CARDIOVASCULAR DISEASE PREDICTION

INTRODUCTION

- Cardiovascular Disease is one the widespread health issues in the healthcare system
- Our job is to perform queries and generate the best possible binary classification model to predict people who are likely to have Cardiovascular Diseases.
- Our dataset consists of 13 columns and 69,301 rows
- All our columns are of numeric type

BUSINESS QUESTIONS

- Do Systolic and Diastolic Blood Pressures lead to Cardiovascular Disease?
- Does Cholesterol Level Play a vital role?
- Is Glucose Level Responsible significantly?
- Are smoking tobacco and drinking alcohol playing their part?
- Can Cardiovascular Diseases be attributed to Physical Inactivity?
- Do demographic Variables like gender and age have their roles?
- Could Body Mass Indicator(BMI) be a critical player as well?
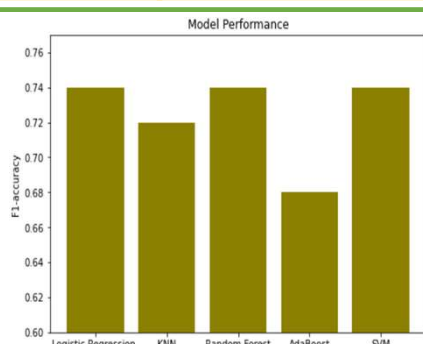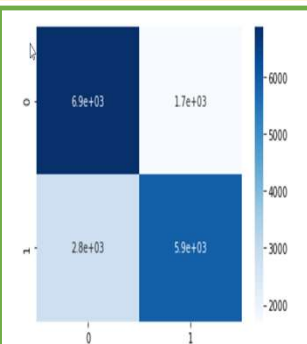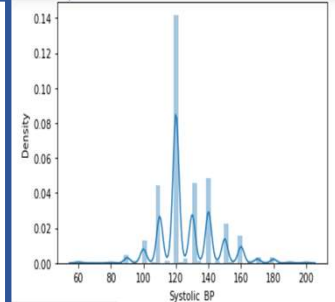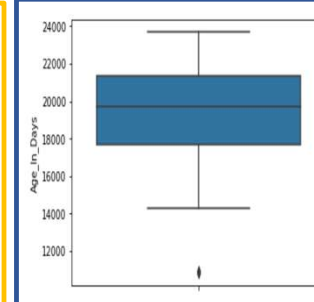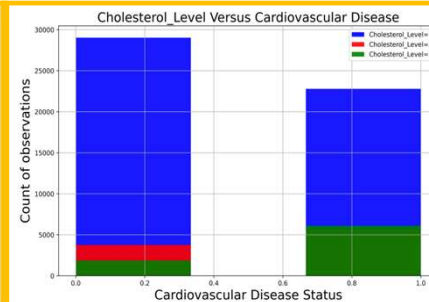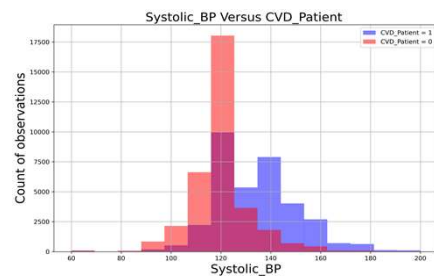
OBJECTIVES

- Identification of the features which impacts Cardiovascular Diseases
- Develop a binary classification model with the highest performance
- Conclude whether our hypothesis holds good

UNIVARIATE ANALYSIS

Distribution of CVD Patient
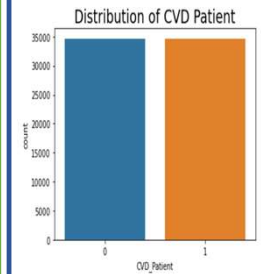
BIVARIATE ANALYSIS

Feature	P-value < 0.05
Physically_Active	YES
Gender	YES
Smoker	YES
Cholesterol_Level	YES
Glucose_Level	YES
Drinker	NO
Systolic_BP	YES
Diastolic_BP	YES
Age_In_Years	YES
BMI	YES

Systolic_BP Versus CVD_Patient

Cholesterol_Level Versus Cardiovascular Disease


```
              precision    recall  f1-score   support

           0       0.71      0.80      0.75      8573
           1       0.78      0.68      0.72      8728

    accuracy                           0.74     17301
   macro avg       0.74      0.74      0.74     17301
weighted avg       0.74      0.74      0.74     17301
```


Model Performance

CONCLUSIONS

- The result suggests that Logistic Regression, Random Forest and SVM Classification performed better
- Only one of our hypothesized parameters(alcohol) that were in the dataset marginally showed statistical insignificance.
- Some of our hypothesized parameters were absent in our dataset and inclusion of those might improve the model performance