Prediction of Heart Disease based on Clinical Factors

Group 6
Masimukku, Satya Aditya
Nalamothu, Prathyusha
Nandala, Sumana Sree
Krishnagiri Tuppal, Venu Gopalan

Instructor
Ali, Syed

INFO 5810 – Data Analysis and Knowledge Discovery

October 7, 2023

Abstract

With increasing applications of Machine Learning Algorithms and Techniques, we can understand and analyze data more efficiently. This application has reach to other sectors, such as medical field. With advancements in medicine, data related to the body and its state from patients can be collected with high precision. This data, combined with AI/ML techniques could be helpful in understating the relation between various factors and health complications. This could be helpful in predicting and treating crucial health issues before they become chronic or severe. One such case is Cardiac Arrest which is caused by Coronary Artery Disease (also known as Coronary Heart Disease). There are multiple factors that could cause coronary heart disease. This paper focuses on those factors such as Glucose Level, Diabetic Type, Pressure of Blood flow, Cholesterol in the body and other clinical factors that describe the human body. By training, accuracy of Machine learning models can be tested and by studying them, we can know which model suits our dataset. After comparing multiple models, the most efficient model with highest accuracy is consider supporting the claims of relationship and impact of the clinical factors on coronary heart disease.

Introduction

Background: Diabetes mellitus is a group of diseases that affects the usage of glucose (sugar is blood). The main impact that diabetes has on body is increase in level of glucose in blood. Even though glucose acts as source of energy for the functioning of body, and brain and helps in developing tissues and muscles, excess of sugar in blood leads to health problems that could be chronic. Complications from diabetes includes cardiovascular disease, increased risk for kidney failure, blindness, mortality, etc. It is also predicted that diabetics could affect more that 690 million adults by the end of 2045 (Cole & Florez, 2020). There are many factors that affect the functioning of body in presence of diabetes. If we consider the major risk among the issues that are impacted by the diabetes, Coronary Artery Disease (also known as Coronary Heart Disease) that is caused by the thickening of the walls of the artery, making it difficult for the blood to flow and leading to cardiac arrest can be considered he health complication with highest importance.

Objective: The main aim of this study is to predict the possibility of a person being affected by coronary artery disease under various clinical factors that define the human body functioning.

Scope: This report focuses on studying and understanding the occurrence or coronary heart disease among the people with diabetes along with other clinical factors such as Age, Height, Weight, Blood components (such as Iron, RBC Count/Hematocrit), Glucose, Cholesterol, etc.

Structure: Firstly, giving the review of the research that has been conducted in this aspect, the document contains the details of the Data collection and the steps used to process it (Cleaning and Analysis). Results with related images and graphs are presented. Then, other aspects and opinions are discussed with a concluding statement. Future works and Acknowledgements are mentioned, and document is concluded with References and Appendices.

Literature Review

Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques Resolving Class Imbalance is a used a dataset that contained around 37000 values to perform analysis. By using the Random Forest Algorithm and Analysis of Variance techniques, the authors concluded that attributes related to body such as BMI, Cholesterol, Age, etc. The authors tackled the issue of class imbalance by assigning weight classes and data was standardized using scikit-learn tool. Random Forest Algorithm achieved highest accuracy of 90% and score for Area Under the Curve (from scikit-learn tool) has 98% accuracy. Authors mentioned that these accuracy scores can be further improved by using other enhanced techniques.

Data Collection and Processing

Data Sources: Source of the dataset that is used for analysis and prediction is from Kaggle (https://www.kaggle.com). In this website, we found a dataset '***Cardiac Data***' that has the records of 37079 patients (observations) distributed across 51 attributes, each describing a clinical factor that could affect the functioning of the human body.

Data Integrity: As the Dataset *Cardiac Data NHanes 65c8df6e-2* found in Kaggle is data that is retrieved from website National Health and Nutrition Examination Survey (NHANES) under National Center for Health Statistics (part of Center for Disease Control and Prevention - CDC). It contains the data from the year 1999 till 2016. This survey is nutritional and health status of both adults and children and takes sample of 5000 people each year across the country (*NHANES - About the National Health and Nutrition Examination Survey*, n.d.).

Data Preprocessing: The dataset *Cardiac Data NHanes 65c8df6e-2* has no missing values in it. It is considered that either the data was carefully recorded, or the dataset is already cleaned (removing/replacing missing values). Not all the clinical records are related to Diabetes. Few attributes such as *Iron*, *Protein*, *Uric. Acid* (components) are not used for training the model. So, these attributes are eliminated and only the columns that are used for processing the data is considered. This makes it easy for the model to understand and oversee data properly. We have columns such as *Gender*, *Diabetes*, *Blood-Rel-Diabetes*, *Blood-Rel-Stroke*, *CoronaryHeartDisease,* that are categorical variables which are in numeric (int64) format. These are converted into categorical type (*Gender* from {1, 2} to {M, F}) and *Diabetes* from {1, 2, 3} to {Type-1, Type-2, Type-3} in new column – *DiabetesType,* refer Appendix A)

Results

Descriptive Statistics: For the record of 37,079 patients, we used 21 columns out of 51 that are related to sugar levels and glucose. The data belongs to patients with minimum age of 20 and maximum of 85 with average of 48 Years. Out of 37,079 patients, 19,032 belong to Gender 2 (considered as Female - F) and 18,047 belong to Gender 1 (considered as Male -M). For this data, patients with Type-1 Diabetes are 4,144 and Type-2 and Type-3 are 32,227 and 708, respectively. Measure of a few among the used attributes are as follows:

- Body-Mass-Index: Ranges from 13.18 to 130.21 with average of 28.83 BMI

- Total-Cholesterol: Ranges from 1.53 to 14.09 with average of 5.08 mmol/L
- Glucose: Average Glucose is 5.6 mmol/L with 1.05 as least and 34.25 as most
- Glycohemoglobin: Lowest level is 2% and Highest is 18.8% and has a mean of 5.68%

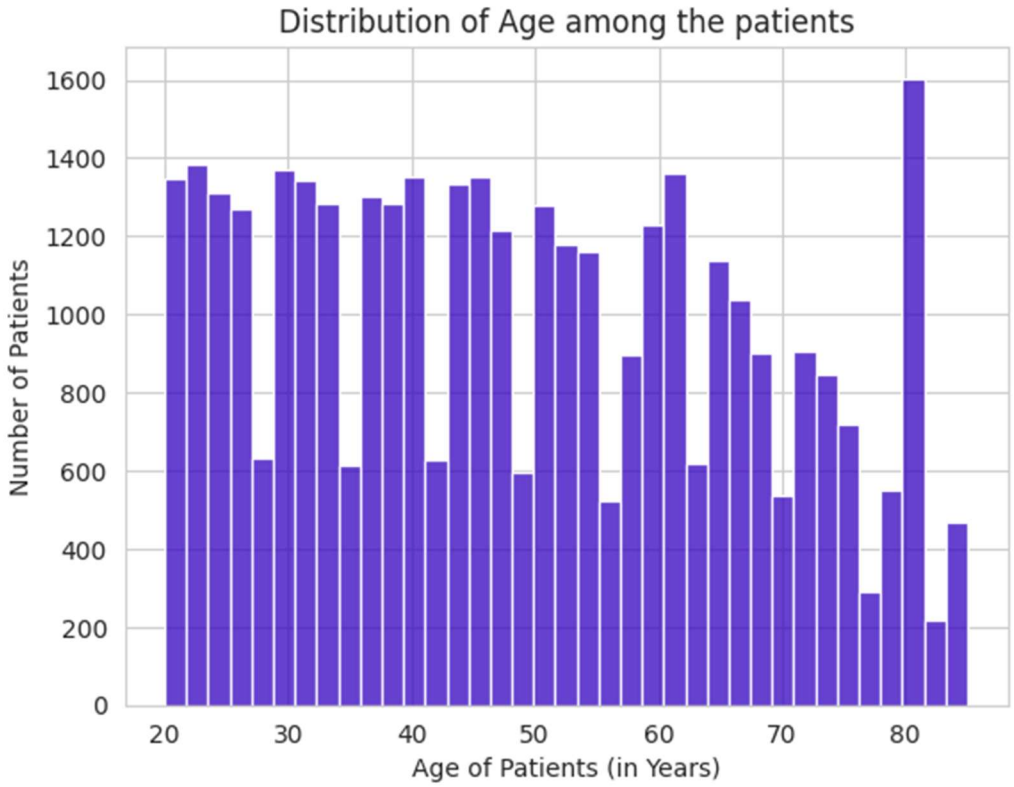
The patients who are affected by Coronary Artery Disease (CoronaryHeartDisease) are 1,508 among the population of 37,079.

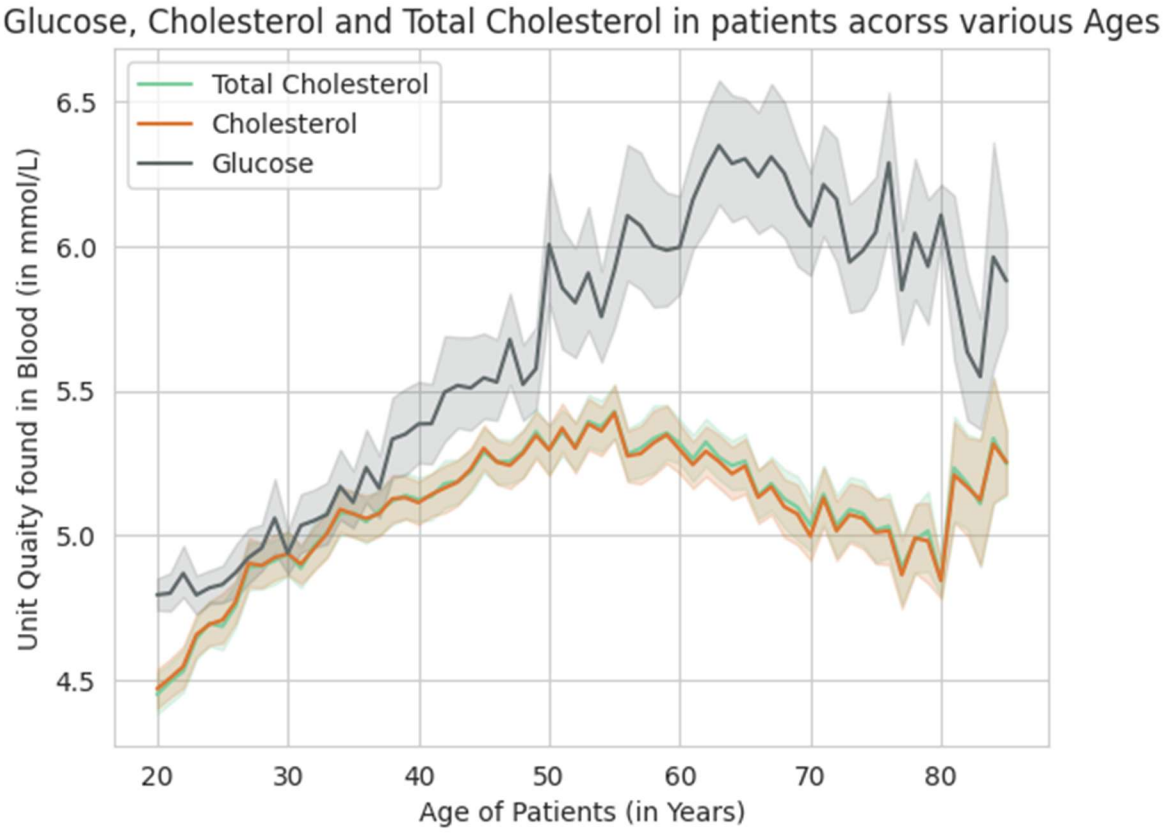
Inferential Statistics: Following are the inferences that are made after observing the data:

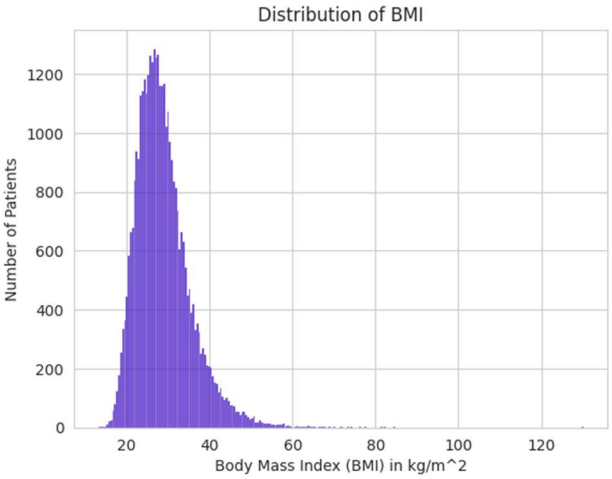
- Data collected is mostly related to earlys and mid's of the age groups. Late's of the data is less when compared to the early's and mid's. i.e., Data related to early 30s (31,32,33) and mid 30s (34,35,36,37) is more focused and data on late 30s (38,39) is less focused.
- This pattern is not followed age of 65 and surprisingly, the age group that has highest number of samples are early 80s. (find the visualization for reference)
- Most of the attributes follow Normal Distribution and slightly skewed towards right (except for Mean-Cell-Vol).
- Glucose and Cholesterol followed same pattern (increasing with age) till 50 Years. After that, Cholesterol is found less when comparing to Glucose Levels. Glucose levels increased till it hit 70 years.
- Glucose levels fluctuated (for Type-3 and Type-1 Diabetic Patients) for the age group 20-40 and then followed more stable pattern later.
- Glycohemoglobin and Glucose follow the same pattern when compared against age groups.
- It is also observed that the average Glucose levels are higher among Males when compared to Females.
- More patients for Type-1 and Type-3 Diabetes were from age group 20-50 had high level of Cholesterol and for age group 55-85, Type-2 patients were more in number with more cholesterol.

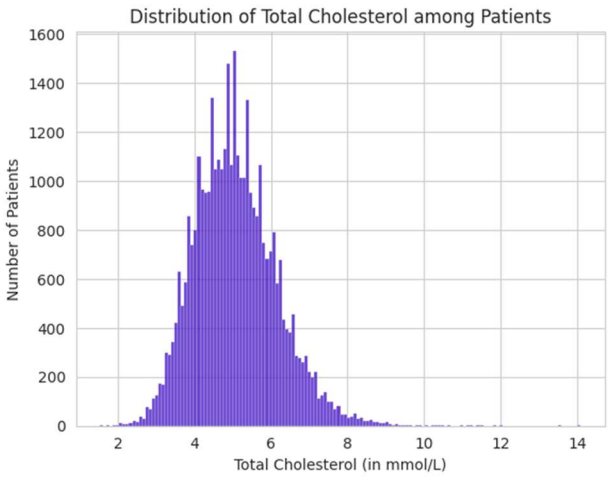
Visualizations:

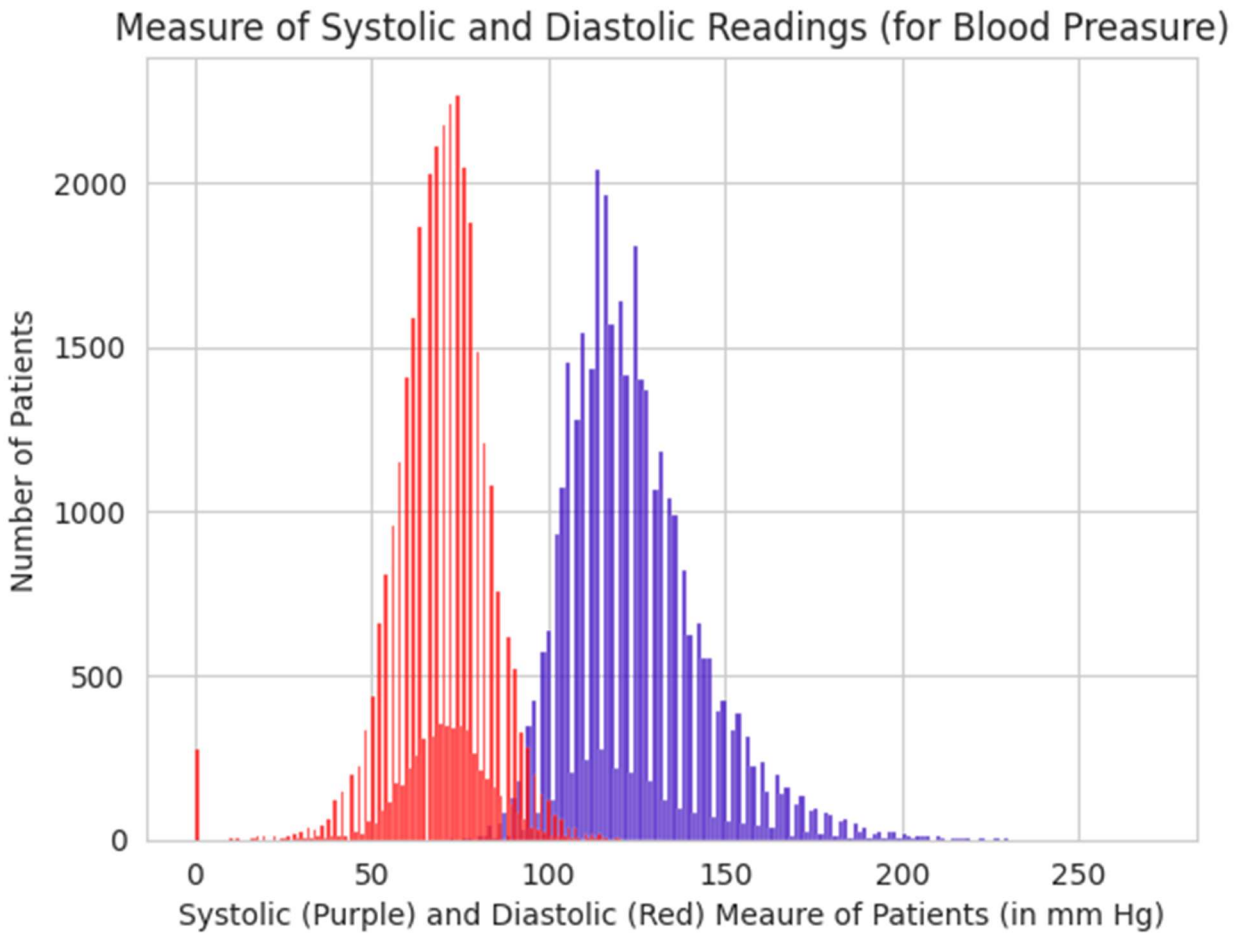
Following are the graphs (Line Graphs and Histograms) that are plotted as part of EDA and understanding the data for inferences. Most of the graphs are study of how Glucose and Cholesterol appear over various ages and histograms that help in understanding the behavior of the attributes that are used in the dataset.

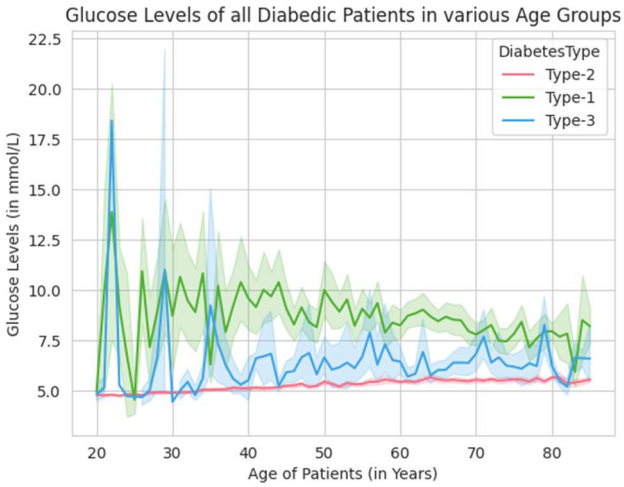

Distribution of Age among the patients
Number of Patients
0
200
400
600
800
1000
1200
1400
1600
20
30
40
50
60
70
80
Age of Patients (in Years)

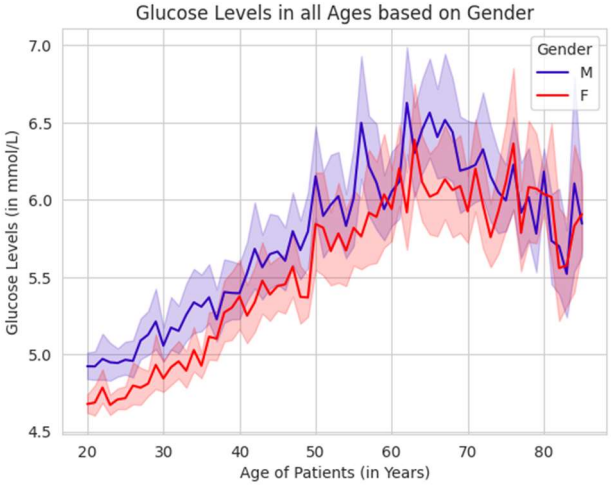

Glucose, Cholesterol and Total Cholesterol in patients acorss various Ages
Total Cholesterol
Cholesterol
Glucose
Unit Quaity found in Blood (in mmol/L)
4.5
5.0
5.5
6.0
6.5
20
30
40
50
60
70
80
Age of Patients (in Years)

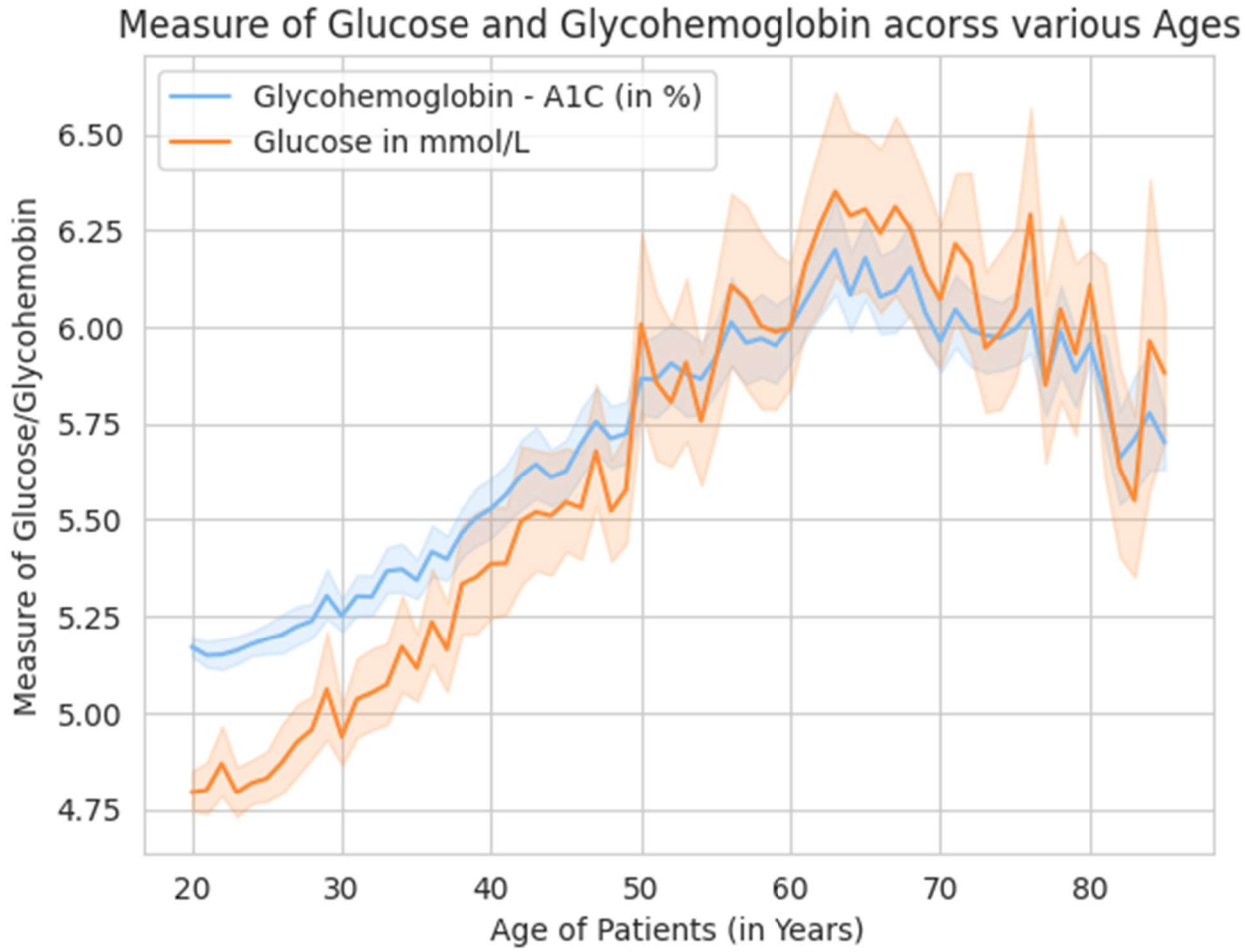

Distribution of BMI
Number of Patients
Body Mass Index (BMI) in kg/m^2
0
200
400
600
800
1000
1200
20
40
60
80
100
120


Distribution of Total Cholesterol among Patients
Number of Patients
Total Cholesterol (in mmol/L)
0
200
400
600
800
1000
1200
1400
1600
2
4
6
8
10
12
14

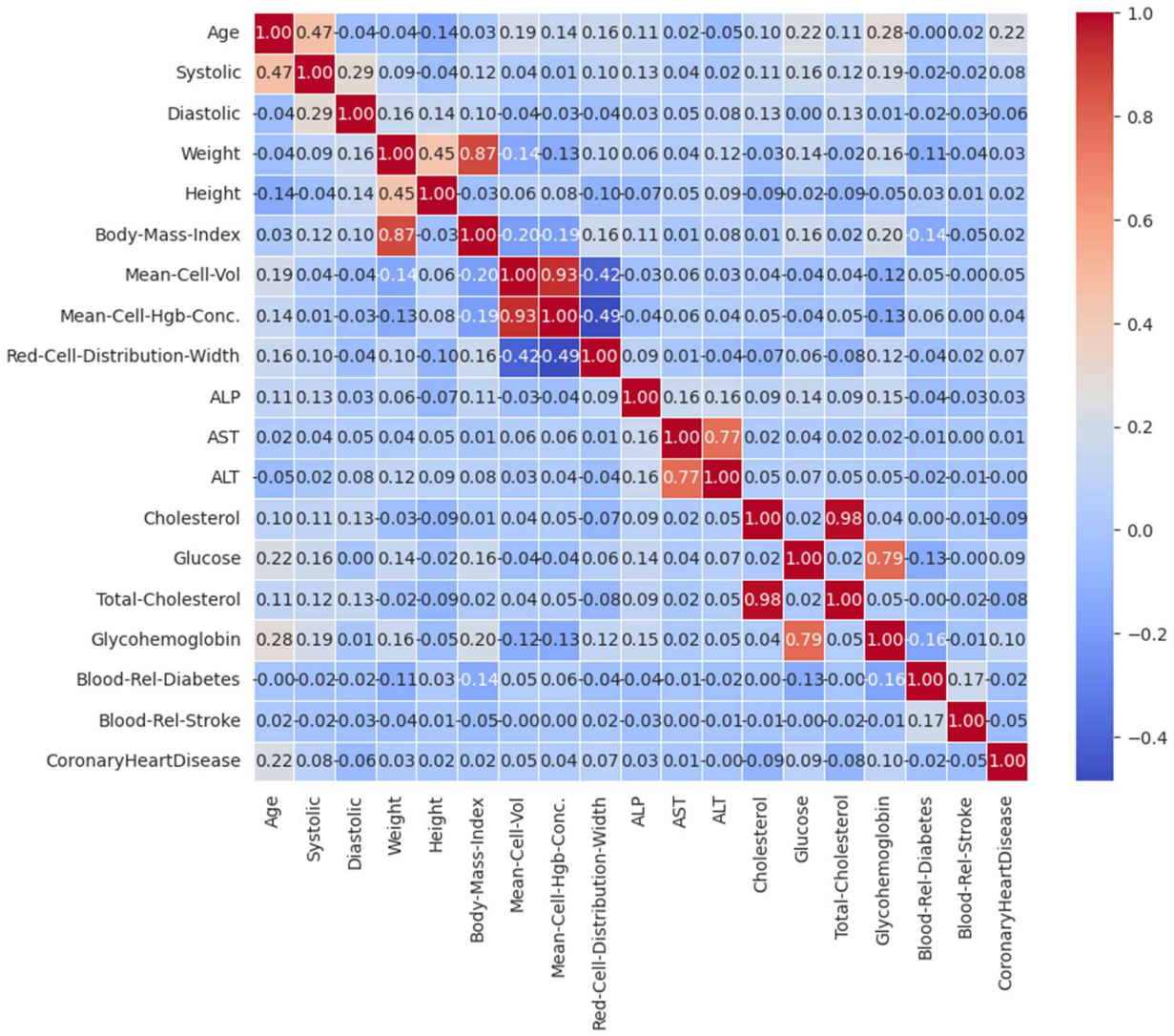

Measure of Systolic and Diastolic Readings (for Blood Preasure)
Number of Patients
Systolic (Purple) and Diastolic (Red) Meaure of Patients (in mm Hg)
0
500
1000
1500
2000
0
50
100
150
200
250


Glucose Levels of all Diabedic Patients in various Age Groups
DiabetesType
Type-2
Type-1
Type-3
Glucose Levels (in mmol/L)
Age of Patients (in Years)


Glucose Levels in all Ages based on Gender
Gender
M
F
Glucose Levels (in mmol/L)
Age of Patients (in Years)


Measure of Glucose and Glycohemoglobin acorss various Ages
Glycohemoglobin - A1C (in %)
Glucose in mmol/L
Measure of Glucose/Glycohemobin
Age of Patients (in Years)

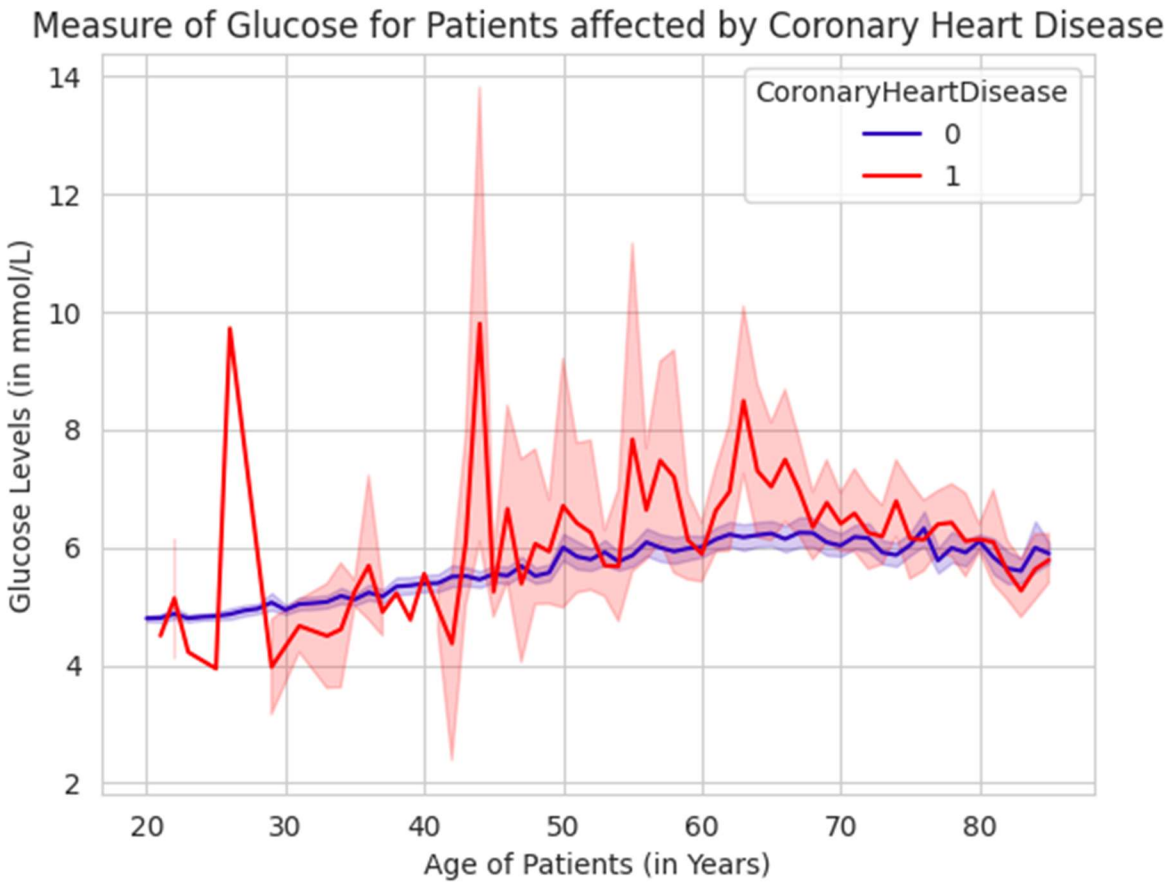
Correlation Matrix for 21 Variables:



Discussions

While performing Exploratory Data Analysis (EDA), it was observed that few of the records were out of the norms and practically possible cases under medical terms. For example, the highest value under '*Body-Mass-Index'* is '130.21'. Medically, BMI at most could be around 100 and for a healthy person, it is in the range of 18.5 – 25. This could be a case of incorrect entry, calculation error or a slight chance that data is real and an outlier. This information should be handled properly to achieve accurate results.

Also, when compared with the contents of the paper and finding, it is observed that on an average, patients who suffer Coronary Heart Disease have more Glucose Levels compared to the patients who do not suffer from this disease (ref. graph below).



Even though the above graph supports the paper, considering how data is distributed and ratio between patients with Coronary Heart Disease and without Coronary Heart Disease (1,508:35,571), we cannot strongly conclude whether the data is biased or not. We can get a conclusive statement regarding the relation and impact of diabetes on people with coronary heart disease, but the model developed needs to be evaluated further.

Conclusion

With all the data mentioned and results, we can conclude that the patients that are affected by diabetes have high chance of getting a Cardiac Arrest. This possibility is high as we observed that diabetes along with other factors have high probability of leading to the case of coronary artery disease. We can further support this statement by performing analysis using machine learning algorithms and techniques on the dataset. With a properly trained model, and testing it (to enhance the model), we can solidify the claims of relation between diabetes and cardiac arrest. This also concludes that people who are diabetic and have high glucose and cholesterol should be careful and monitor their health regularly to avoid cardiac related issues.

Future Work

Currently, only 21 variables were used to study and analyze from the given list of attributes. Further medical enhancements in the field of diabetes may help is finding the relationship between the unused attributes, thus enhancing the accuracy of the model when implemented. This can help in achieving highest accuracy for the model that is trained and tested. New and enhanced models can also be used to obtain the same. The paper which was used as reference had Random Forest Algorithm achieve 90% accuracy. We will try to obtain more accuracy or use a model that is more efficient than Random Forest Algorithm.

References

Diabetes - Symptoms and causes - Mayo Clinic. (2023, September 15). Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Bansal, N. (2015). Prediabetes diagnosis and treatment: A review. *World Journal of Diabetes*, *6*(2), 296. https://doi.org/10.4239/wjd.v6.i2.296

K. A. Hasan and M. A. M. Hasan, "Prediction of Clinical Risk Factors of Diabetes Using Multiple Machine Learning Techniques Resolving Class Imbalance," *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, DHAKA, Bangladesh, 2020, pp. 1-6, doi: 10.1109/ICCIT51783.2020.9392694.

Appendices

Sample Code Snippets for Data Processing and Visualization

- Importing Packages and Reading Dataset

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from scipy import stats as st
rd = pd.read_excel('/content/CardiacData.xlsx') #RawData as rd
#Top 10 Rows in Dataframe rd
rd.head(10)
```

- Processing

```
rd['DiabetesType'] = rd['Diabetes']
rd['DiabetesType'] = rd['DiabetesType'].map({1: "Type-1", 2: "Type-2", 3:
"Type-3"})
rd['DiabetesType']
```

- Selecting Required Columns

```
#Selecting required Columns
rd1 = rd[['Gender','Age','Systolic','Diastolic','Weight','Height','Body-
Mass-Index','Mean-Cell-Vol','Mean-Cell-Hgb-Conc.','Red-Cell-Distribution-
Width','ALP','AST','ALT','Cholesterol','Glucose','Total-
Cholesterol','Glycohemoglobin','DiabetesType','Blood-Rel-Diabetes','Blood-
Rel-Stroke','CoronaryHeartDisease']]
```

- Visualization

```
sns.barplot(x='DiabetesType', y='Age', data = rd1)
plt.xticks(rotation = 60)
plt.show()
```

```
#Histogram - Age
sns.histplot(data = rd1, x = 'Age')
plt.xlabel("Age of Patients (in Years)")
plt.ylabel("Number of Patients")
plt.title("Distribution of Age among the patients")
```

```
#Histogram - BP
sns.histplot(data = rd1, x = 'Systolic', label = 'Systolic Measure')
```

```
sns.histplot(data = rd1, x = 'Diastolic', label = 'Diastolic Measure')
plt.xlabel("Systolic (Purple) and Diastolic (Red) Meaure of Patients (in
mm Hg)")
plt.ylabel("Number of Patients")
plt.title("Measure of Systolic and Diastolic Readings (for Blood
Pressure)")
```

```
#Line Graph - Glucose + Glycohemoglobin
sns.lineplot(x = 'Age', y = 'Glycohemoglobin', data  = rd1, label =
'Glycohemoglobin - A1C (in %)', color = '#7EB9F2')
sns.lineplot(x = 'Age', y = 'Glucose', data  = rd1, label = 'Glucose in
mmol/L', color = '#FF8C38')
plt.legend(loc = "upper left")
plt.title("Measure of Glucose and Glycohemoglobin acorss various Ages")
plt.xlabel("Age of Patients (in Years)")
plt.ylabel("Measure of Glucose/Glycohemobin")
```

```
#Correlation Matrix
plt.figure(figsize=(10, 8))
sns.heatmap(cor_mat,annot=True, cmap='coolwarm', fmt=".2f",
linewidths=0.5)
```

```
#Main Line Graph - Glucose + CoronaryHeartDisease
sns.lineplot(x = 'Age', y = 'Glucose', data = rd1, hue =
'CoronaryHeartDisease')
plt.xlabel("Age of Patients (in Years)")
plt.ylabel("Glucose Levels (in mmol/L)")
plt.title("Measure of Glucose for Patients affected by Coronary Heart
Disease")
```