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

December 11, 2023

Table of Contents

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.

Methodology

Model Selection: As we are trying to find how the factors influence the occurrence of heart disease, machine learning techniques – *Decision Tress* and *Random Forest* Algorithms that is used in case of Classification.

Decision Tree

A decision tree is a graphical model that is used in machine learning to make decisions. It analyzes a complex decision-making process into a series of simpler decisions that are structured in a tree-like form. Each node represents a decision based on a particular attribute, with branches representing various outcomes. This hierarchical technique simplifies categorization, regression, and problem-solving tasks across multiple domains, while also providing transparency and interpretability in the decision-making process.

Random Forest

Random Forest is a machine learning method that builds numerous decision trees during training and outputs the mode of the classes for classification or mean prediction for regression. To increase the model's resilience and decrease overfitting, each tree is constructed using a random subset of the training set, and characteristics are chosen at random to split nodes. It improves accuracy and resilience by merging predictions from several individual trees and boosting the overall model’s performance.

Tools and Validation Techniques: To test the performance and run models, we use Python Programming and Google Colab as IDE. For implementation of algorithms, we used existing packages in python i.e. *sklearn* for dividing data into train and test datasets (to validate the efficiency of the models) and implementing Decision Tree (sklearn.tree) and Random Forest Algorithm (sklearn.ensemble), and packages seaborn, matplotlib.pyplot, numpy and pandas to visualize, load, process and handle data.

Accuracy for the Algorithms were calculated to validate and find the efficiency and compare both the algorithm’s processing and effectiveness when working on classification data related to heart diseases.

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 or heart diseases. Few attributes such as *Annual-Family-Income*, *Phosphorus, 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*,

Results

Descriptive Statistics: For the record of 37,079 patients, we used 18 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 early's 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.

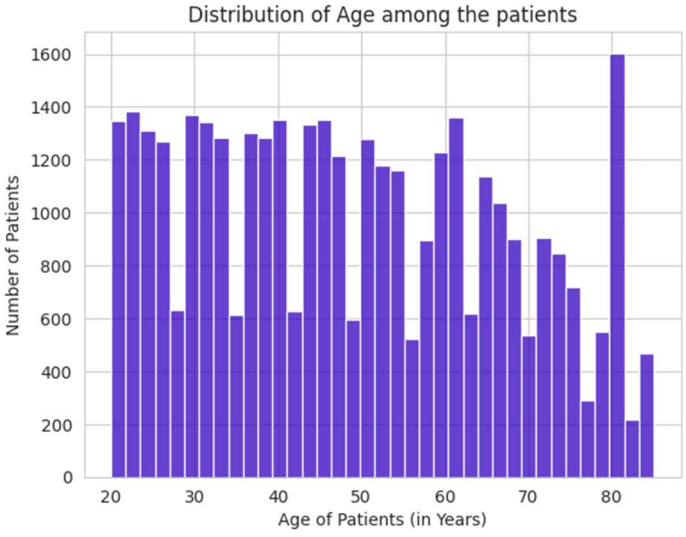
After performing machine learning, following accuracies were achieved for both Decision Trees and Random Forest algorithms:

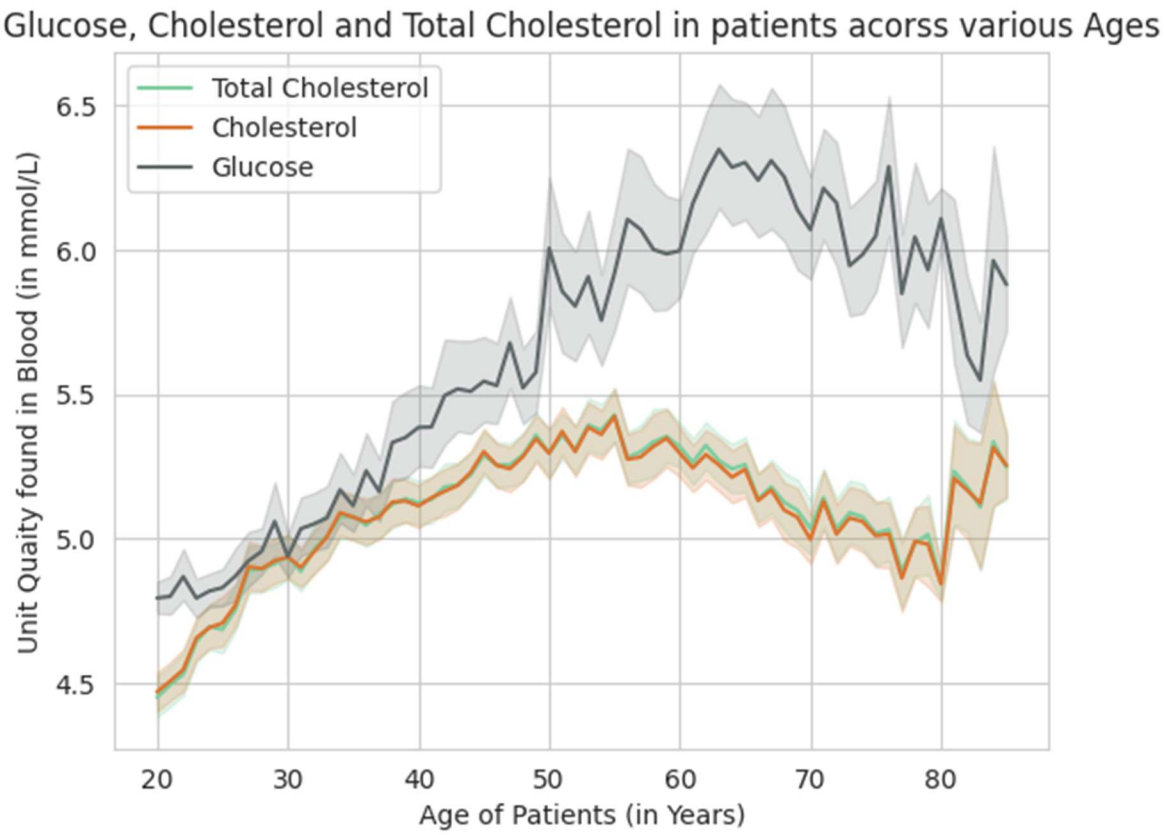
Decision Tree – 92% Accurate, stating that the results obtained when tested for a heart disease are 92% sure and there is 8% chance of it being false-positive.

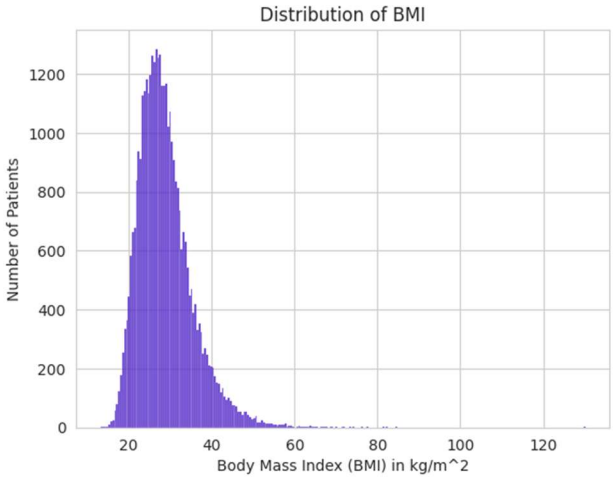
Random Forest – 96% Accurate, stating that the results obtained when tested for a heart disease are 92% sure and there is 4% chance of it being false-positive.

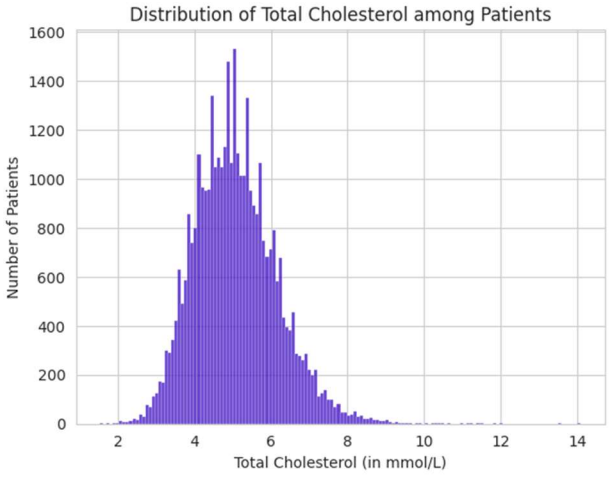
Visualizations:

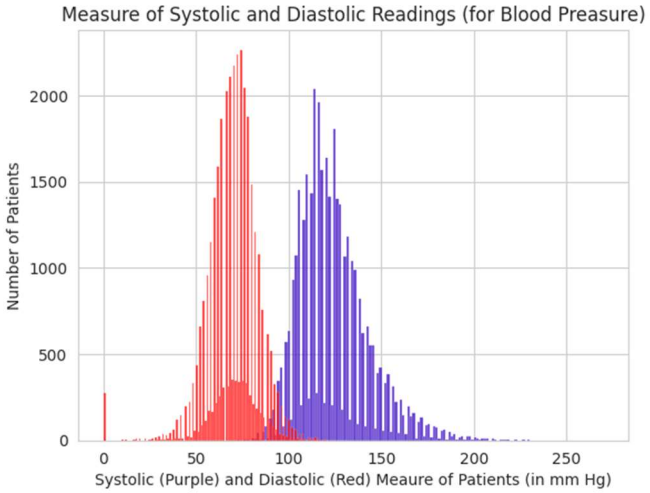
Following are the graphs (Line Graphs and Histograms) that are plotted as part of EDA and understanding the data for inferences. Most of the graphs are study of how Glucose and Cholesterol appear over various ages and histograms that help in understanding the behavior of the attributes that are used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

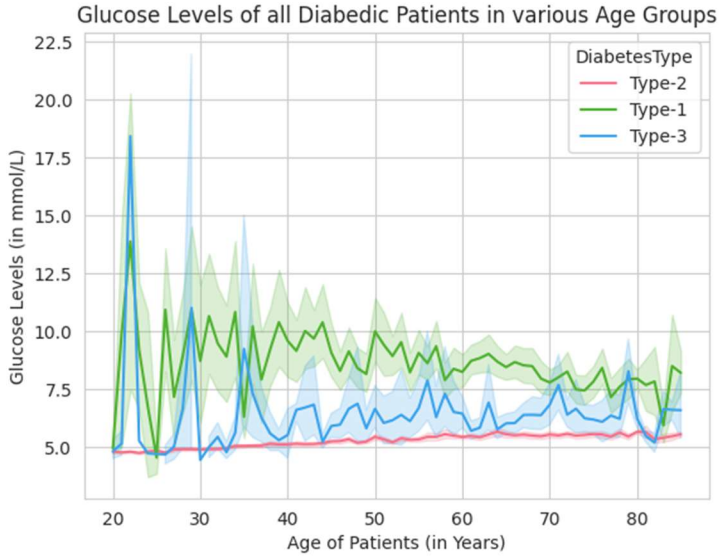


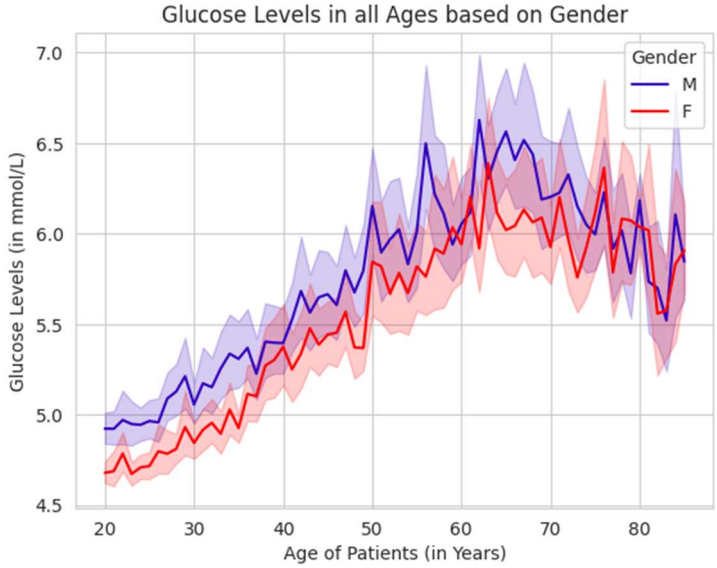

Glucose, Cholesterol and Total Cholesterol in patients acorss various Ages
Total Cholesterol
Cholesterol
Glucose
Unit Quaity found in Blood (in mmol/L)
6.5
6.0
5.5
5.0
4.5
20
30
40
50
60
70
80
Age of Patients (in Years)

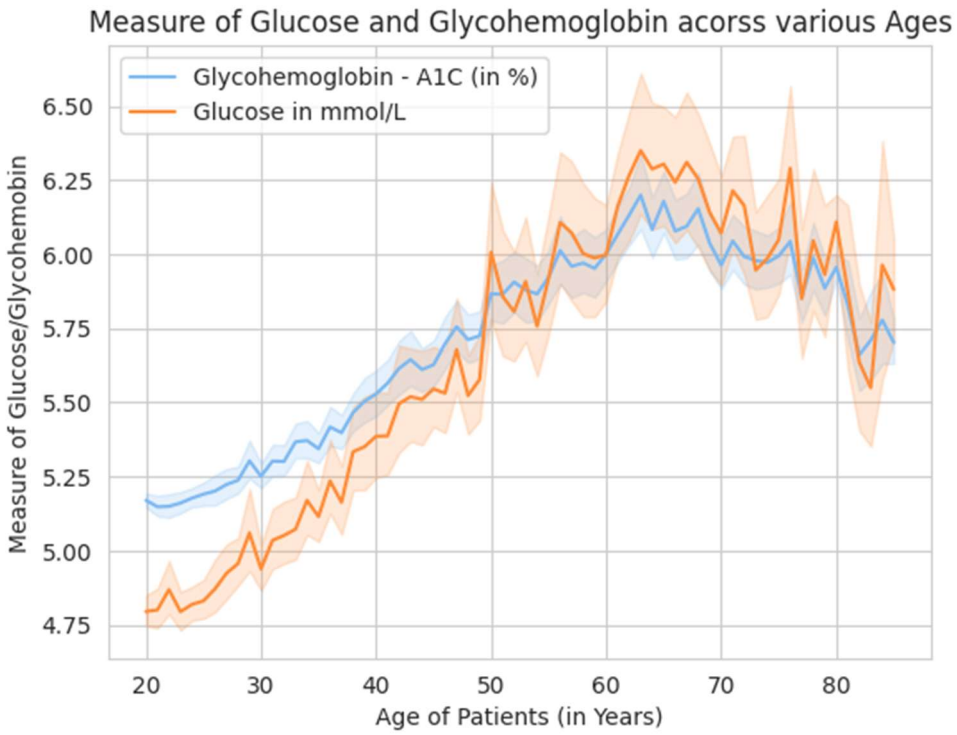

Distribution of BMI
Number of Patients
1200
1000
800
600
400
200
0
20
40
60
80
100
120
Body Mass Index (BMI) in kg/m^2


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

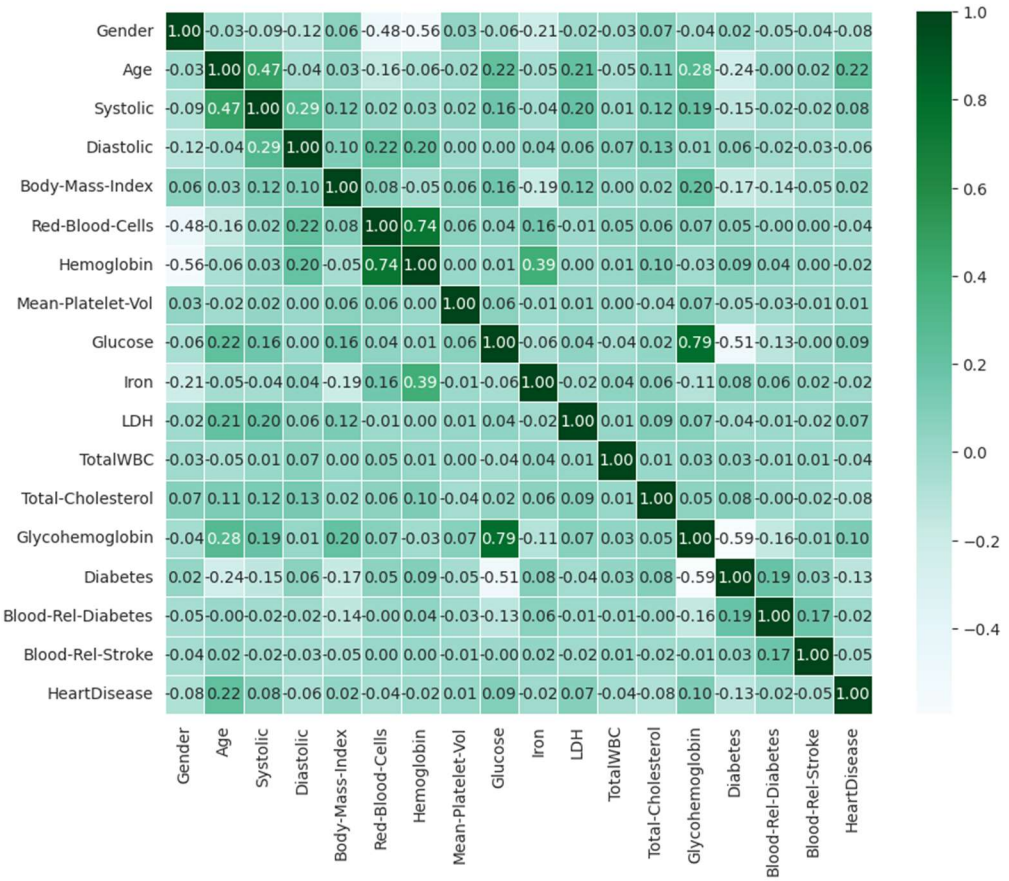

Measure of Systolic and Diastolic Readings (for Blood Preasure)
Number of Patients
2000
1500
1000
500
0
0
50
100
150
200
250
Systolic (Purple) and Diastolic (Red) Meaure of Patients (in mm Hg)


Glucose Levels of all Diabedic Patients in various Age Groups
DiabetesType
Type-2
Type-1
Type-3
Glucose Levels (in mmol/L)
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
20
30
40
50
60
70
80
Age of Patients (in Years)

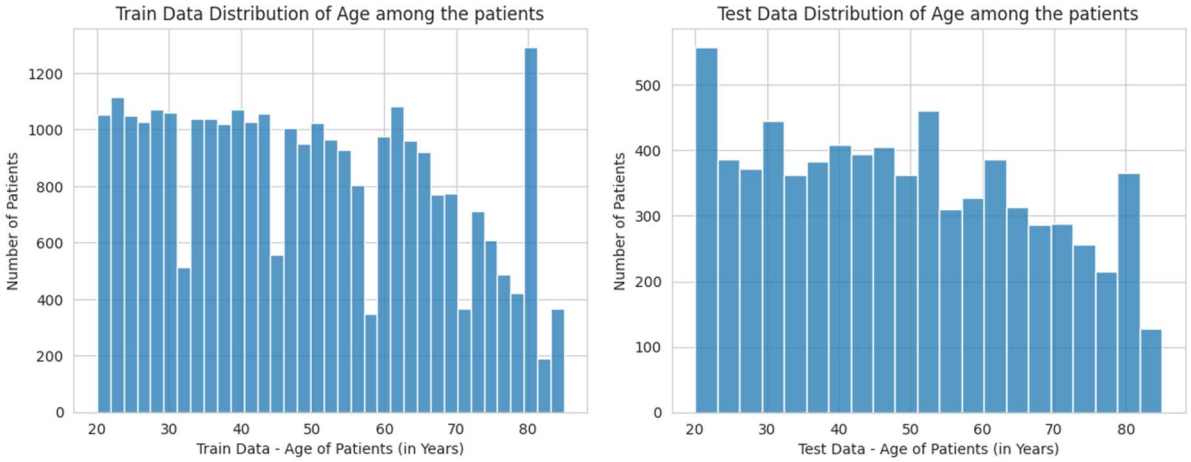

Glucose Levels in all Ages based on Gender
Gender
M
F
Glucose Levels (in mmol/L)
7.0
6.5
6.0
5.5
5.0
4.5
20
30
40
50
60
70
80
Age of Patients (in Years)

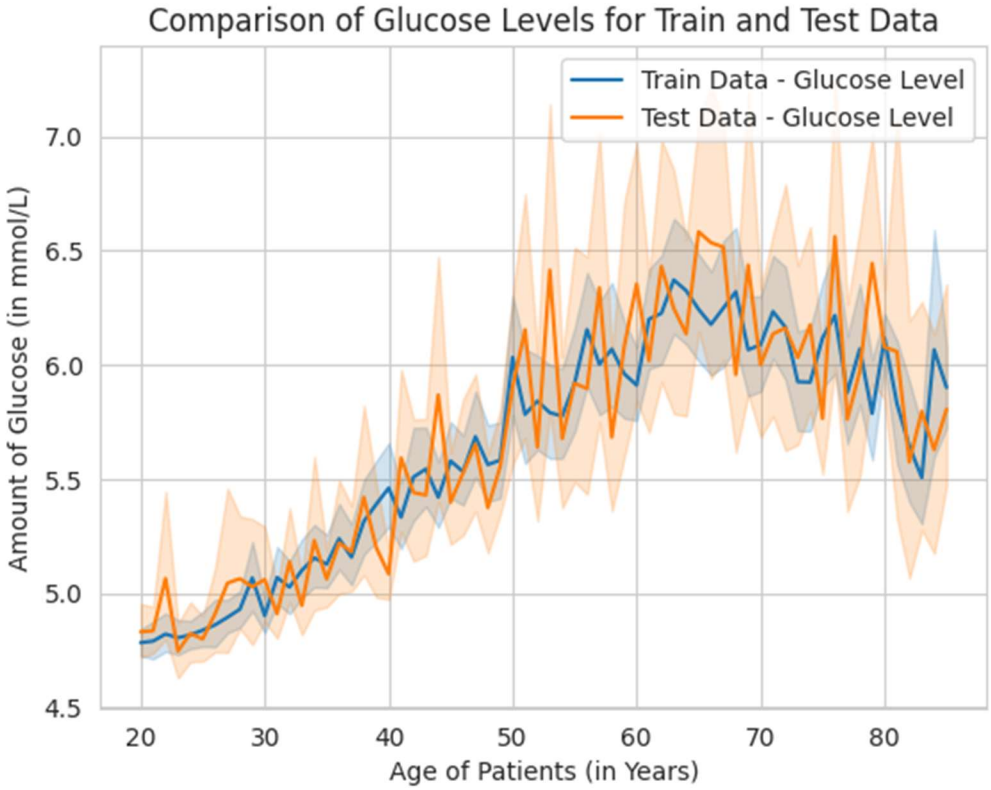


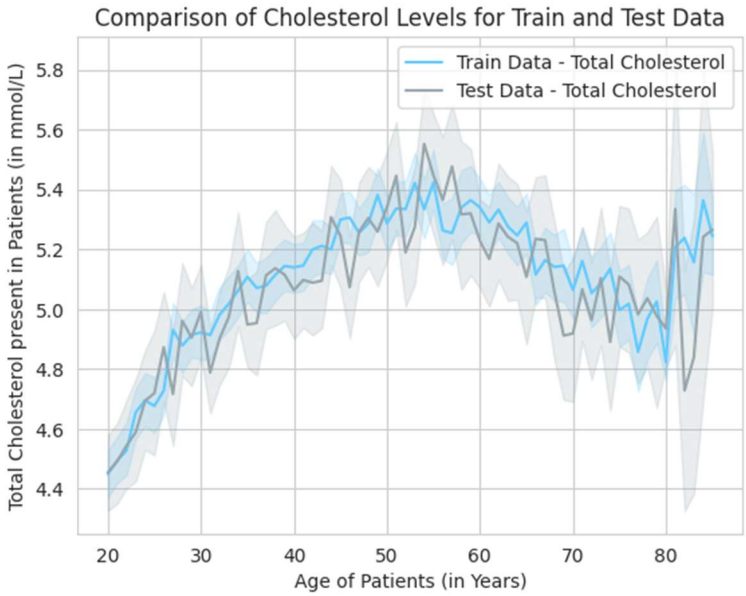
Correlation Matrix for 18 Variables:



Train and Test Data Comparisons:



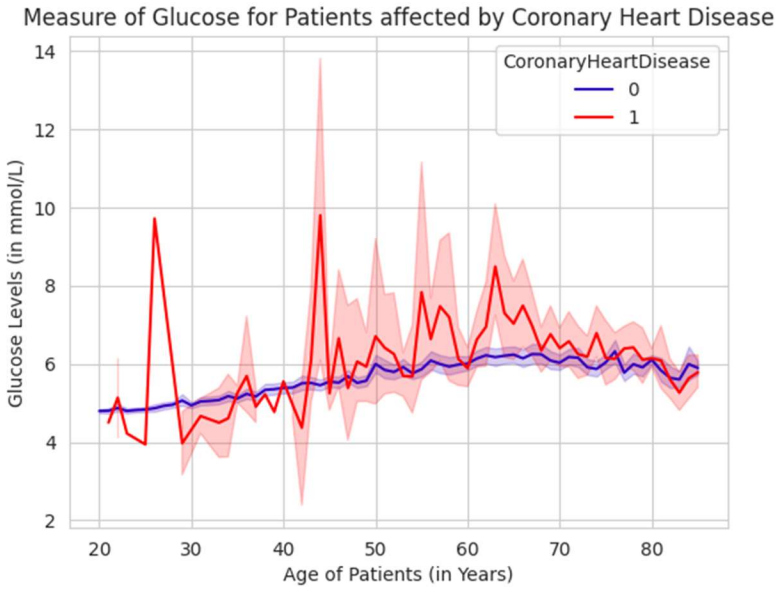




Discussions

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. It was also observed that Random Forest Algorithm achieved more accuracy than Decision Tree Algorithm.

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 18 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. Even though we had high accuracy, the other parameters like Precision, Recall and F-1 Score were not within the range of a good result. To get better results, additional parameters can be passed so that these scores can be improved and the cases of results being false-positive or true-negative can be reduced and accurate data can be provided. This is essential as we are dealing with the information related to the health of people.

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

- Loading Dataset:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

#Loading Dataset
CardiacData = pd.read_excel('/content/CardiacData.xlsx')
print(CardiacData.head())
```

- Deleting unused Columns

```
#Removing unused columns:
newData = CardiacData.drop(['SEQN','Annual-Family-Income','Ratio-Family-Income-Poverty','X60-sec-pulse','Weight','Height','Mean-Cell-Vol','Mean-Cell-Hgb-Conc.','Mean-cell-Hemoglobin','Platelet-count','Segmented-Neutrophils','Hematocrit','Red-Cell-Distribution-Width','Albumin','ALP','AST','ALT','Cholesterol','Creatinine','GGT','Phosphorus','Bilirubin','Protein','Uric.Acid','Triglycerides','HDL','Vigorous-work','Moderate-work','Health-Insurance'], axis=1)
```

- Data Preprocessing

```
#Modifying WBC
newData['TotalWBC'] = newData['White-Blood-Cells'] + newData['Lymphocyte'] + newData['Monocyte'] + newData['Eosinophils'] + newData['Basophils']
```

```
newData = newData.drop(['White-Blood-
Cells','Lymphocyte','Monocyte','Eosinophils','Basophils'], axis=1)
newData.describe
```

```
#Modifying Diabetes Type 3 to Type 2
newData['Diabetes'] = newData['Diabetes'].map({1: 1, 2: 2, 3: 2})
newData.describe
```

- Obtaining Train and Test Data

```
#Considering HeartDisease as Y and remaining data as X

X = newData.drop(['GenderType', 'DiabetesType', 'HeartDisease'], axis=1)
y = newData['HeartDisease']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=8)
featureList = list(newData.drop(['GenderType', 'DiabetesType',
'HeartDisease'], axis=1).columns)
```

- Implementing Decision Tree

```
#DecisionTreeClassifier
dtree = DecisionTreeClassifier()

# train the model using the training sets
dtree.fit(X_train, y_train)

# make predictions using the testing set
y_pred = dtree.predict(X_test)

# calculate accuracy
print(f'Accuracy: {accuracy_score(y_test, y_pred)}')
```

- Implementing Random Forest

```
from sklearn.ensemble import RandomForestClassifier

rForest = RandomForestClassifier()

rForest.fit(X_train, y_train)

y_pred2 = rForest.predict(X_test)
print(f'Accuracy: {accuracy_score(y_test, y_pred2)}')
```