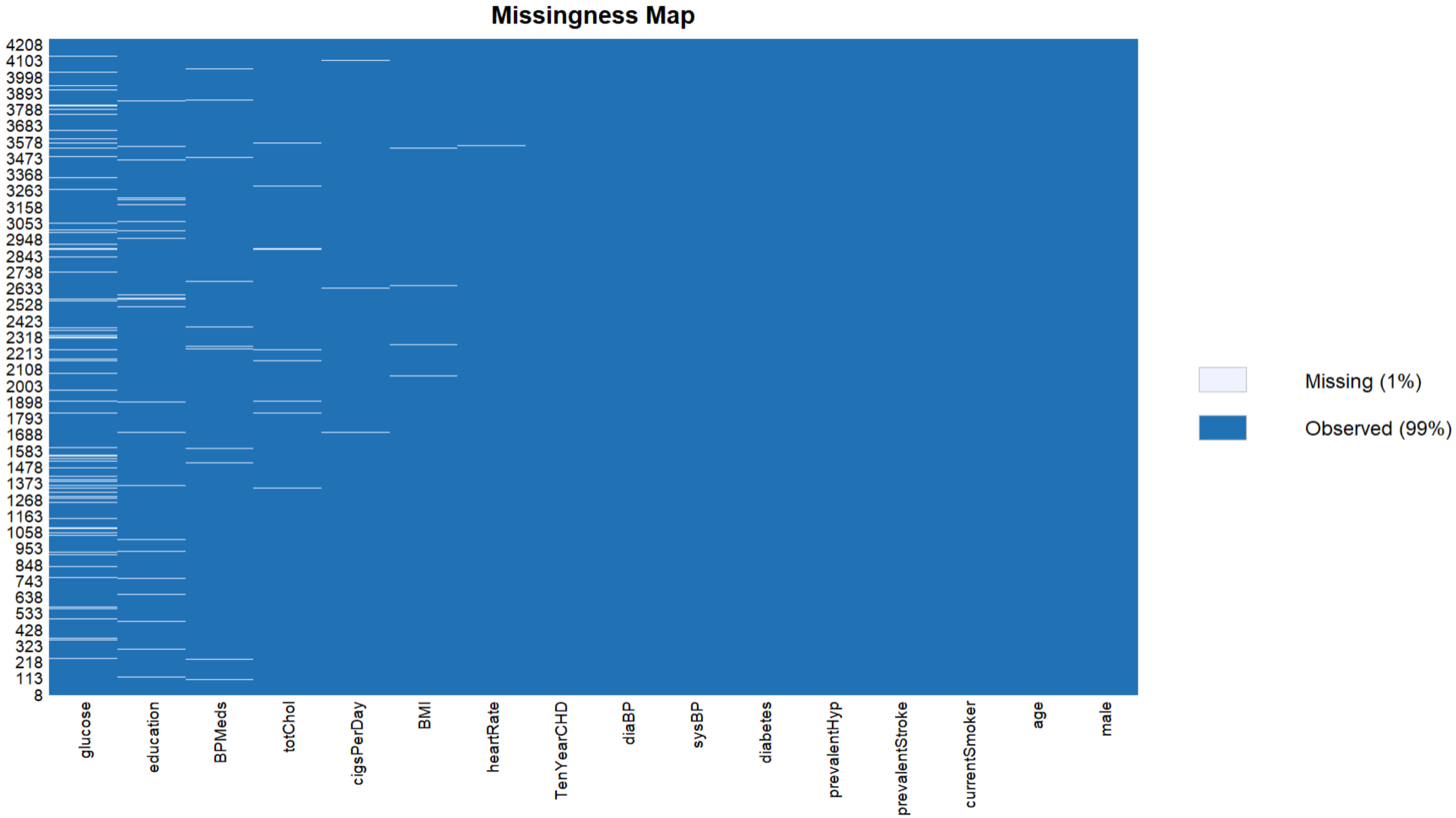
# Research Questions

*What are the top 5 features that cause Cardio Vascular Diseases?*

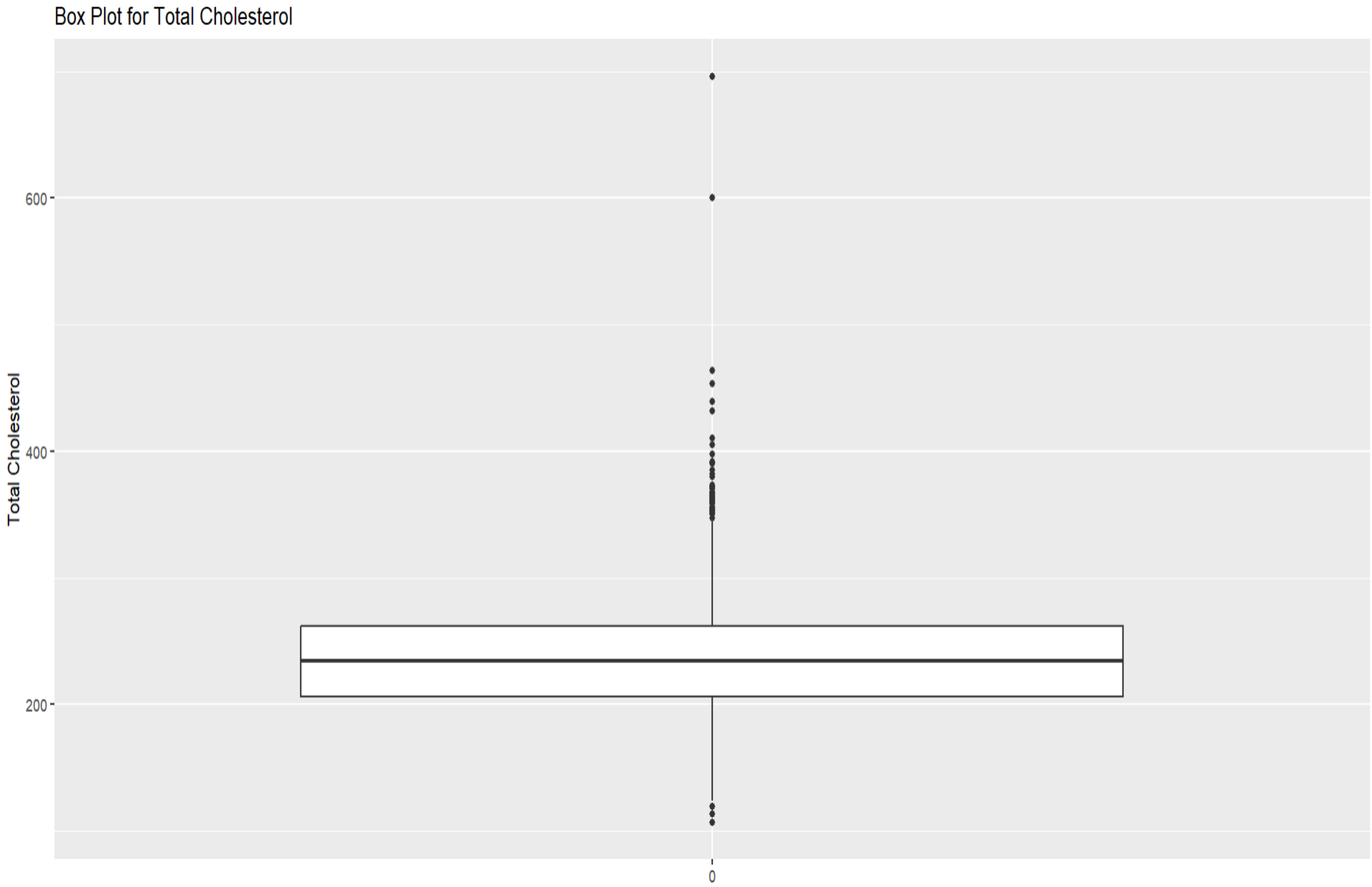
*How does age affect the risk factor?*

*What role does Blood Pressure have in predicting the Heart Disease risk?*

# Data Cleaning: Handling Missing Values



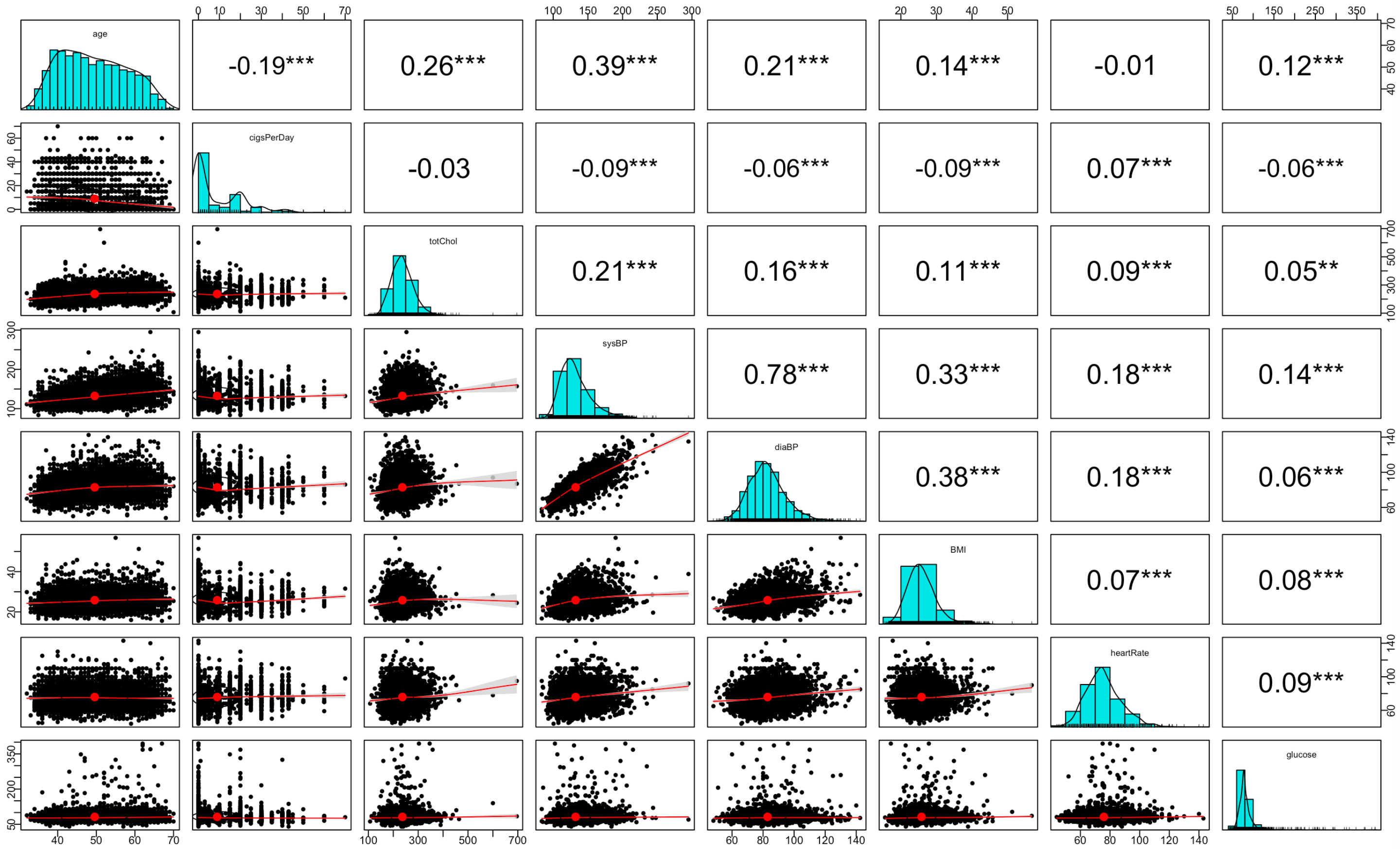
# Data Cleaning:



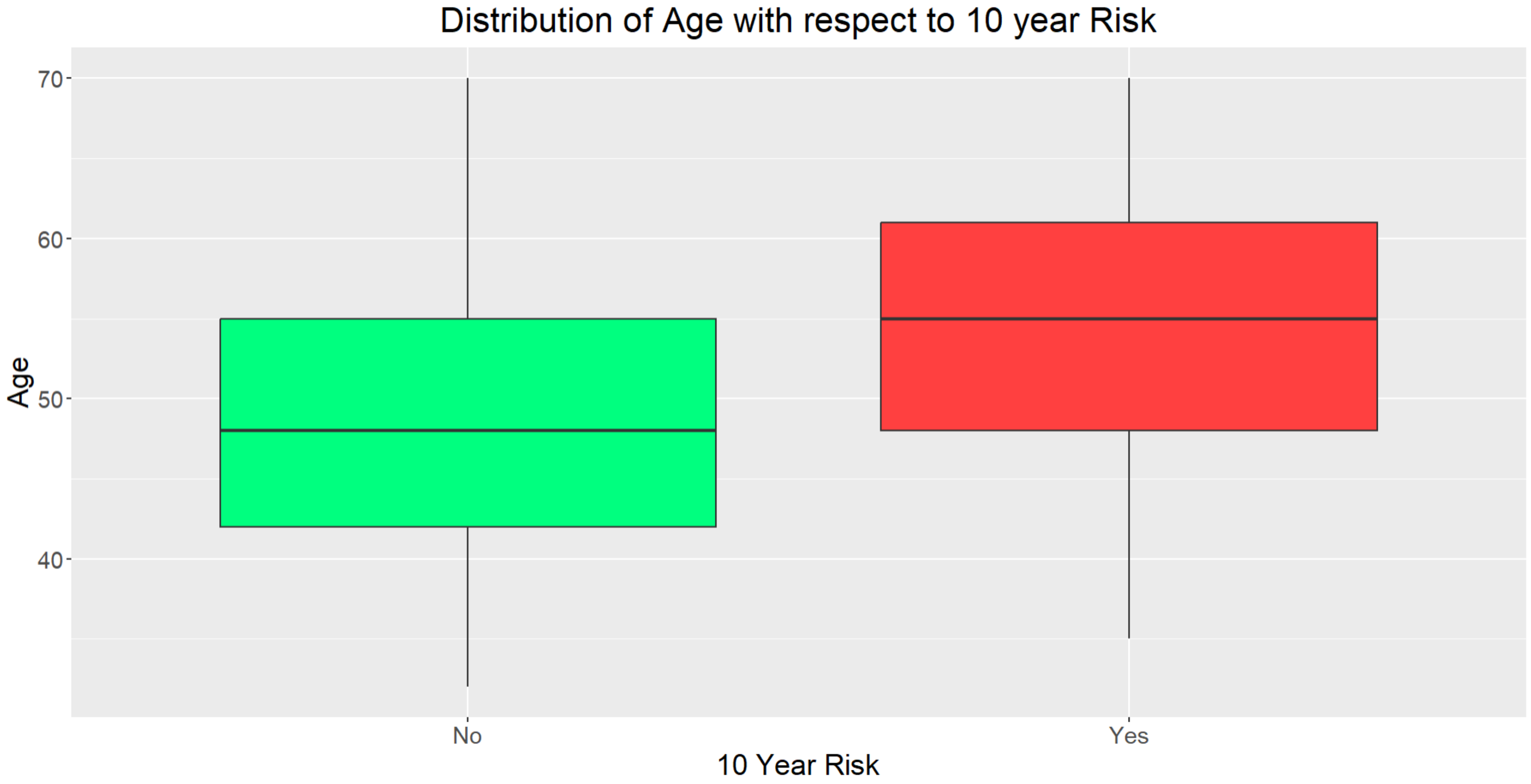
- Outlier Detection
- Removal of unnecessary columns

# Exploratory Data Analysis

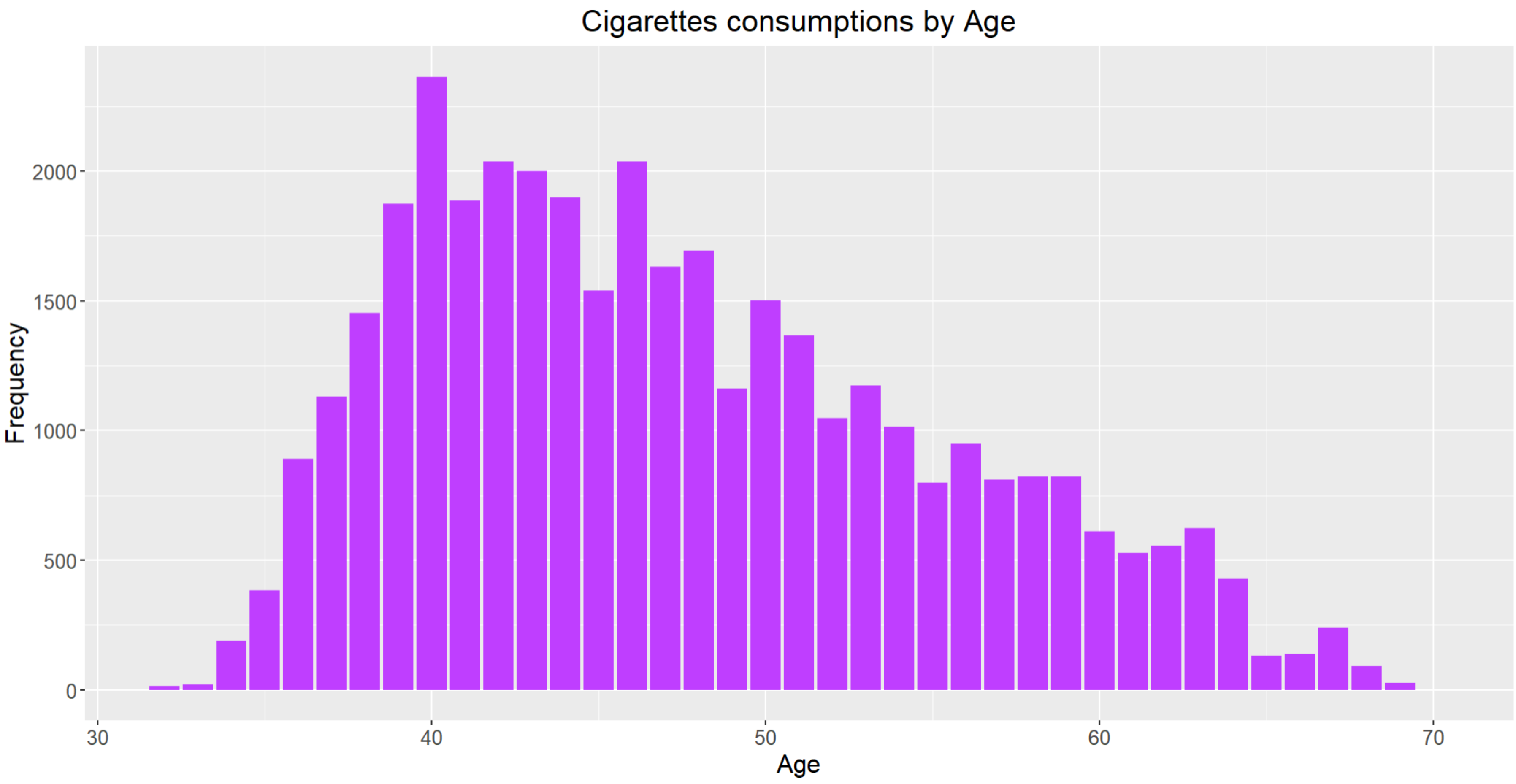
## Correlation Matrix



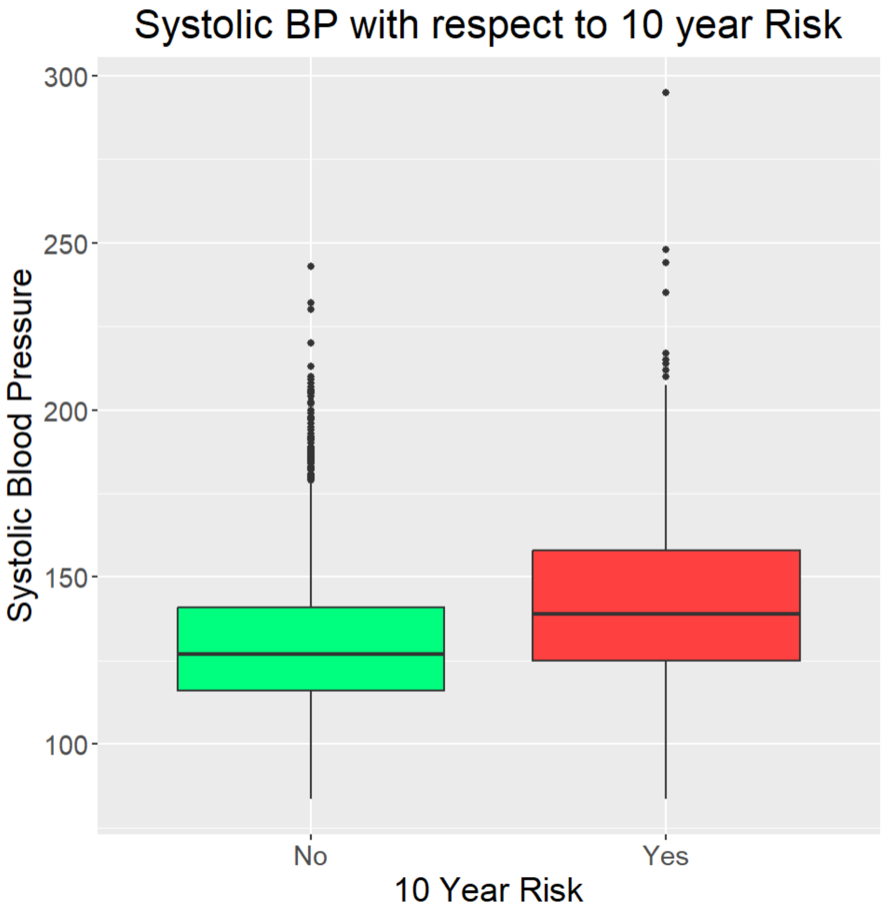
# Box Plot : Age vs 10-Year Risk



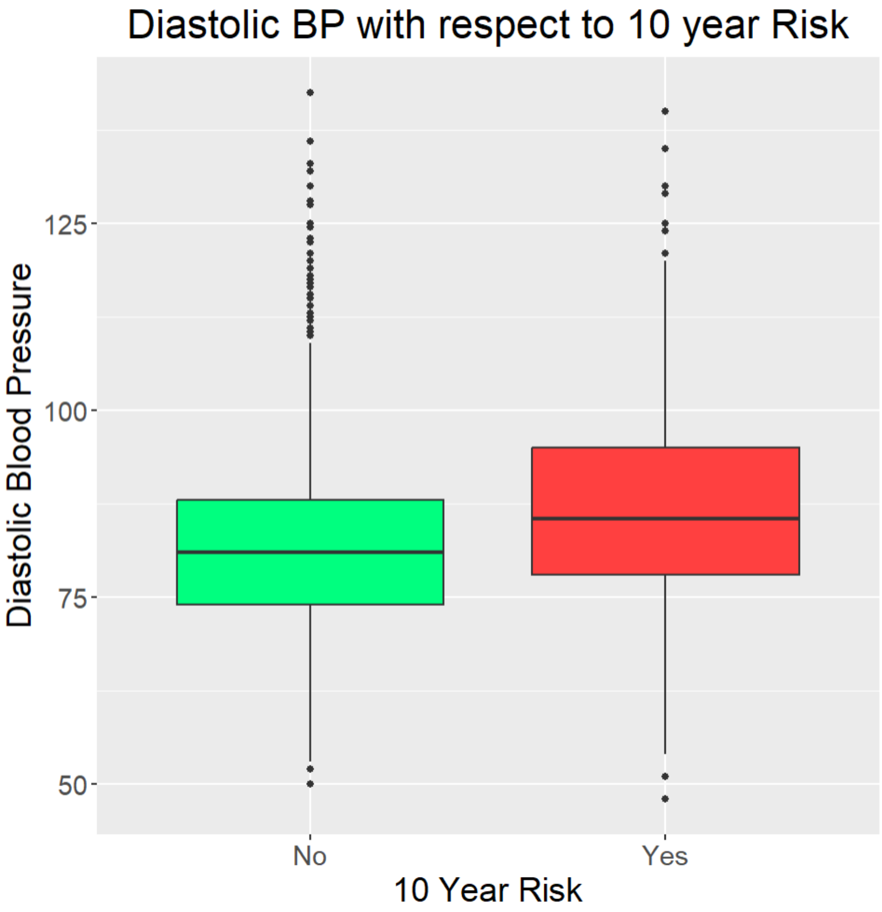
# Bar Graph : Cigarettes Consumption vs AGE



# Box Plot : BP vs 10-Year Risk

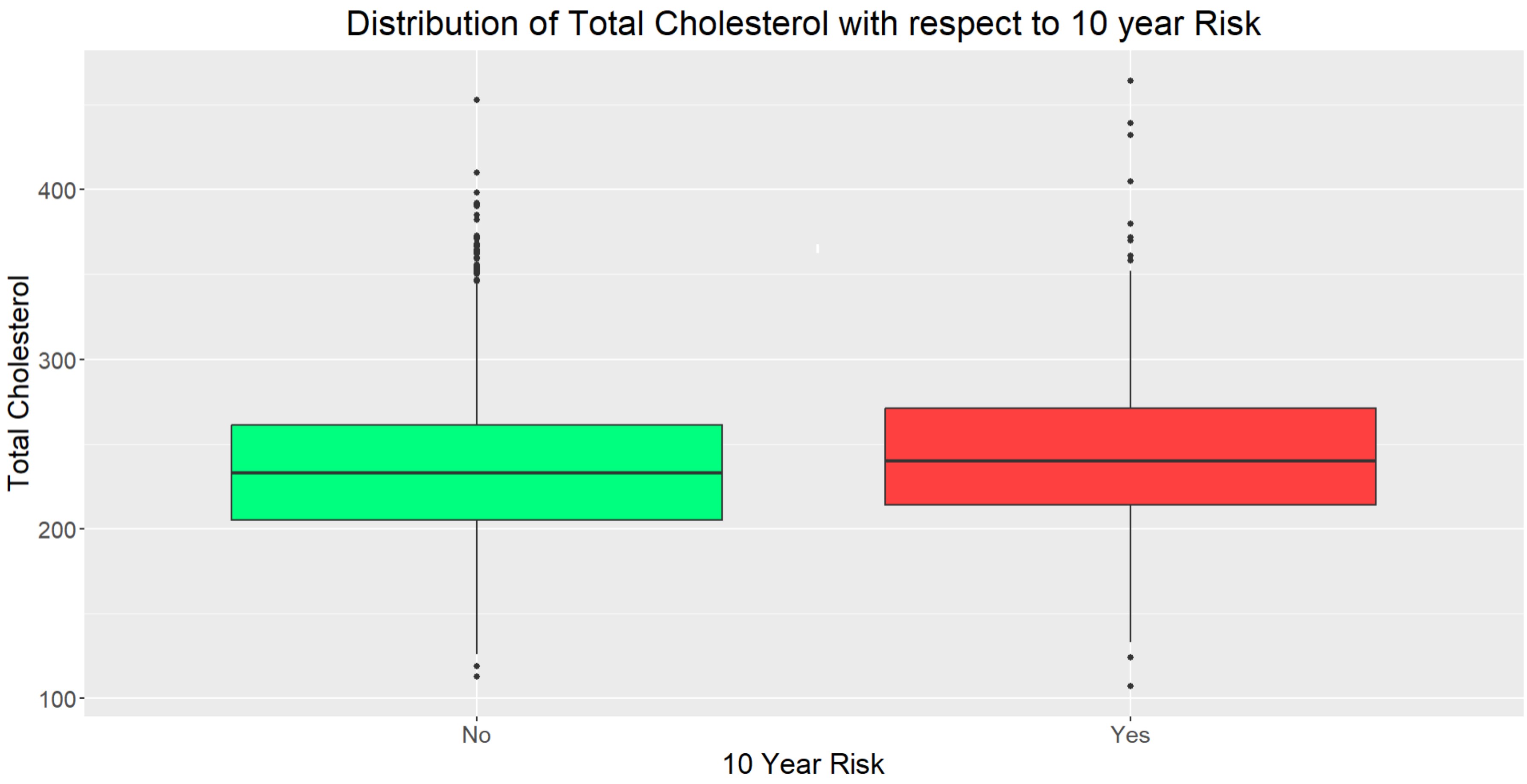


Systolic BP with respect to 10 year Risk

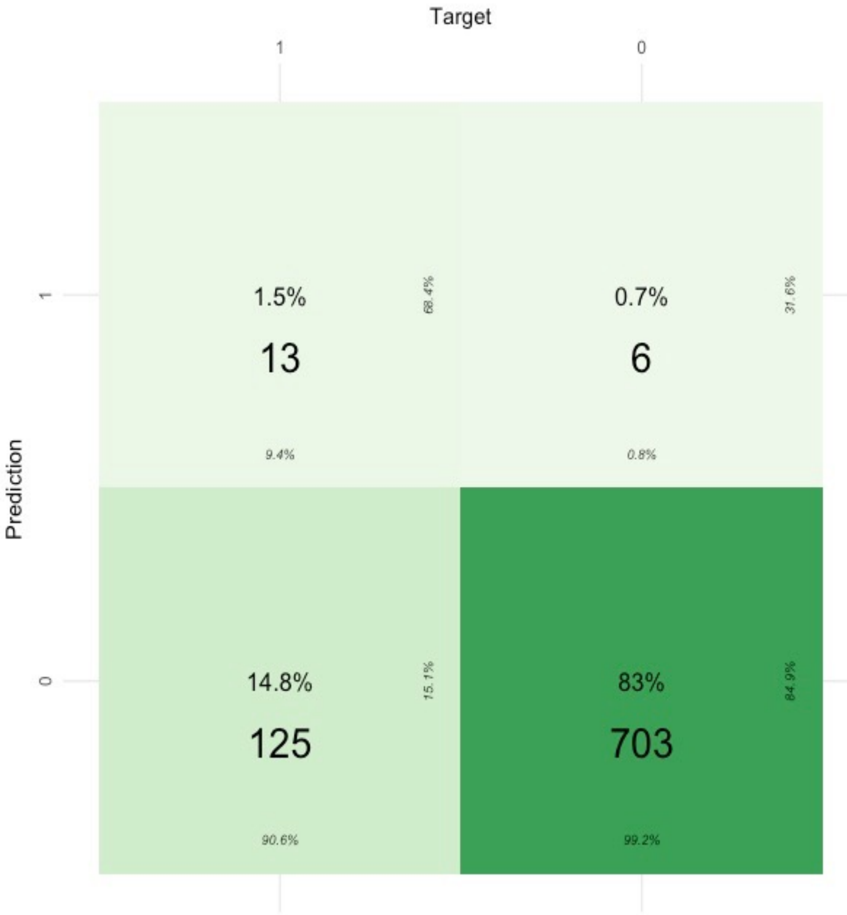


Diastolic BP with respect to 10 year Risk

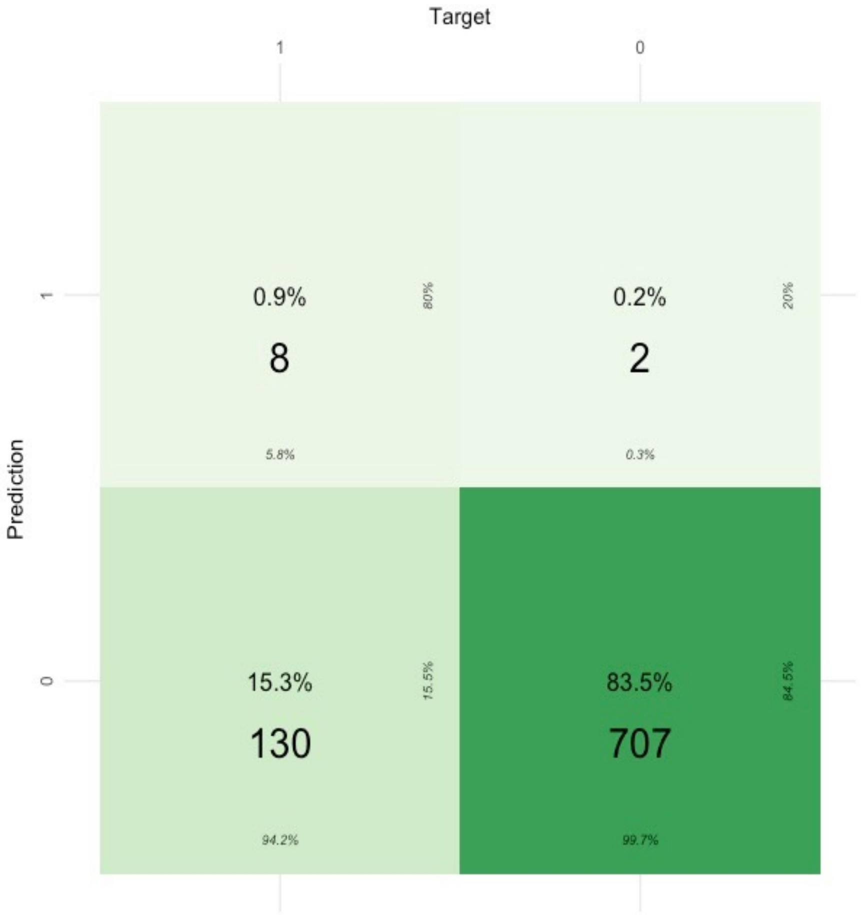
# Box Plot : Total Cholesterol vs 10-Year Risk



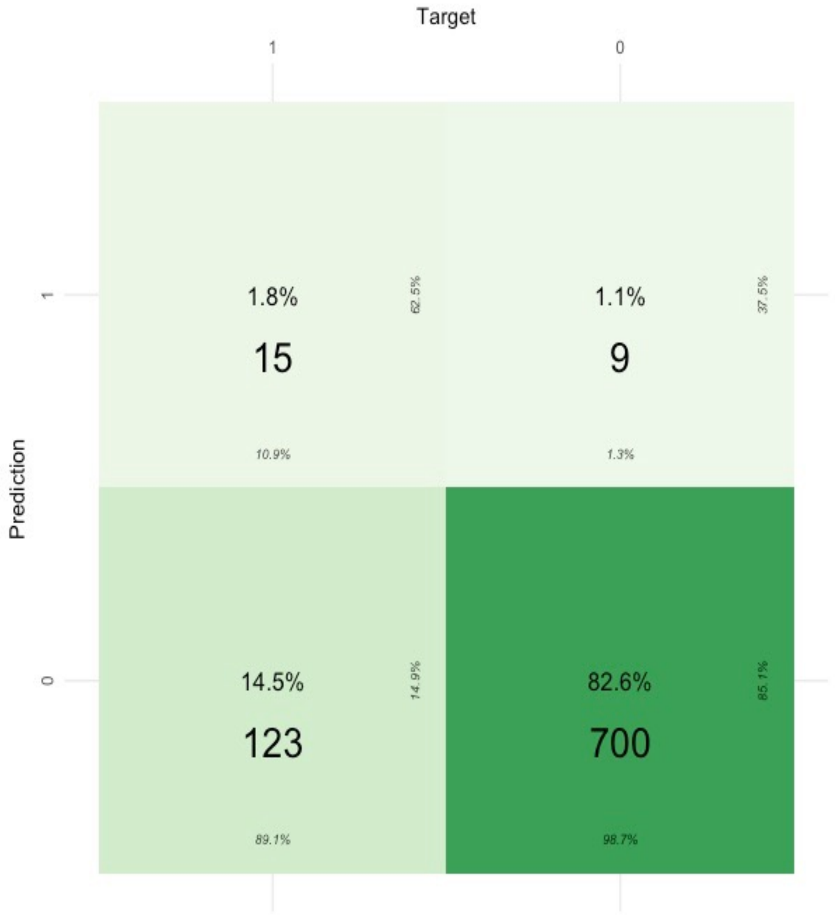
# Modeling: Comparison of Results



Logistic Regression

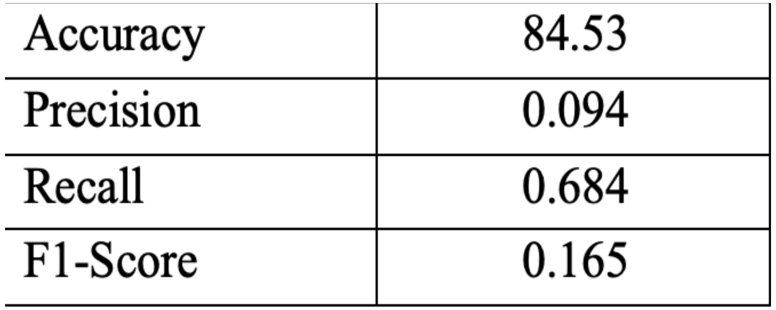


KNN



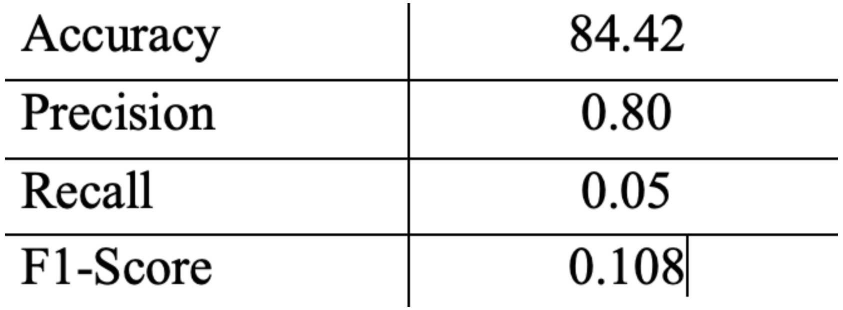
Random Forest

# Modeling: Comparison of Scores



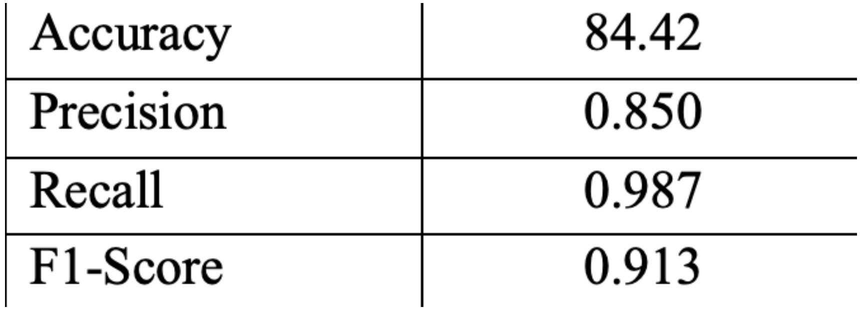
| | |
|---|---|
| Accuracy | 84.53 |
| Precision | 0.094 |
| Recall | 0.684 |
| F1-Score | 0.165 |

Logistic Regression



| | |
|---|---|
| Accuracy | 84.42 |
| Precision | 0.80 |
| Recall | 0.05 |
| F1-Score | 0.108 |

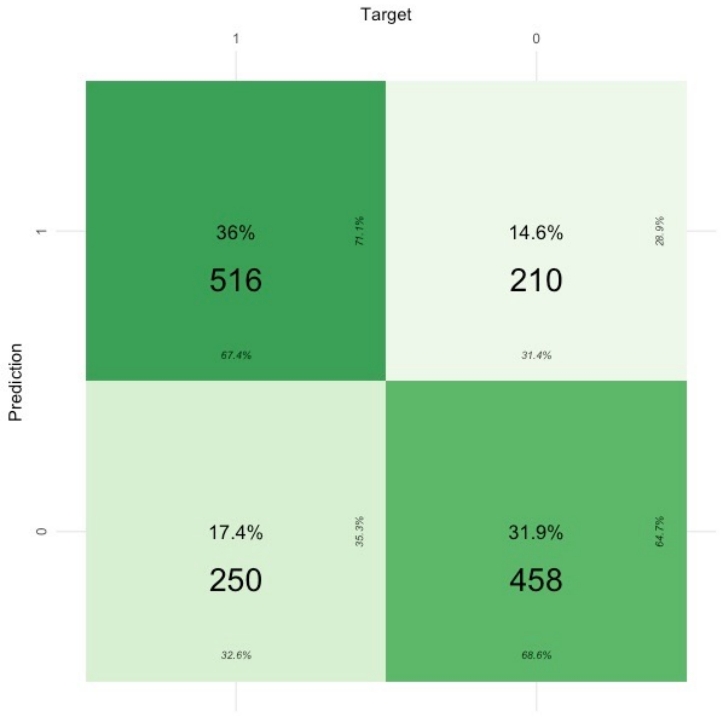
KNN



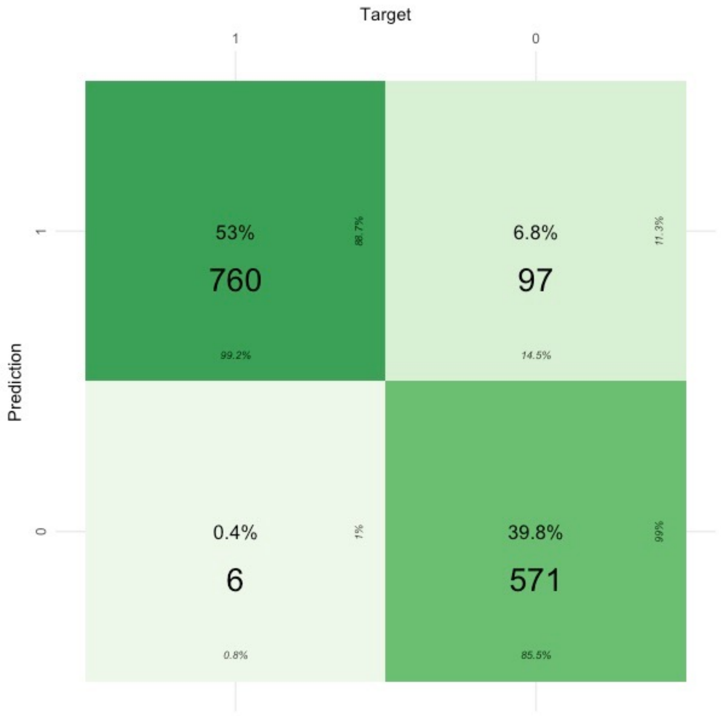
| | |
|---|---|
| Accuracy | 84.42 |
| Precision | 0.850 |
| Recall | 0.987 |
| F1-Score | 0.913 |

Random Forest

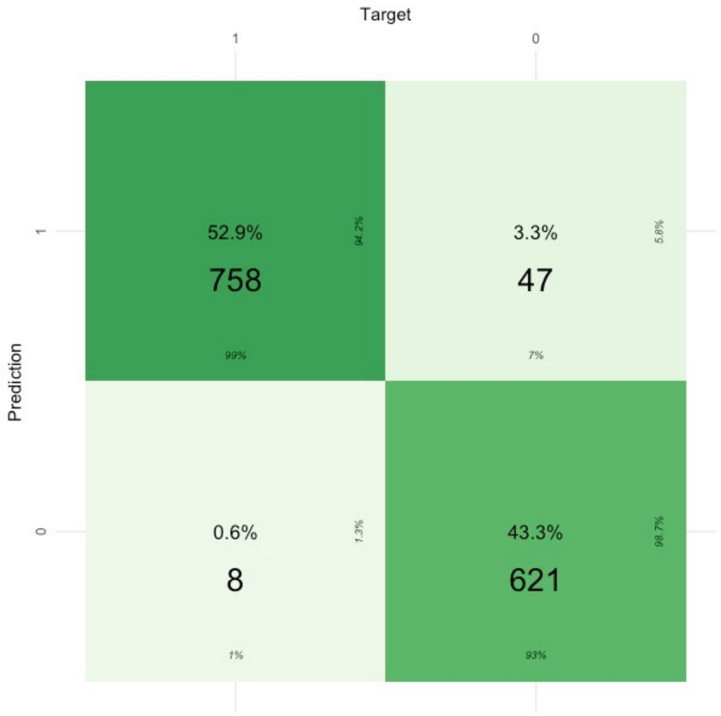
# Modeling: Comparison of Results (SMOTE)



Logistic Regression

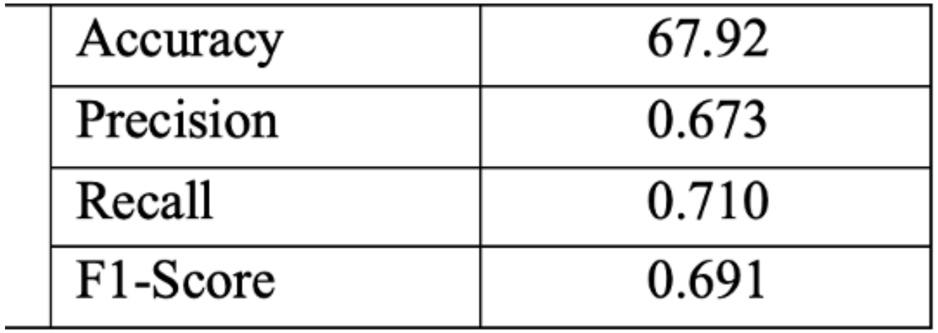


KNN



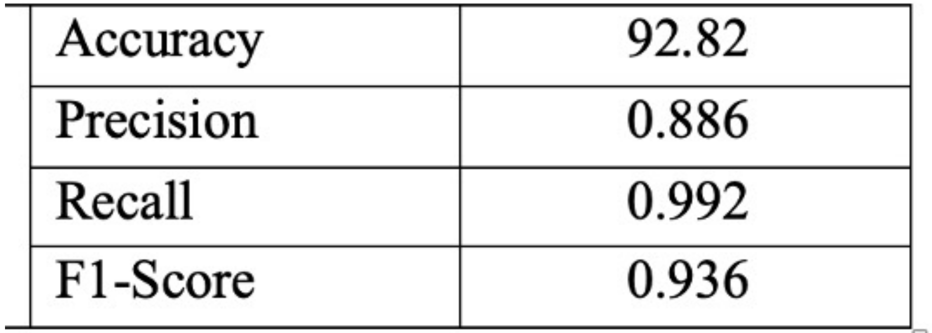
Random Forest

# Modeling: Comparison of Results (SMOTE)



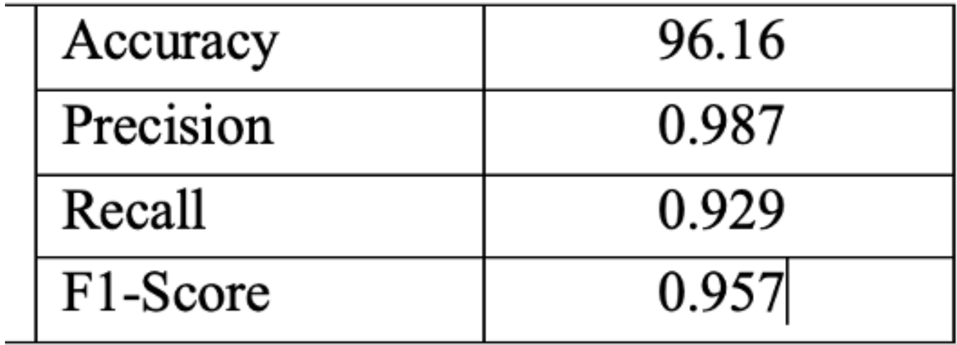
| | |
|---|---|
| Accuracy | 67.92 |
| Precision | 0.673 |
| Recall | 0.710 |
| F1-Score | 0.691 |

Logistic Regression



| | |
|---|---|
| Accuracy | 92.82 |
| Precision | 0.886 |
| Recall | 0.992 |
| F1-Score | 0.936 |

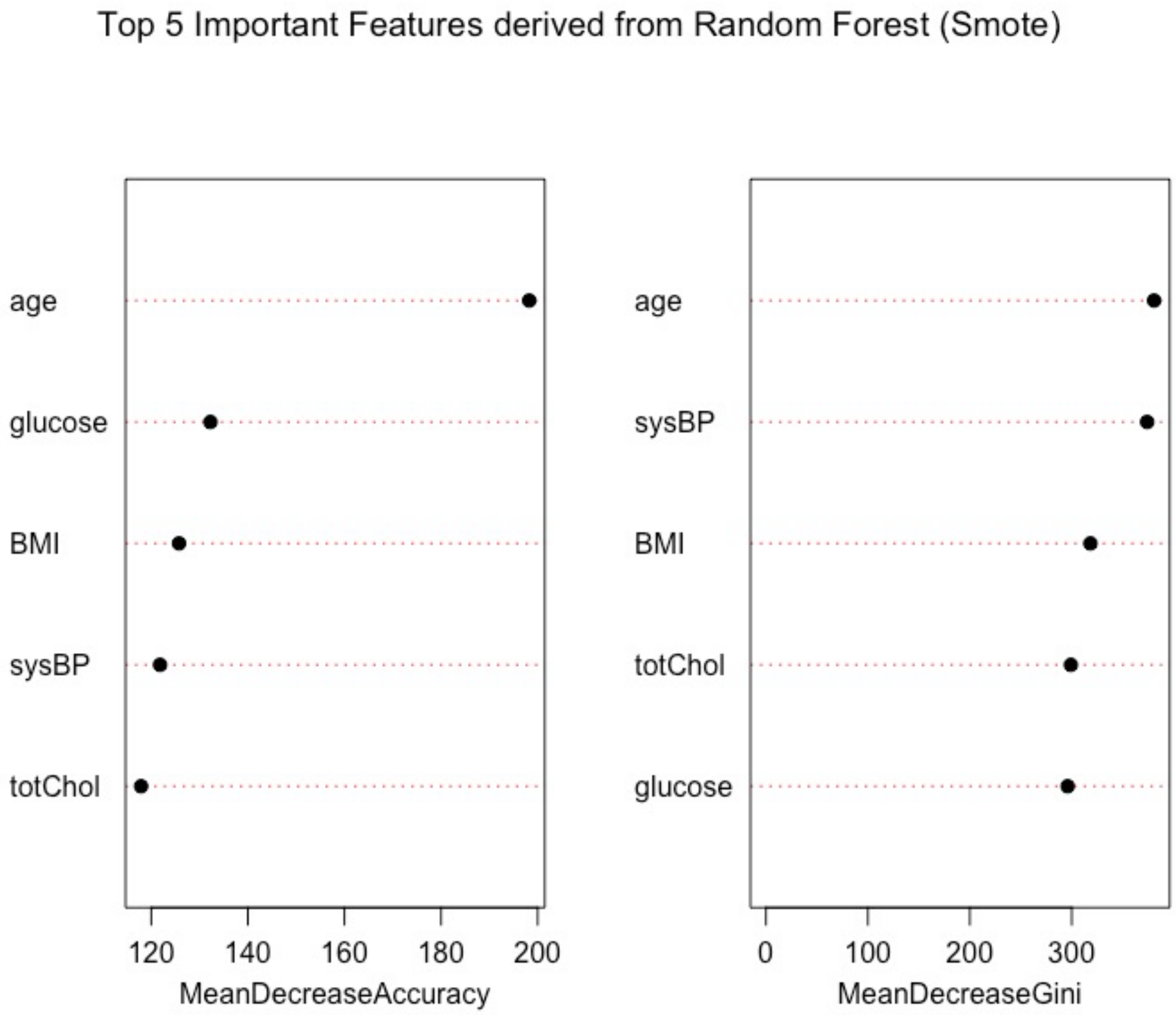
KNN



| | |
|---|---|
| Accuracy | 96.16 |
| Precision | 0.987 |
| Recall | 0.929 |
| F1-Score | 0.957 |

Random Forest

# Feature importance



# Findings:

- Age, Glucose level, Blood Pressure, BMI, Total Cholesterol are the top 5 features
- Higher that age greater the chance of getting the Heart Disease.
- Increase in Systolic Blood Pressure and Diastolic Blood Pressure will increase the change of getting the heart disease.

Thank You