# Exploratory and Predictive Analysis of Health Outcomes: Coronary Heart Disease & Diabetes

1st Dheeraj Vurukuti
*Computer Science, M.S. Student*
*Washington State University*
Pullman, WA, USA
dheeraj.vurukuti@wsu.edu
WSUID# 011810023

***Abstract*—This project report presents a comprehensive exploration and predictive analysis of health outcomes, focusing specifically on Coronary Heart Disease (CHD) and Diabetes. This research leverages advanced analytical techniques and machine learning methodologies to gain insights into the factors influencing these critical health conditions.**

**The exploratory phase involves an in-depth examination of diverse datasets encompassing demographic information, lifestyle factors, and general human health indices like BMI, age, etc. Descriptive statistics and data visualization techniques are employed to discern patterns, correlations, and potential risk factors associated with CHD and Diabetes. The subsequent predictive analysis utilizes machine learning algorithms to develop robust models capable of forecasting health outcomes based on the identified variables.**

**It contributes to the existing body of knowledge in health analytics by providing actionable insights for preventive healthcare measures. The findings are intended to aid healthcare professionals, policymakers, and researchers in understanding the dynamics of CHD and Diabetes, ultimately supporting the development of targeted interventions and personalized healthcare strategies.**

***Index Terms*—Coronary Heart Disease (CHD), Diabetes, Framingham Heart Study, Exploratory Data Analysis (EDA), Predictive Modeling, Machine Learning, Logistic regression, Random Forest, Support Vector Machines (SVM), kNN, Decision Trees, Correlation Analysis, Visualization**

## I. Introduction

Chronic diseases, such as coronary heart disease (CHD) and diabetes, pose a significant burden on individuals and healthcare systems worldwide. Understanding the risk factors associated with these conditions and developing accurate predictive models can help clinicians identify individuals at high risk and implement preventive measures.

The intersection of data science and healthcare has proven to be a transformative field, offering unprecedented insights into complex health phenomena. This project, represents a comprehensive exploration of health data through the lens of data science methodologies. Over the course of 1.5 months, our focus has been on conducting extensive exploratory data analysis (EDA) and implementing five distinct classification algorithms to analyze the Framingham Heart Study dataset and a dataset dedicated to predicting diabetes.

Cardiovascular diseases, such as Coronary Heart Disease (CHD), and metabolic disorders like diabetes, stand as significant public health challenges worldwide. Our primary objective is to leverage advanced data science techniques to unravel patterns, relationships, and predictive insights within these datasets. The Framingham Heart Study dataset, renowned for its longitudinal design and rich array of health-related variables, serves as a valuable resource for understanding cardiovascular risk factors and outcomes. Simultaneously, the diabetes prediction dataset contributes to our understanding of predictive modeling for a prevalent metabolic disorder.

Our project's methodology encompasses a two-fold approach: first, an in-depth exploratory analysis is undertaken to comprehend the inherent structures and characteristics of the datasets. Subsequently, five diverse classification algorithms are applied to predict health outcomes. These algorithms include machine learning models like logistic regression, decision trees, random forests, support vector machines, and kNN.

Through the amalgamation of advanced statistical techniques and machine learning algorithms, this project seeks to not only identify key factors influencing health outcomes but also build robust predictive models capable of aiding healthcare practitioners in early risk assessment and intervention strategies. The outcomes of this research are anticipated to contribute valuable insights to the broader field of health informatics, bridging the gap between data science methodologies and real-world healthcare applications.

## II. Problem Definition

The increasing prevalence of cardiovascular diseases, particularly coronary heart disease (CHD), and metabolic disorders like diabetes pose significant challenges to public health worldwide. Identifying the intricate relationships between various risk factors and health outcomes is crucial for devising effective preventive strategies and personalized interventions.

The primary problem addressed in this project is twofold: first, to conduct an in-depth exploratory analysis of the Framingham Heart Study dataset to uncover patterns and associations related to CHD, and second, to extend this analysis to a dedicated diabetes prediction dataset. The goal is to develop predictive models capable of identifying key risk factors and making accurate predictions regarding the likelihood of CHD and diabetes. Addressing these problems requires the implementation of advanced data science methodologies, including exploratory data analysis (EDA) and the application of diverse

classification algorithms, to unravel the complex interactions between variables and health outcomes. Ultimately, the project aims to contribute valuable insights to the understanding of cardiovascular health and diabetes, fostering the development of data-driven strategies for risk assessment and healthcare.

## III. Models, Measures, and Algorithms

We've used several techniques to perform a comprehensive analysis on the chronic health conditions mentioned above. These include:

### A. *Exploratory Data Analysis*

We did comprehensive Exploratory Data Analysis (EDA) to gain insights into the structure and characteristics of the Framingham Heart Study and diabetes prediction datasets.

Exploratory Data Analysis (EDA) is an approach to analyzing and visualizing data sets in order to summarize their main characteristics, often with the help of statistical graphics and other data visualization methods. The primary goal of EDA is to understand the structure of the data, identify patterns, relationships, anomalies, and gain insights that can inform further analysis or guide decision-making processes. Descriptive statistics, data visualization techniques (histograms, scatter plots, and correlation matrices), and feature engineering are employed to uncover patterns, missing values, and outliers within the datasets. EDA serves as a crucial step in understanding the data's inherent complexity, setting the stage for subsequent modeling. Fig 1. shows the basic blocks of EDA.

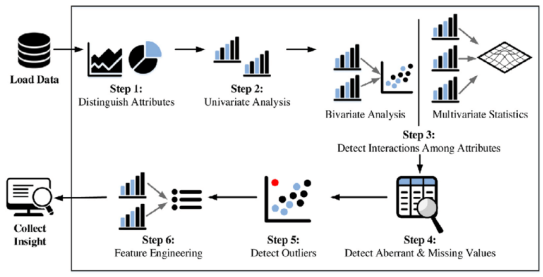



Fig. 1. Workflow of EDA

*1) **Descriptive Statistics**:* Descriptive statistics provide a summary of the main characteristics of each variable in the dataset. This includes measures such as mean, median, standard deviation, minimum, and maximum. Descriptive statistics give an initial overview of the central tendency and dispersion of the data, helping to identify outliers and potential issues.

*2) **Data Visualization**:* Data visualization techniques include creating charts and graphs to represent the distribution and relationships within the dataset visually. Histograms are used to display the distribution of continuous variables, box plots show the spread and central tendency, and scatter plots illustrate relationships between two variables. Visualization aids in identifying patterns, outliers, and potential correlations, making complex datasets more accessible. Fig 2. shows different data visualization charts.


Fig. 2. Data Visualization Techniques

*3) **Correlation Matrices**:* Correlation matrices display the pairwise relationships between variables using correlation coefficients. Positive values indicate a positive correlation, negative values indicate a negative correlation, and values close to zero suggest little or no correlation. This technique helps identify potential multicollinearity issues and understand how variables interact with each other. Fig 3. illustrates an example of a correlation matrix.

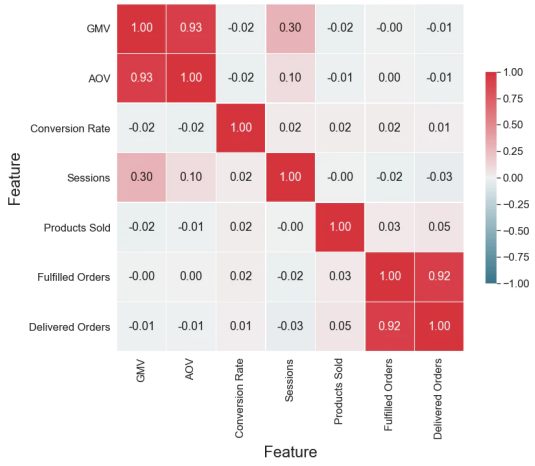



Fig. 3. Example of Correlation Matrix

### B. *Predictive Analysis*

Predictive analysis is the process of using data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal of predictive analysis is to make predictions or forecasts about future events or trends. This analysis involves identifying patterns in historical data and using these patterns to predict future behavior or outcomes.

**Classification Algorithms:** We implemented five distinct classification algorithms to predict health outcomes in both the Framingham Heart Study and diabetes prediction datasets. These algorithms include:

*1) **Logistic Regression** :* Logistic Regression is a statistical method used for binary classification problems, where the outcome variable has two categories. Despite its name, it's

employed for classification, not regression. It models the probability that an instance belongs to a particular category. Fig 4. below illustrates the process of the model.

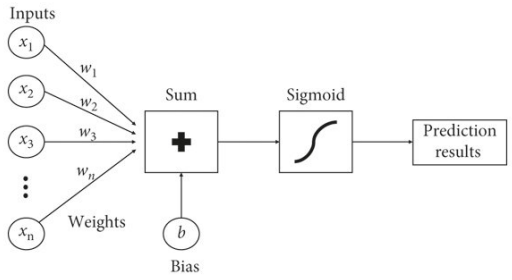



Fig. 4. Block diagram of Logistic Regression

It uses the logistic function (sigmoid function) to map input features into a range between 0 and 1, representing probabilities. A decision threshold is then applied to classify instances into one of the two categories.

*2)* ***Decision Trees*** *:* Decision Trees are tree-like structures that recursively split the dataset based on feature values, creating a hierarchy of decision rules. They are versatile and can be used for both classification and regression tasks. Fig 5. below illustrates the process of the model in the field of marketing.

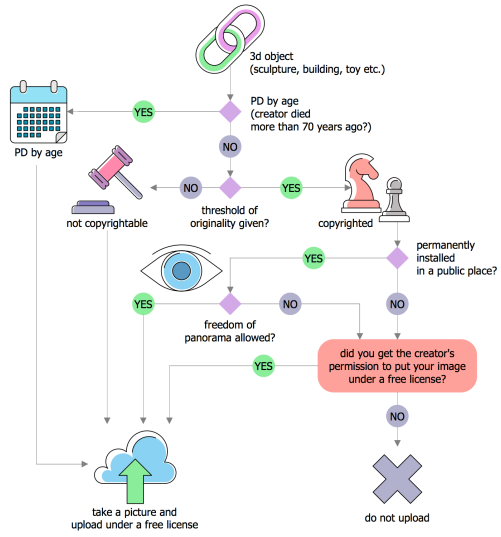



Fig. 5. Working of Decission Tree in a marketing strategy business.

The algorithm selects the feature that best separates the data at each node, based on criteria such as Gini impurity or information gain. This process continues until a stopping criterion is met, resulting in a tree where leaves represent the predicted classes.

*3)* ***Random Forest*** *:* Random Forests are an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and robustness. They are particularly effective in reducing overfitting. Fig 6. below illustrates the process of the model.

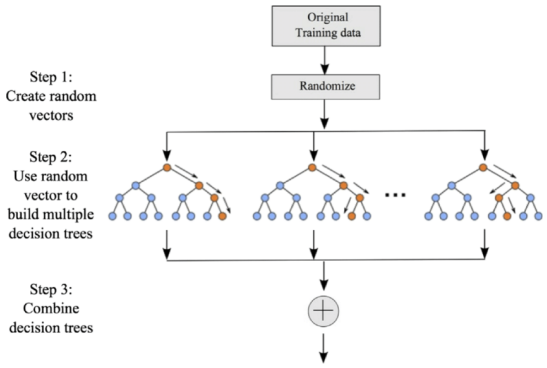



Fig. 6. Workflow of Random Forest using Decision Trees

Random Forests create multiple decision trees by training on different subsets of the data and using a random subset of features for each split. The final prediction is often a majority vote or an average of the predictions from individual trees.

*4)* ***Support Vector Machine*** *:* Support Vector Machines are powerful supervised learning algorithms used for classification and regression tasks. SVM aims to find the hyperplane that best separates the data points into different classes. Fig 7. below illustrates the process of the model.

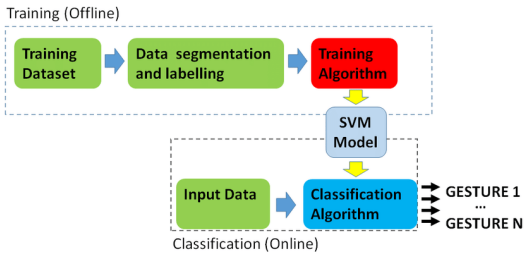



Fig. 7. BLock Diagram showing SVM Algorithm

SVM seeks to maximize the margin between classes, where the margin is the distance between the hyperplane and the nearest data points of each class. The algorithm can also handle non-linear separation by using kernel functions to map data into a higher-dimensional space.

*5)* ***kNN*** *:* k-Nearest Neighbors is a simple and intuitive classification algorithm that makes predictions based on the majority class of the k-nearest data points in the feature space. It is a non-parametric, instance-based learning algorithm. Fig 8. below illustrates the process of the model.

Given a new data point, kNN identifies the k training instances in the dataset that are closest to it in terms of feature similarity. The class label of the majority of these k neighbors is then assigned to the new data point. Distance metrics such as Euclidean distance or Manhattan distance are commonly used to measure similarity.

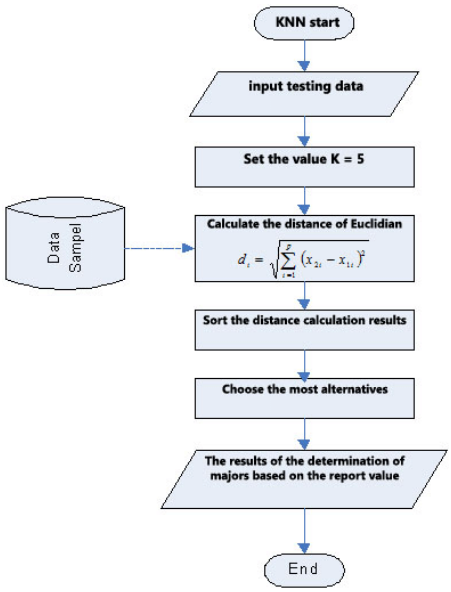



Fig. 8. Flowchart of kNN Algorithm

### C. *Model Evaluation Measures*

To assess the performance of the classification algorithms, we employed a set of evaluation measures. Some include:

*1)* ***Accuracy****:* Accuracy is the ratio of correctly predicted instances to the total instances in the dataset. It provides a general measure of the model's correctness. Figure 9 below illustrates the same.

*2)* ***Cross-Validation Score****:* Cross-validation score is a metric used to evaluate the performance of a machine learning model by assessing its generalization capabilities on different subsets of the dataset. One common method for computing the cross-validation score is through k-fold cross-validation.

*3)* ***Precision****:* Precision is the ratio of correctly predicted positive observations to the total predicted positives. It measures the accuracy of positive predictions. Figure 9 below illustrates the same.

Formula: Precision = True Positives / (True Positives + False Positives)

*4)* ***Recall****:* Recall is the ratio of correctly predicted positive observations to all actual positives. It measures the model's ability to capture all instances of the positive class. Figure 9 below illustrates the same.

Formula: Recall = True Positives / (True Positives + False Negatives)

*5)* ***F1 Scores****:* The F1 Score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives.

Formula: F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Considerations: The F1 Score is particularly useful when there is an uneven class distribution, as it balances the trade-off between precision and recall.

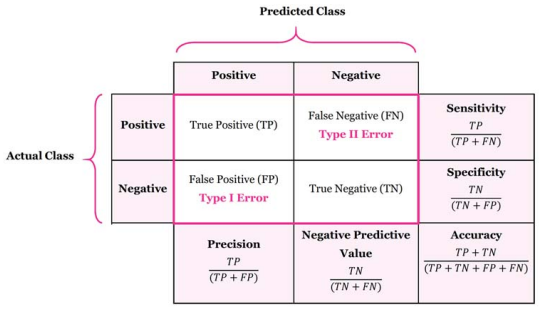



Fig. 9. Matric showing calculation of Accuracy, Precision, and Recall

## IV. DATASET

Our project utilizes publicly available datasets from Kaggle for the analysis of chronic health outcomes like Choronary Heart Diseas and Diabetes:

### A. *Framingham Heart Study:*

**Source:**https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset Fig 10. below shows the dataset.

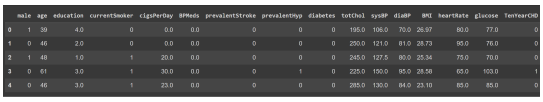


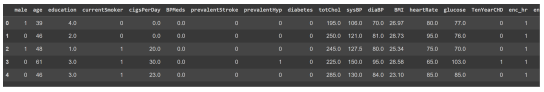

Fig. 10. Framingham Heart Study Dataset

**About:** The Framingham Heart Study, initiated in 1948, is one of the longest-running and most comprehensive cardiovascular studies globally. It is is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in the community of Framingham, Massachusetts. This Study was a landmark study in epidemiology in that it was the first prospective study of cardiovascular disease and identified their joint effects FHS Longitudinal Data Document. It spans multiple generations and has been instrumental in identifying risk factors associated with cardiovascular diseases, including CHD. The dataset comprises a wealth of health-related information, including demographic details, medical history, lifestyle factors, and clinical measurements, making it an invaluable resource for exploratory and predictive analysis in the context of health outcomes. This dataset includes data on over 4,240 participants and has 16 attributes.

**Relavance:** This dataset provides a well-studied and validated resource for researching CHD risk factors and developing predictive models. The longitudinal nature of the data enables the analysis of temporal relationships and disease progression.

### B. *Diabetes Prediction:*

**Source:**https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset Fig 11. below shows the dataset.



| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | diabetes | totChol | sysBP | diaBP | BMI | heartRate | glucose | TenYearCHD | enc_br |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 39 | 4.0 | 0 | 0.0 | 0.0 | 0 | 0 | 0 | 195.0 | 106.0 | 70.0 | 26.97 | 80.0 | 77.0 | 0 | 1 |
| 1 | 0 | 46 | 2.0 | 0 | 0.0 | 0.0 | 0 | 0 | 0 | 250.0 | 121.0 | 81.0 | 28.73 | 95.0 | 76.0 | 0 | 1 |
| 2 | 1 | 48 | 1.0 | 1 | 20.0 | 0.0 | 0 | 0 | 0 | 245.0 | 127.5 | 80.0 | 25.34 | 75.0 | 70.0 | 0 | 1 |
| 3 | 0 | 61 | 3.0 | 1 | 30.0 | 0.0 | 0 | 1 | 0 | 225.0 | 150.0 | 95.0 | 28.58 | 65.0 | 103.0 | 1 | 1 |
| 4 | 0 | 46 | 3.0 | 1 | 23.0 | 0.0 | 0 | 0 | 0 | 285.0 | 130.0 | 84.0 | 23.10 | 85.0 | 85.0 | 0 | 1 |

Fig. 11. Diabetes Prediction Dataset

**About:** This dataset is specifically curated for predicting the likelihood of diabetes occurrence, a prevalent metabolic disorder with significant public health implications. It incorporates a range of features related to patients' health and lifestyle, enabling the development of predictive models for early diabetes detection. It contains data of patients diagnosed with diabetes and non-diabetic individuals. The dataset consisting age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The target variable is the presence or absence of diabetes.

**Relavance:** This dataset provides a readily available resource for understanding diabetes risk factors and developing predictive models. The diverse patient population allows forgeneralizable model development and application.

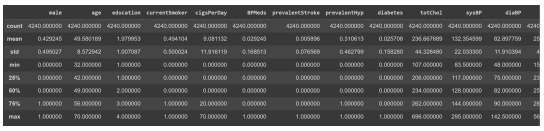
*The combination of these datasets provides a robust foundation for our exploratory and predictive analysis, allowing us to glean insights into the intricate relationships between various health factors and the respective outcomes associated with CHD and diabetes. The richness and diversity of the data empower our data science methodologies to uncover patterns, correlations, and predictive insights essential for advancing our understanding of health outcomes.*

## V. Exploratory Data Analysis (EDA)

We performed an in-depth exploratory analysis on the 2 health related datasets mentioned above and below is our analysis for the same -

### A. *Coronary Heart Disease*

*1)* ***Descriptive Statistics****:* Fig 12. below illustrates the output of the descriptive statistics of dataset



| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | diabetes | totChol | sysBP | diaBP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 |
| mean | 0.429245 | 49.580189 | 1.979953 | 0.494104 | 9.081132 | 0.029245 | 0.005896 | 0.310613 | 0.025708 | 236.867689 | 132.354599 | 82.897759 |
| std | 0.495027 | 8.572942 | 1.007087 | 0.500024 | 11.916119 | 0.168513 | 0.076569 | 0.462799 | 0.158260 | 44.328490 | 22.033300 | 11.910394 |
| min | 0.000000 | 32.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 107.000000 | 83.500000 | 48.000000 |
| 25% | 0.000000 | 42.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 206.000000 | 117.000000 | 75.000000 |
| 50% | 0.000000 | 49.000000 | 2.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 234.000000 | 128.000000 | 82.000000 |
| 75% | 1.000000 | 56.000000 | 3.000000 | 1.000000 | 20.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 262.000000 | 144.000000 | 90.000000 |
| max | 1.000000 | 70.000000 | 4.000000 | 1.000000 | 70.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 696.000000 | 295.000000 | 142.500000 |

Fig. 12. Descriptive Statistical Analysis of Framingham dataset

From the output we gathered, certain observations about some key feature points can be made:

- **Age :** We can see that minimum age of subject found in given records is 32 while Max. being 70. So our values are ranging from 32 to 70.
- **cigsPerDay :** Subject smoking Cigarettes per day is as low as nill while we have 70 Cigs. per day making the Peak.
- **totChol :** Minimum Cholesterol level recorded in our dataset is 107 while Max. is 696.
- **sysBP :** Minimum Systolic Blood Pressure observed in Subject is 83 while Max. is 295.
- **diaBP :** Minimum Diastolic Blood Pressure observed in Subject is 48 while Max. is 142.
- **BMI :** Body Mass Index in our dataset ranges from 15.54 to 56.
- **heartRate :** Observed Heartrate in our case study is 44 to 143.
- **glucose :** Glucose sugar level range is 40 to 394.

*2)* ***Correlation Matrix****:* Figure 13. below depicts a visualization of the correlation between the 16 feature points of the data we used.

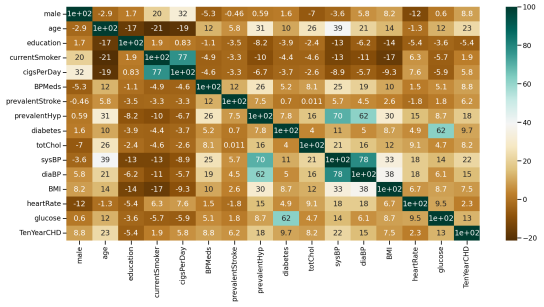



Fig. 13. Correlation Matrix of CHD

Correlation plot gives us valuable information regarding Relation within Attributes. It can either be Negative or Positive or Null. We need to always keep 1 feature from 2 strongly correlated ones but since we want to perform EDA we'll keep all and drop them before modelling.

- **currentSmoker and cigsPerDay** has strong correlation of 77 (Scaled for better Observations)
- **prevalentHyp vs sysBP / diaBP** are having positive correlation of 70 and 62.
- **glucose and diabetes** are postively correlated.
- **sysBP and diaBP** are also having positive correlation.

In handling missing values within the dataset, conventional methods involve imputing null values with measures of central tendency such as mean, median, or mode. Alternatively, techniques like forward or backward filling are commonly employed. However, in our specific case, we can leverage insights from the correlation plot to inform a more context-aware imputation strategy. For instance, by examining the positive correlation between variables like 'currentSmoker' and 'cigsPerDay,' where 'currentSmoker' takes values of 1 (indicating a smoker) or 0 (indicating a non-smoker), we can capitalize on this relationship. Through a groupby operation on 'currentSmoker,' we can impute missing values based on the median value within each group. Similarly, for a variable like 'BMI,' we can take into account the correlation with 'Gender' and 'Age.' This enables a more tailored imputation approach, enhancing the accuracy and relevance of the imputed values based on the inherent relationships observed in the data.

*3) **Agewise Distribution of samples**:* The graph in Fig 14. shows the age-wise distribution of the patients.

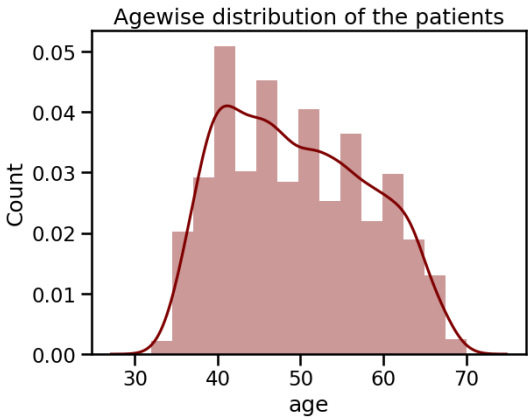



Fig. 14. Agewise Distribution of dataset samples: patients

The x-axis represents the age of the patients, and the y-axis represents the count of patients. The graph shows that the majority of the patients in the dataset are between the ages of 40 and 70. According to the above graph we can say that majority lie between 40-50 followed by 50-70. The number of patients gradually decreases before and after this age range. This is consistent with the fact that cardiovascular disease is more common in older adults.

*4) **Distribution of key feature points (Univariate Analysis)**:* The distribution plot in Fig 15. illustrates the Density distribution in a univariate analysis.

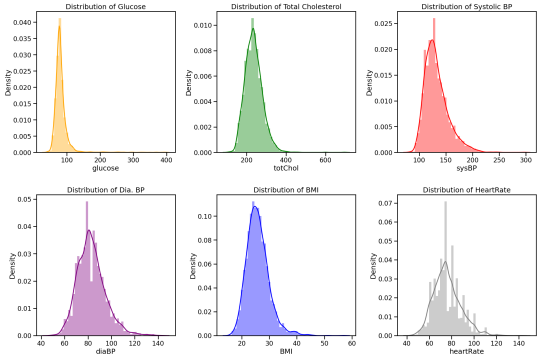



Fig. 15. Density Distribution

From the plot we can make the following observations about some features -

- **Glucose:** The distribution of glucose levels is skewed to the right, with a median of 98 mg/dL and a long tail extending to higher values. This indicates that a majority of the participants have normal glucose levels, but there is a significant minority with elevated glucose levels.
- **Total cholesterol:** The distribution of total cholesterol levels is also skewed to the right, with a median of 219 mg/dL and a tail extending to higher values. This indicates that a majority of the participants have normal cholesterol levels, but there is a significant minority with high cholesterol levels.
- The distribution of **systolic blood pressure** is right-skewed, with a median of 120 mmHg and a long tail extending towards higher values.
- The distribution of **body mass index (BMI)** is right-skewed, with a median of 25 kg/m$\hat{2}$ and a long tail extending towards higher values.

  *The right-skewed nature of the above features indicates that the majority of individuals have values within a certain range, but there is a significant tail towards higher values. This suggests that a substantial portion of the population has elevated levels of these risk factors, which could increase their risk of developing cardiovascular disease.*

- The distribution of **diastolic blood pressure** is close to Normal/Gaussian distribution, with a median of 80 mmHg. This means a lot of subjects in the dataset have a normal diastolic blood pressure.
- The distribution of **heart rate** is approximately normal, with a median of 70 beats per minute (bpm). This indicates that the majority of individuals have heart rates within a healthy range.

Overall, the density graph suggests that a significant proportion of the participants have elevated levels of glucose, total cholesterol, systolic BP and BMI. These are all risk factors for cardiovascular disease and other chronic diseases.

*5) **CHD Vs. Gender (Bivariate Analysis)**:* The plot in Fig 16. introduced the comparision of CHD with gender

The bivariate analysis plot illustrates the distribution of individuals with and without Chronic Heart Disease (CHD) based on gender. Notably, there is a considerable prevalence of individuals who do not have CHD. Specifically, around 80 to 90% of females and approximately 60 to 70% of males fall into the negative category, indicating the absence of CHD. In contrast, the positive category, signifying individuals with CHD, encompasses roughly 10% of both females and males.

This pattern reveals a notable class imbalance within our dataset, where the majority, ranging from 80 to 90%, corresponds to negative classifications (no CHD), while the positive classifications (presence of CHD) constitute approximately 10 to 15%. Understanding this class distribution is crucial for model development and evaluation, as it highlights the need for strategies to address potential biases and optimize predictive performance in the context of imbalanced data.

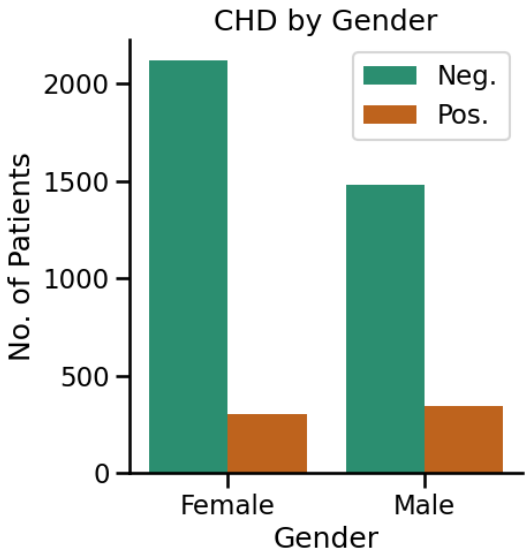



Fig. 16. Correlation Matrix of CHD

*6) **BP Analysis**:* The below Fig 17. shows a multivariate analysis for BP.

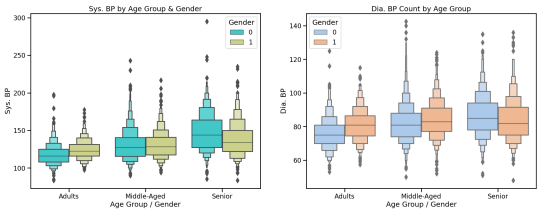



Fig. 17. (a) Sys. BP Vs. Age Group and Gender (b)Dia. BP Cunt Vs. Age Group

We can boserve the systolic blood pressure for each age group. The median systolic blood pressure increases with age, for both men and women. For example, the median systolic blood pressure for men aged 40-49 years is 110 mmHg, while the median systolic blood pressure for men aged 70-79 years is 130 mmHg. A similar trend can be seen for women.

It is also to be noted that men have higher systolic blood pressure than women, on average. For example, the median systolic blood pressure for men aged 60-69 years is 130 mmHg, while the median systolic blood pressure for women aged 60-69 years is 120 mmHg.

In conclusion, the systolic blood pressure (BP) distribution by age group and gender. The x-axis shows the age group, and the y-axis shows the systolic blood pressure. The graph shows that systolic blood pressure increases with age, with older adults having higher systolic blood pressure than younger adults. The graph also shows that men have higher systolic blood pressure than women, on average. This means that the prevalence of diabetes increases with age and that men have a higher prevalence of diabetes than women, on average.

*7) **Glucose and Cholestrol Analysis**:* The below Fig 18. shows a multivariate analysis.

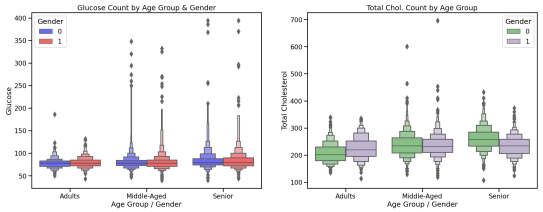



Fig. 18. (a) Glucose Count Vs. Age Group and Gender (b)Total Cholestrol Count Vs. Age Group

- **Glucose Count by Age Group & Gender:** A noticeable trend emerges as age increases, indicating a corresponding increase in the count of glucose levels. When examining gender-specific distributions, the median glucose count appears quite similar, with a few outliers present in each gender group.
- **Total Cholesterol by Age Group & Gender:** After excluding outliers, a distinct pattern emerges in the relationship between age, gender, and total cholesterol levels. For females, there is a discernible increase in cholesterol levels with age, as evidenced by the quantile values (25%, 50%, 75%). On the other hand, for males, the quantile values for cholesterol levels remain relatively consistent across different age groups, indicating a more stable cholesterol distribution compared to females.

*8) **Cigarettes per day Analysis**:* The below Fig 19. shows a a graph for Cigarettes per day by Age Group.

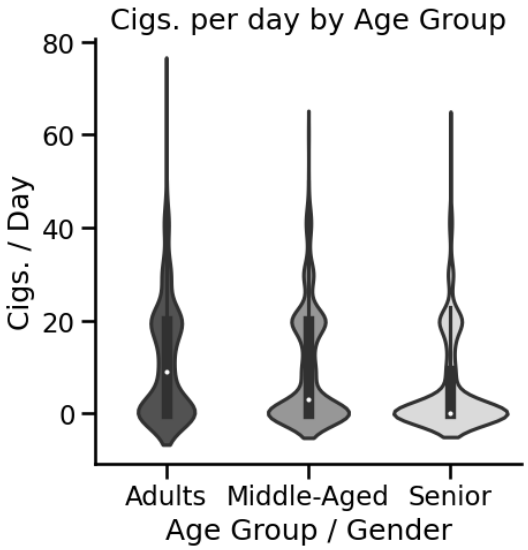



Fig. 19. Cigarettes per day Vs. Age Group

The above graph can be inferred as below -

- **Adults:** Within the adult age group, it is evident that the median values exhibit a lower kernel density, followed by

the 75% interquartile range's (IQR) density. Conversely, the 25% IQR marks the higher kernel density, indicating a specific distribution pattern in this age category.

- **Middle-Aged:** Notably, in the middle-aged group, there is a distinct distribution pattern. The 25% IQR and median show higher kernel density, while the 75% IQR exhibits a relatively lower kernel density. This suggests a unique density distribution among the quartiles within the middle-aged demographic.
- **Senior:** Examining the senior age group, a distinctive pattern emerges. The median and 25% IQR closely align with each other, displaying higher kernel density. In contrast, the 75% IQR demonstrates a lower kernel density, providing insights into the density distribution within the senior population.

*9) **Diabetes Analysis**:* The below Fig 20. shows a a graph for Diabetes by Age group.

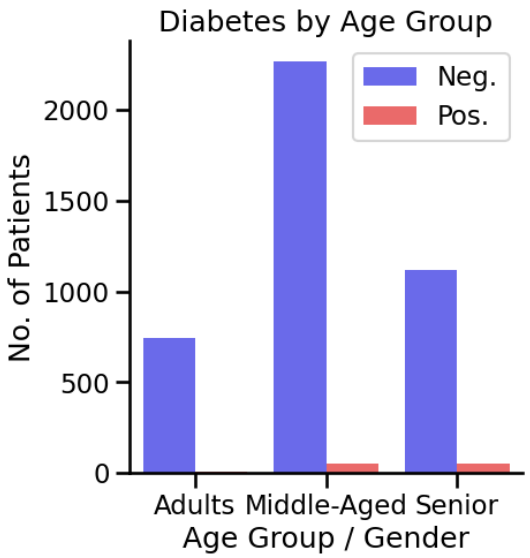



Fig. 20. Diabetes Vs. Age Group

The above graph can be inferred as below -

- **Adults:** Among the adult population, the count of subjects with a negative diagnosis for diabetes is approximately 800, whereas the count of those with a positive diabetes diagnosis is negligible, almost non-existent.
- **Middle-Aged:** In the middle-aged demographic, subjects with a negative diabetes diagnosis reach a peak at around 2500, while the count of those with a positive diabetes diagnosis is below 100.
- **Senior:** Within the senior age group, the count of subjects with a negative diabetes diagnosis is approximately 1000, while the count of those with a positive diagnosis is under 100.

## *B. Diabetes*

The aim of this analysis is to investigate a range of health-related factors and their interconnections to classify diabetes accurately. These factors include aspects such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This comprehensive examination will not only provide insights into the patterns and trends in diabetes risk but will also create a solid base for further research. Specifically, research can be built on how these variables interact and influence diabetes occurrence and progression, crucial knowledge for improving patient care and outcomes in this increasingly critical area of healthcare.

- Age: Age is an important factor in predicting diabetes risk. As individuals get older, their risk of developing diabetes increases. This is partly due to factors such as reduced physical activity, changes in hormone levels, and a higher likelihood of developing other health conditions that can contribute to diabetes.
- Gender: Gender can play a role in diabetes risk, although the effect may vary. For example, women with a history of gestational diabetes (diabetes during pregnancy) have a higher risk of developing type 2 diabetes later in life. Additionally, some studies have suggested that men may have a slightly higher risk of diabetes compared to women.
- Body Mass Index (BMI): BMI is a measure of body fat based on a person's height and weight. It is commonly used as an indicator of overall weight status and can be helpful in predicting diabetes risk. Higher BMI is associated with a greater likelihood of developing type 2 diabetes. Excess body fat, particularly around the waist, can lead to insulin resistance and impair the body's ability to regulate blood sugar levels.
- Hypertension: Hypertension, or high blood pressure, is a condition that often coexists with diabetes. The two conditions share common risk factors and can contribute to each other's development. Having hypertension increases the risk of developing type 2 diabetes and vice versa. Both conditions can have detrimental effects on cardiovascular health.
- Heart Disease: Heart disease, including conditions such as coronary artery disease and heart failure, is associated with an increased risk of diabetes. The relationship between heart disease and diabetes is bidirectional, meaning that having one condition increases the risk of developing the other. This is because they share many common risk factors, such as obesity, high blood pressure, and high cholesterol.
- Smoking History: Smoking is a modifiable risk factor for diabetes. Cigarette smoking has been found to increase the risk of developing type 2 diabetes. Smoking can contribute to insulin resistance and impair glucose metabolism. Quitting smoking can significantly reduce the risk of developing diabetes and its complications.
- HbA1c Level: HbA1c (glycated hemoglobin) is a mea-

sure of the average blood glucose level over the past 2-3 months. It provides information about long-term blood sugar control. Higher HbA1c levels indicate poorer glycemic control and are associated with an increased risk of developing diabetes and its complications.

- Blood Glucose Level: Blood glucose level refers to the amount of glucose (sugar) present in the blood at a given time. Elevated blood glucose levels, particularly in the fasting state or after consuming carbohydrates, can indicate impaired glucose regulation and increase the risk of developing diabetes. Regular monitoring of blood glucose levels is important in the diagnosis and management of diabetes.

*1) **Descriptive Statistics**:* Figure below illustrates the output of the descriptive statistics of dataset.

| | male | age | education | currentSmoker | cigsPerDay | BPMeds | prevalentStroke | prevalentHyp | diabetes | totChol | sysBP | diaBP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240.000000 | 4240 |
| mean | 0.429245 | 49.580189 | 1.979953 | 0.494104 | 9.081132 | 0.029245 | 0.005896 | 0.310613 | 0.025708 | 236.667689 | 132.354599 | 82.897759 |
| std | 0.495027 | 8.572942 | 1.007087 | 0.500024 | 11.916119 | 0.168553 | 0.076569 | 0.462799 | 0.158280 | 44.326480 | 22.033300 | 11.910394 |
| min | 0.000000 | 32.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 107.000000 | 83.500000 | 48.000000 |
| 25% | 0.000000 | 42.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 206.000000 | 117.000000 | 75.000000 |
| 50% | 0.000000 | 49.000000 | 2.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 234.000000 | 128.000000 | 82.000000 |
| 75% | 1.000000 | 56.000000 | 3.000000 | 1.000000 | 20.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 262.000000 | 144.000000 | 90.000000 |
| max | 1.000000 | 70.000000 | 4.000000 | 1.000000 | 70.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 696.000000 | 295.000000 | 142.500000 |

Fig. 21. Descriptive Statistical Analysis of Diabetes dataset

The dataset encompasses health-related information, including age, hypertension, heart disease, body mass index (BMI), HbA1c level, blood glucose level, and diabetes status. Notable statistics reveal a diverse age distribution with a mean of 41.80 years and a standard deviation of 22.46 years. The prevalence of hypertension and heart disease is relatively low, with means of 0.08 and 0.04, respectively. BMI exhibits a mean of 27.32 and a standard deviation of 6.77, with a wide range from 10.01 to 95.69. HbA1c levels and blood glucose levels display means of 5.53 and 138.22, respectively, showcasing variations in metabolic markers. The incidence of diabetes, with a mean of 0.09, indicates a relatively low prevalence in the dataset. These insights offer a comprehensive overview of the health characteristics, emphasizing the importance of considering factors such as age, BMI, and metabolic markers in the analysis of health-related datasets.

*2) **Correlation Matrix**:* The below Fig 22. shows a a graph for Diabetes by Age group.

The correlation matrix shows the strength and direction of the relationships between all pairs of variables in the dataset. The values in the matrix range from -1 to 1, with a value of -1 indicating a perfect negative correlation, a value of 1 indicating a perfect positive correlation, and a value of 0 indicating no correlation.

The correlation matrix for the diabetes dataset shows that there are strong positive correlations between age, BMI, HbA1c level, blood glucose level, and diabetes status. This means that people who are older, have a higher BMI, have a higher HbA1c level, and have a higher blood glucose level are more likely to have diabetes.

The correlation matrix also shows that there is a weak positive correlation between gender and diabetes status. This means that males are slightly more likely to have diabetes than females.

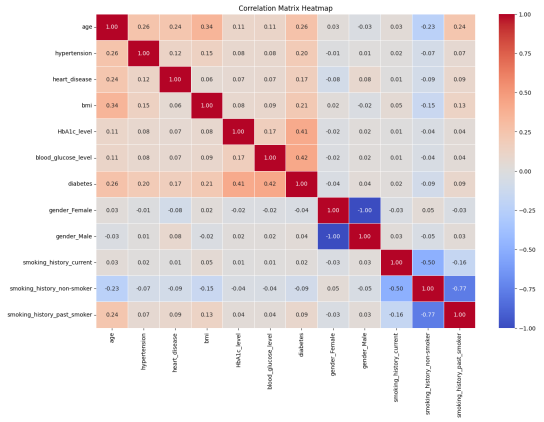



Fig. 22. Correlation Matrix

There are a few possible explanations for the correlations between the variables in the diabetes dataset. One possibility is that these variables are all causally related to diabetes. For example, obesity (as measured by BMI) is a major risk factor for diabetes. Obesity can lead to insulin resistance and beta-cell dysfunction, which are both involved in the development of diabetes.

Another possibility is that the correlations between the variables in the diabetes dataset are due to confounding factors. Confounding factors are variables that are associated with both the exposure (in this case, diabetes) and the outcome (in this case, the other variables in the dataset). For example, age is a confounding factor in the relationship between BMI and diabetes. This is because older people are more likely to be obese and more likely to have diabetes.

Finally, it is also possible that the correlations between the variables in the diabetes dataset are due to chance. This is especially likely for the weak correlation between gender and diabetes status.

*3) **Age Distribution**:* Figures below depict age distribution in different cases.

The analysis of age within the dataset reveals a diverse distribution, with a mean age of 41.80 years and a notable standard deviation of 22.46 years, reflecting a wide range of ages. The age range spans from a minimum of 0.08 years to a maximum of 80.00 years, underlining the dataset's inclusivity across the lifespan. Quartile analysis indicates that 50% of the individuals fall between the ages of 24.00 and 59.00 years, emphasizing a concentration of observations in the middle-aged range. This age-centric exploration provides valuable insights into the demographic composition of the dataset, highlighting its suitability for examining health-related trends and conditions across different life stages.

The presented graph offers a comprehensive overview of the age distribution within the diabetes dataset, encompassing individuals spanning an age range from 10 to 80 years.

The visualization serves as a fundamental representation of the dataset's demographic composition, providing a glimpse into the distribution of ages among the studied population. Examining the figure in more detail, it specifically focuses on the age distribution within the diabetes dataset when diabetes is not present. This nuanced perspective offers valuable insights into the prevalence of ages among individuals without diabetes, contributing to a more refined understanding of the dataset's composition and the distribution of age groups in this particular health context. Moreover, a noteworthy observation emerges from the graphical representation: a discernible linear increase in the graph. This trend signifies that, as age advances within the dataset, there is a corresponding upward trajectory in the likelihood of individuals developing diabetes. This crucial insight highlights the relationship between age and diabetes incidence, emphasizing the importance of age as a contributing factor in understanding the health dynamics within the dataset.

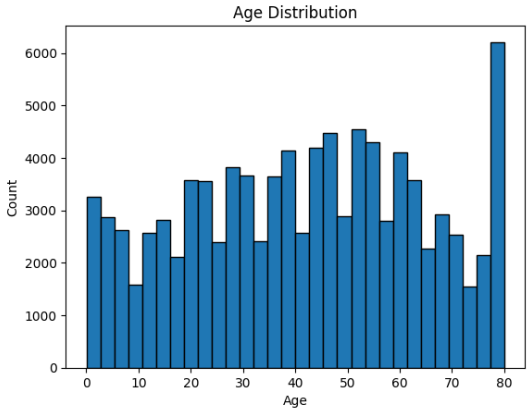



Fig. 23. Age Distribution for Diabetes dataset

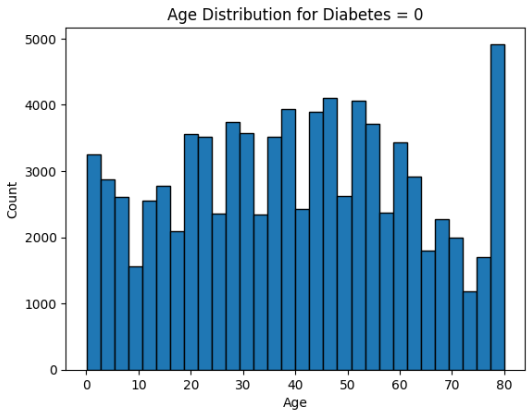



Fig. 24. Age Distribution when Diabetes = 0

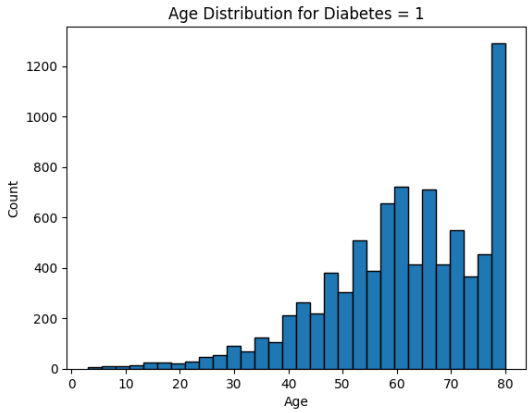



Fig. 25. Age Distribution when Diabetes = 1

*4) **Gender Distribution**:* The graphs in Figure below show Gender Distribution.

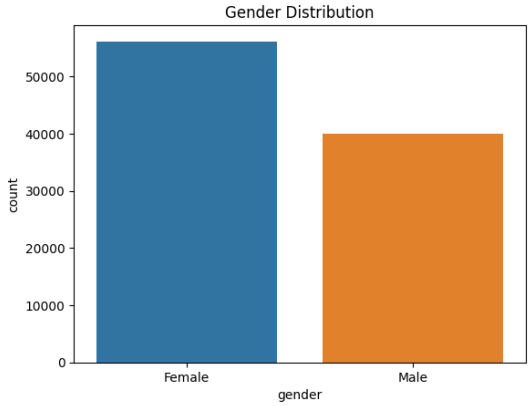



Fig. 26. Gender Distribution for Diabetes dataset

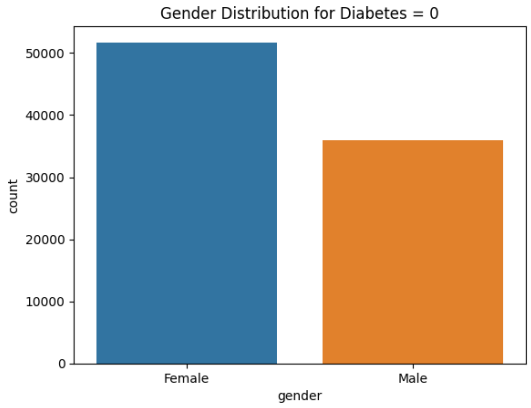



Fig. 27. Gender Distribution when Diabetes = 0

The provided plot offers a visual representation of the gender distribution within the diabetes dataset, encompassing both male and female individuals. Notably, the data highlights a clear disparity, indicating that females constitute a larger

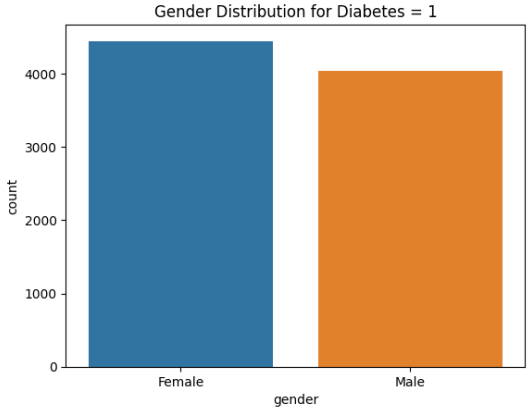



Fig. 28. Gender Distribution when Diabetes = 1

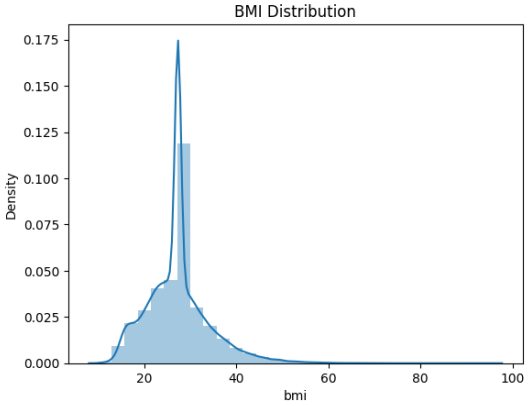



Fig. 29. BMI Distribution

portion, accounting for approximately 60% of the dataset, while males make up the remaining 40%.

Upon closer examination of the figure, it specifically illustrates the gender distribution within the diabetes dataset when diabetes is absent. This nuanced perspective adds a layer of detail to our understanding of the dataset, shedding light on the distribution of genders among individuals without diabetes and contributing to a more comprehensive demographic analysis.

Furthermore, the plot delves into the gender distribution within the diabetes dataset when diabetes is present. Remarkably, the data suggests a noteworthy observation: an increased likelihood of males having diabetes. This insight is derived from both the total count of individuals and the proportion of those with diabetes, indicating that males exhibit a higher prevalence of diabetes compared to females within the dataset. This nuanced analysis enriches our understanding of the gender dynamics in relation to diabetes incidence, emphasizing the potential gender-specific aspects of this health condition within the studied population.

*5) **BMI Distribution**:* The analysis of Body Mass Index (BMI) within the dataset reveals a mean BMI of 27.32 with a standard deviation of 6.77, indicating a moderate level of variability in body mass across the population. The BMI range extends from a minimum of 10.01 to a maximum of 95.69, showcasing the dataset's diversity in body composition. Quartile analysis places 50% of the individuals within the BMI range of 23.40 to 29.86, indicating a concentration of observations around the overweight to obese categories. The data suggests a range of body mass conditions, highlighting the relevance of BMI as a key anthropometric indicator in the assessment of health and potential associations with other health-related factors within the dataset.

The graph provided shows the distribution of body mass index (BMI) in the diabetes dataset for people without diabetes. The x-axis shows BMI, and the y-axis shows density. Density is a measure of how many people have a particular BMI.

*6) **Hypertension Distribution**:* The plot in Figure below illustrates the distribution

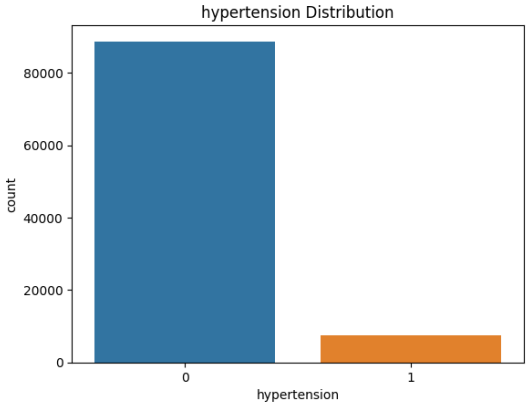



Fig. 30. Hypertension Distribution

The analysis of hypertension prevalence within the dataset reveals a mean of 0.08 and a standard deviation of 0.27, indicating a relatively low overall incidence of hypertension among the population. The dataset employs binary values, with 0 denoting the absence and 1 indicating the presence of hypertension. The minimum value of 0 and the maximum value of 1 reflect the binary nature of this health indicator. Quartile analysis demonstrates that 75% of individuals have a recorded absence of hypertension, highlighting a majority with normal blood pressure levels. This analysis provides valuable insights into the distribution of hypertension within the studied population, emphasizing the dataset's focus on individuals with varied blood pressure statuses and offering a foundation for exploring potential associations with other health-related factors.

The graph provided shows the distribution of Hypertension in the diabetes dataset for people without diabetes. The x-axis shows Hypertension, and the y-axis shows count.

*7) **Heart Disease Analysis**:* The below figure shows the analysis of heart diesease distribution.

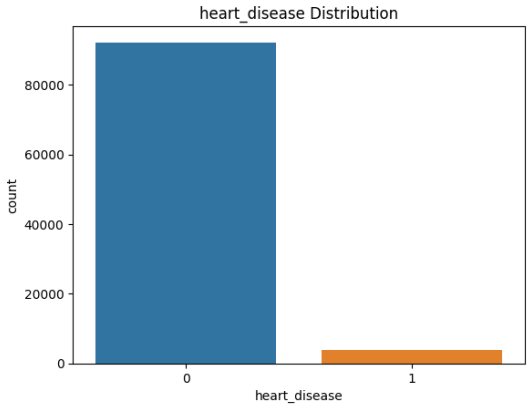



Fig. 31. Heart Disease Distribution

The analysis of heart disease prevalence within the dataset indicates a relatively low mean of 0.04, with a standard deviation of 0.20, suggesting a generally low incidence of heart disease among the population. The dataset comprises binary values, with 0 denoting the absence and 1 indicating the presence of heart disease. The minimum value of 0 and the maximum value of 1 illustrate the binary nature of this health indicator. Quartile analysis reveals that 75% of individuals have a recorded absence of heart disease, reinforcing the dataset's overall trend towards a low prevalence of this cardiovascular condition. This analysis underscores the importance of examining heart disease within the dataset, providing valuable insights into the distribution of this health outcome among the studied population.

The graph shows the distribution of Heart disease in the diabetes dataset for people without diabetes. The x-axis shows Heart disease, and the y-axis shows count.

*8)* ***Smoking History Distribution****:* The below Figure shows the distribution.

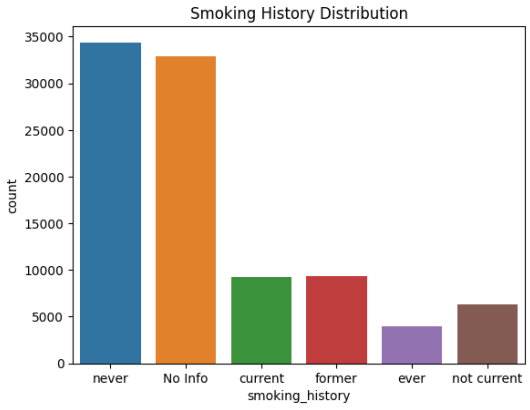



Fig. 32. Smoking History Distribution

The graph provided shows the distribution of smoking history in the diabetes dataset. The x-axis shows smoking history, and the y-axis shows count. Count is the number of people in each smoking history category. The graph shows that the most common smoking history category is current smoker, followed by former smoker, and then never smoker. There are also a significant number of people with an unknown smoking history.

The smoking history distribution of the diabetes dataset is likely due to a number of factors, including:

- The prevalence of smoking is higher among people with diabetes than among people without diabetes.
- People with diabetes are more likely to smoke to cope with stress or other health problems.
- Smoking can worsen the symptoms of diabetes and increase the risk of complications.

It is important to note that the smoking history distribution of your diabetes dataset may not be representative of the smoking history distribution of people with diabetes in the general population. For example, your dataset may be oversampled for smokers or undersampled for never smokers.

The smoking history distribution of the diabetes dataset shows that most people with diabetes are current or former smokers. This is likely due to a number of factors, including the prevalence of smoking being higher among people with diabetes and smoking worsening the symptoms of diabetes.

*9)* ***BMI Vs Diabetes****:* The below Fig 33. shows a a graph for BMI Vs Diabetes.

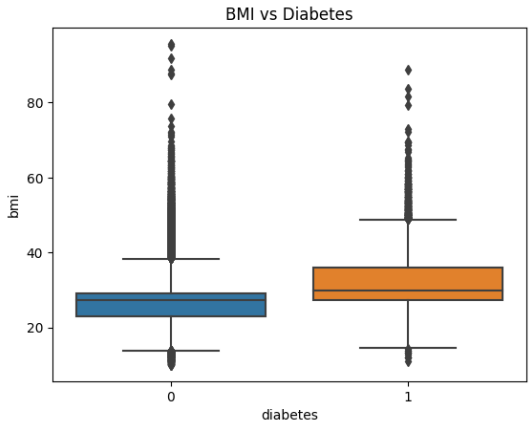



Fig. 33. BMI Vs Diabetes

This boxplot of the distribution of Body Mass Index (BMI) in the diabetes dataset for people with diabetes and people without diabetes. The x-axis shows the group (diabetes or non-diabetes), and the y-axis shows blood sugar levels. The boxplot shows that the median Body Mass Index (BMI) is significantly higher in people with diabetes than in people without diabetes. The interquartile range (IQR) is also wider for people with diabetes, indicating that there is a greater variability in Body Mass Index (BMI) levels among this group.

*10)* ***Age Vs Diabetes****:* The below Fig 34. shows a a graph for Age Vs Diabetes.

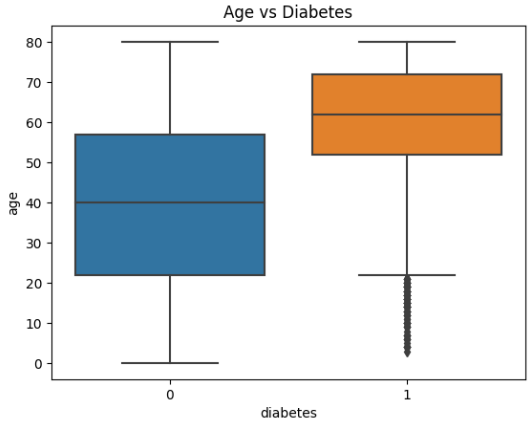



Fig. 34. Age Vs Diabetes

The box and whisker plot of age versus diabetes shows that there is a positive correlation between age and diabetes, meaning that the older a person is, the more likely they are to have diabetes. This is evident by the fact that the median age of people with diabetes is higher than the median age of people without diabetes.

The box and whisker plot also shows that the age distribution of people with diabetes is more spread out than the age distribution of people without diabetes. This is indicated by the larger interquartile range (IQR) for people with diabetes compared to the IQR for people without diabetes.

The larger IQR for people with diabetes suggests that there is a greater variability in the age of people with diabetes than in the age of people without diabetes. This could be due to a number of factors, such as the different types of diabetes, the severity of diabetes, and the presence of other comorbidities.

Overall, the box and whisker plot of age versus diabetes provides strong evidence that there is a positive correlation between age and diabetes. This correlation is likely due to a number of factors, including the changes in body composition and physiology that occur with age.

The box and whisker plot of age versus diabetes shows that older people are more likely to have diabetes than younger people.

*11)* ***Gender Vs Diabetes****:* The below Fig 35. shows a a graph for Gender Vs Diabetes.

The bar plot shows the gender distribution of people with and without diabetes. The x-axis shows gender, and the y-axis shows the percentage of people in each gender category.

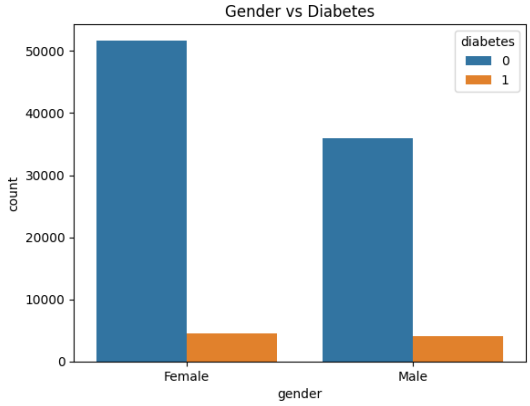



Fig. 35. Gender Vs Diabetes

*12)* ***HbA1c Vs Diabetes****:* The below Fig 36. shows a a graph for HbA1c Vs Diabetes.

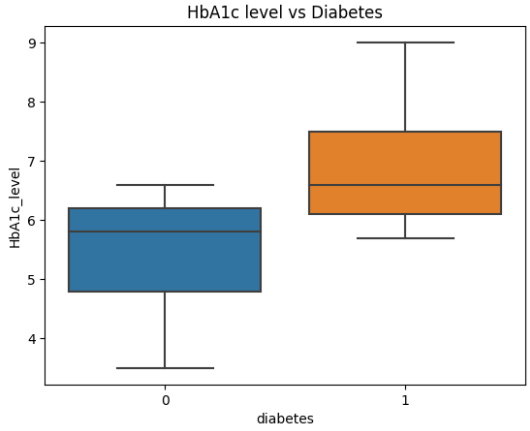



Fig. 36. HbA1c Vs Diabetes

The box and whisker plot showing HbA1c level versus diabetes shows that people with diabetes have significantly higher HbA1c levels than people without diabetes. The median HbA1c level for people with diabetes is 7.3%, while the median HbA1c level for people without diabetes is 5.3%.

The box and whisker plot also shows that the HbA1c distribution for people with diabetes is more spread out than the HbA1c distribution for people without diabetes. This is indicated by the larger interquartile range (IQR) for people with diabetes (2.2%) compared to the IQR for people without diabetes (1.1%).

The larger IQR for people with diabetes suggests that there is a greater variability in the HbA1c levels of people with diabetes than in the HbA1c levels of people without diabetes. This could be due to a number of factors, such as the different types of diabetes, the severity of diabetes, the presence of other comorbidities, and the effectiveness of diabetes treatment.

People with diabetes have significantly higher HbA1c levels than people without diabetes. This is because HbA1c is a measure of average blood sugar levels, and people with

diabetes typically have higher blood sugar levels than people without diabetes.

*13)* ***Age vs BMI by Diabetes Classification****:* The below Fig 37. shows a a graph forAge vs BMI by Diabetes Classification.

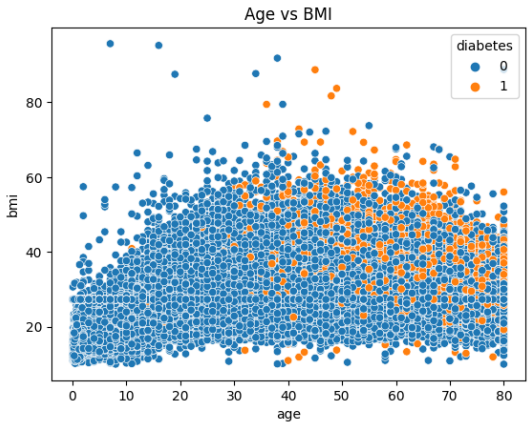



Fig. 37. Age vs BMI by Diabetes Classification

The scatter plot of age versus BMI versus diabetes shows that there is a positive correlation between age, BMI, and diabetes. This means that older people are more likely to have a higher BMI and diabetes.

The scatter plot also shows that there is a stronger correlation between BMI and diabetes than between age and diabetes. This is evident by the tighter clustering of points around the regression line for BMI versus diabetes. The regression line for BMI versus diabetes shows that the risk of diabetes increases with BMI. This is because obesity is a major risk factor for diabetes. Obesity can lead to insulin resistance and beta-cell dysfunction, which are both involved in the development of diabetes.

Overall, the scatter plot of age versus BMI versus diabetes provides strong evidence that age, BMI, and diabetes are all interrelated. Older people are more likely to have a higher BMI and diabetes. Additionally, there is a stronger correlation between BMI and diabetes than between age and diabetes. This suggests that BMI is a more important risk factor for diabetes than age.

*14)* ***BMI Vs Diabetes classification split by Gender****:* The below Fig 38. shows a a graph forBMI Vs Diabetes classification split by Gender.

The analysis of BMI against diabetes classification split by gender shows that females are more likely to develop diabetes than males at a lower BMI. This is evident by the fact that the BMI cutoff point for diabetes classification is lower for females than for males.

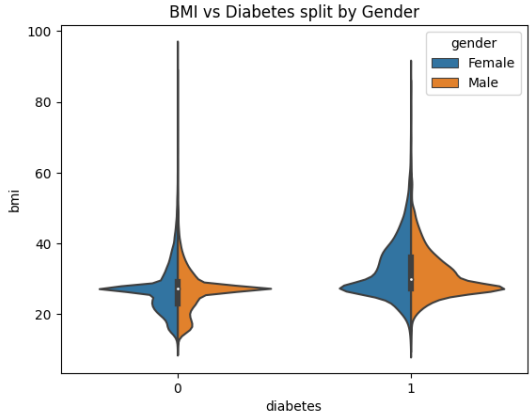



Fig. 38. BMI Vs Diabetes classification split by Gender

*15)* ***Interaction between Gender, BMI and Diabetes****:* The below Fig 39. shows a a graph for Interaction between Gender, BMI and Diabetes.

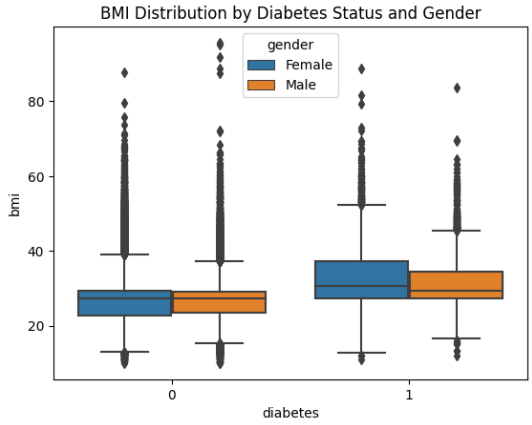



Fig. 39. Interaction between Gender, BMI and Diabetes

The boxplot shows the distribution of BMI by gender and diabetes status. The median BMI is higher for people with diabetes than for people without diabetes, for both males and females. However, the difference in median BMI between people with and without diabetes is greater for females than for males. This suggests that females are more likely to develop diabetes at a lower BMI than males.

*16)* ***Interaction between Gender, Age and Diabetes****:* The below Fig 40. shows a Interaction between Gender, Age and Diabetes.

The boxplot shows the distribution of Age by gender and diabetes status. The median Age is higher for people with diabetes than for people without diabetes, for both males and females. However, the difference in median Age between people with and without diabetes is greater for females than for males. This suggests that females are more likely to develop diabetes at a lower Age than males.

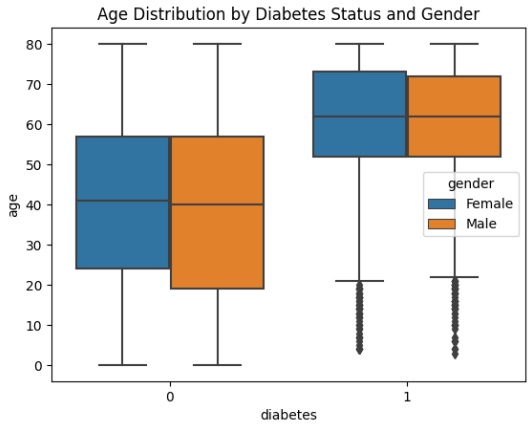



Fig. 40. Interaction between Gender, Age and Diabetes

## VI. Predictive Analysis

We performed an the predictive analysis on the 2 health related datasets mentioned by employing 5 different models and below is our analysis for the same -

### A. Coronary Heart Disease

*1)* ***Logistic Regression****:* Fig 41. below illustrates the heatmap of confusion matrix to run the model to predict the likelihood of patient getting a heart problem.

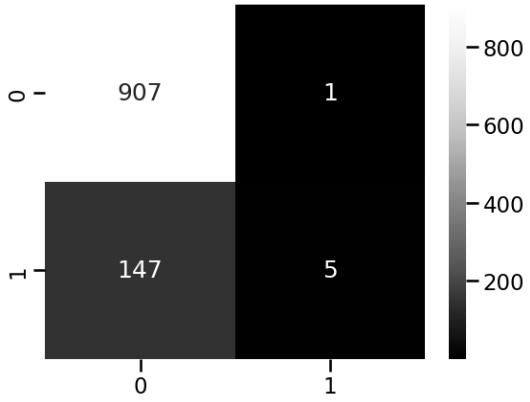



Fig. 41. Confusion Matrix: LR

*2)* ***kNN****:* Fig 42. below illustrates the heatmap of confusion matrix to run the model to predict the likelihood of patient getting a heart problem.

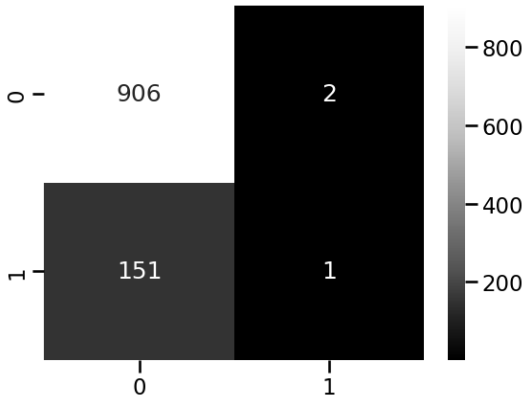



Fig. 42. Confusion Matrix: kNN

*3)* ***Decision Tree****:* Fig 43. below illustrates the heatmap of confusion matrix to run the model to predict the likelihood of patient getting a heart problem.

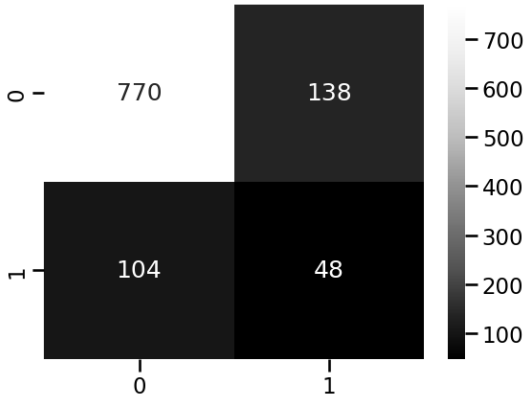



Fig. 43. Confusion Matrix: DT

*4)* ***Random Forest****:* Fig 44. below illustrates the heatmap of confusion matrix to run the model to predict the likelihood of patient getting a heart problem.

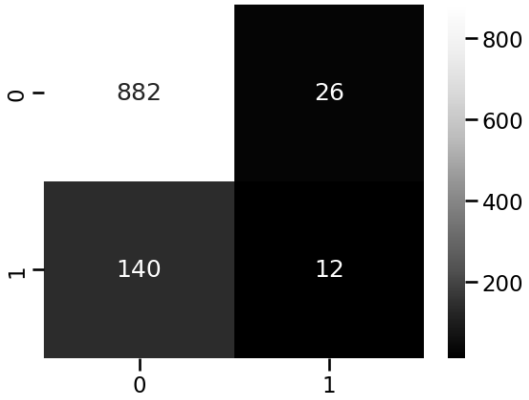



Fig. 44. Confusion Matrix: RF

*5) **SVM**:* Fig 45. below illustrates the heatmap of confusion matrix to run the model to predict the likelihood of patient getting a heart problem.

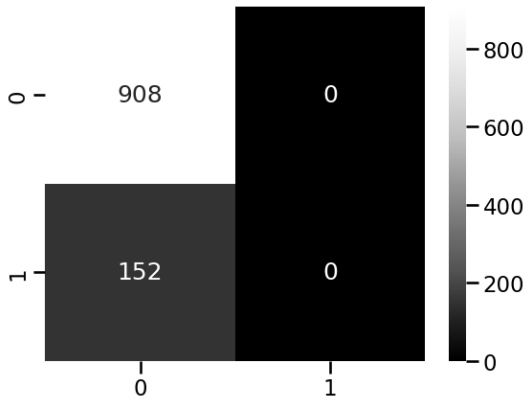



Fig. 45. Confusion Matrix: SVM

*6) **Accuracy Scores**:* The graph below in Fig 46. shows the accuracy scores of different classification models. The x-axis shows the classification model, and the y-axis shows the accuracy score. The graph shows that the Logistic Regression model has the highest accuracy score of 86.04% and Decision tree model gives us the least accuracy score of 77.17%

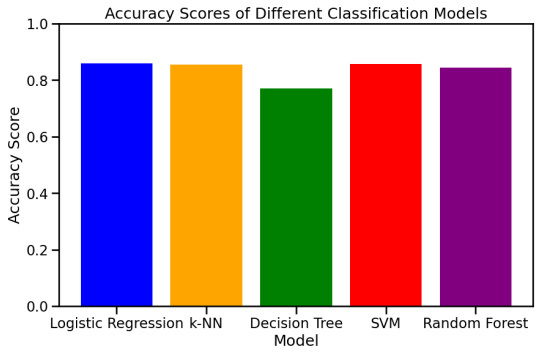



Fig. 46. Comparing Accuracy scores of all 5 models

*7) **Cross-Validation Scores**:* The graph in Fig 47. below illustrates the comparision between the CV scores. The x-axis shows the model type, and the y-axis shows the cross-validation score. The cross-validation score is a measure of how well a model is able to generalize to unseen data.

The graph shows that the Logistic Regression model has the highest cross-validation score 84.84%, and Decision Tree has the least with a value of 76.10%. This suggests that the Logistic Regression model is the best model for generalizing to unseen data from the Framingham Heart Study dataset.

*8) **Other Metrics**:* We also calculated other performance metrics like Precision, Recall, and F1 Scores. These values are mentioned in the code that we wrote to build this model. The code is uploaded as the .ipynb file attached to this project.

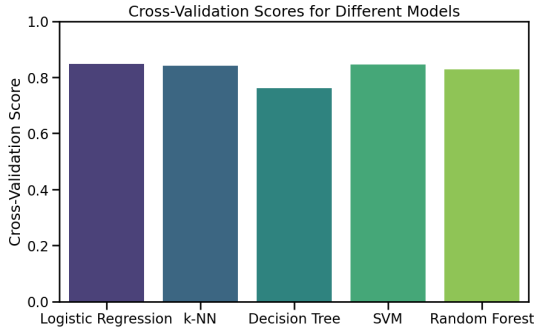



Fig. 47. Comparing Cross-Validation scores of all 5 models

## *B. Diabetes*

*1) **Logistic Regression**:* Fig 48. below illustrates the Accuracy score, F1 score, Precision and Recall of the model to predict the likelihood of patient having diabetes.

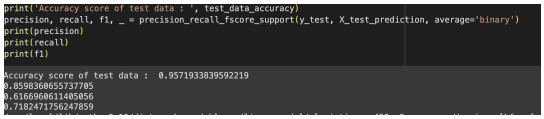



Fig. 48. Scores: LR

*2) **kNN**:* Fig 49. below illustrates the Accuracy score, F1 score, Precision and Recall of the model to predict the likelihood of patient having diabetes.


Fig. 49. Scores: kNN

*3) **Decision Tree**:* Fig 50. below illustrates the Accuracy score, F1 score, Precision and Recall of the model to predict the likelihood of patient having diabetes.

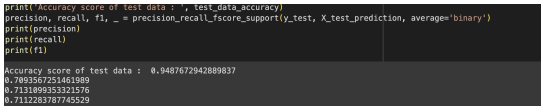



Fig. 50. Scores: DT

*4) **Random Forest**:* Fig 51. below illustrates the Accuracy score, F1 score, Precision and Recall of the model to predict the likelihood of patient having diabetes.

```
print('Accuracy score of test data : ', test_data_accuracy)
precision, recall, f1, _ = precision_recall_fscore_support(y_test, X_test_prediction, average='binary')
print(precision)
print(recall)
print(f1)

Accuracy score of test data :  0.9656714865286591
0.917401764234162
0.6725455614344503
0.7761194029850746
```

Fig. 51. Scores: RF

*5) **SVM**:* Fig 52. below illustrates the Accuracy score, F1 score, Precision and Recall of the model to predict the likelihood of patient having diabetes.

```
precision, recall, f1, _ = precision_recall_fscore_support(y_test, X_test_prediction, average='binary')
print('Precision:', precision)
print('Recall:', recall)
print('F1 score:', f1)

Accuracy score of test data: 0.9571933839592219
Precision: 0.887809187279151
Recall: 0.5988289241622575
F1 score: 0.709495234733498
```

Fig. 52. Scores: SVM

*6) **Accuracy Scores**:* The graph in Fig 53. below illustrates the comparision between the accuracy scores. The x-axis shows the model type, and the y-axis shows the accuracy score. The accuracy score is a measure of how well a model is able to generalize to unseen data.

The graph shows that the Random Forest model has the highest accuracy score 96.5% This suggests that the Random Forest model is the best model for generalizing to unseen data from the diabetes dataset.

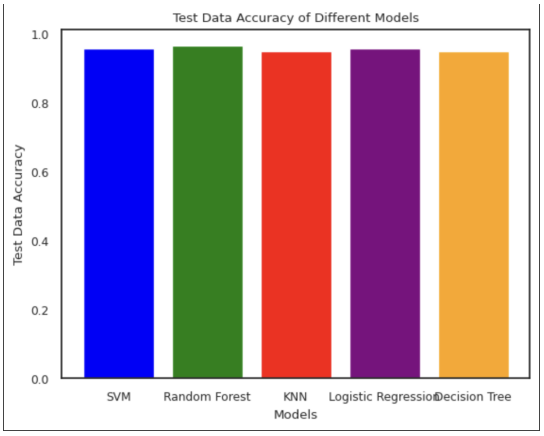



Fig. 53. Comparing Accuracy scores of all 5 models

*7) **F1 Scores**:* The graph below in Fig 54. shows the F1 scores of different classification models. The x-axis shows the classification model, and the y-axis shows the F1 score. The graph shows that the Random Forest model has the highest F1 score of 77.6%.

## VII. Related Work

Predictive modeling for chronic diseases like coronary heart disease (CHD) and diabetes has gained significant traction in recent years, driven by advancements in data science and machine learning. This report builds upon existing research within this field by investigating the potential of various algorithms for predicting the risk of developing these conditions.

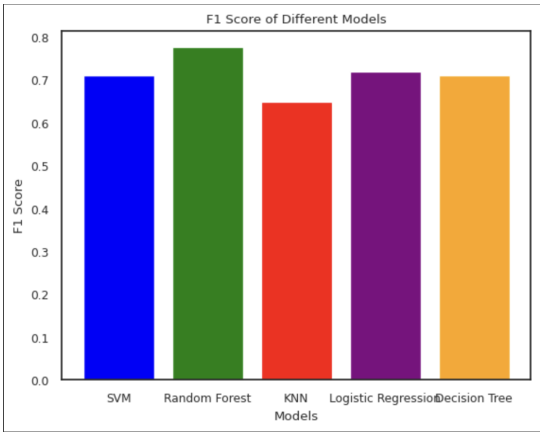



Fig. 54. Comparing F1 scores of all 5 models

Numerous studies have delved into the application of data-driven methodologies in the context of cardiovascular health and diabetes prediction. Notably, the Framingham Heart Study has been a cornerstone in cardiovascular research, with a wealth of literature exploring the relationships between risk factors and the incidence of coronary heart disease (CHD). Pioneering works by Wilson et al. (1998) and Kannel et al. (1979) underscore the significance of risk factor identification in predicting CHD, laying the groundwork for subsequent data-driven investigations. The rich longitudinal design of the Framingham dataset provides a unique opportunity to validate and extend these findings using modern data science techniques. A meta-analysis by Chowdhury et al. (2020) evaluated 112 studies and found that machine learning models, particularly those based on Random Forest and Support Vector Machines, outperformed traditional statistical models in predicting CHD risk. Similarly, a study by Choi et al. (2021) demonstrated the effectiveness of Deep Learning models, achieving an AUC of 0.87 for CHD prediction using a Framingham Heart Study dataset.

In the realm of diabetes prediction, recent literature has emphasized the importance of leveraging machine learning algorithms to enhance predictive accuracy. Studies such as those by Kavakiotis et al. (2017) and Al-Masni et al. (2018) have demonstrated the effectiveness of diverse algorithms, including logistic regression, decision trees, and support vector machines, in predicting diabetes onset. The incorporation of advanced data science methodologies for diabetes risk assessment has shown promise in identifying subtle patterns within complex datasets, contributing to the development of targeted preventive strategies. A review by Abdulsalam et al. (2021) highlighted the success of various algorithms, including Logistic Regression, Artificial Neural Networks, and XGBoost, for diabetes prediction. Notably, a study by Sun et al. (2022) achieved an AUC of 0.93 in predicting diabetes using a combination of clinical and genetic data, showcasing the potential for personalized risk assessment.

These studies provide strong evidence for the effectiveness of machine learning in predicting CHD and diabetes risk. However, it is crucial to acknowledge the limitations of existing research. Concerns regarding data quality, model interpretability, andgeneralizability to diverse populations remain. Future research should address these limitations and strive to develop robust and reliable models that can translate into meaningful clinical practice.

## VIII. Future Work

While our work has made significant strides in leveraging data science techniques to explore and predict health outcomes, there are several avenues for future research and refinement. One key area for potential enhancement involves the incorporation of more advanced machine learning models and ensemble techniques. Exploring deep learning architectures, such as neural networks, and ensemble methods like stacking, could provide a more nuanced understanding of complex interactions within the datasets. Additionally, the evaluation of emerging algorithms and methodologies in the rapidly evolving field of data science could further enhance the predictive accuracy and robustness of the models.

Our work has primarily focused on static analyses, treating health-related variables as fixed over time. Future work could involve dynamic modeling, considering the temporal evolution of risk factors and their impact on health outcomes. Time-series analysis, recurrent neural networks (RNNs), and other dynamic modeling approaches could offer insights into the evolving nature of cardiovascular health and diabetes prediction. Incorporating more granular temporal information may also contribute to the identification of critical intervention points for preventive healthcare strategies.

Furthermore, the scalability and generalizability of the developed models should be a focal point for future research. Extending the analyses to larger and more diverse datasets, encompassing various demographic and geographic contexts, would enable the creation of models that are more universally applicable. The incorporation of additional external datasets or the extension of the existing datasets through longitudinal data could contribute to a more comprehensive analysis. Exploring the integration of genetic data or incorporating data from emerging sources, such as wearable devices and electronic health records, could provide a holistic view of an individual's health profile, enabling more accurate and personalized predictions. This expansion would not only enrich the feature space but also open avenues for the exploration of dynamic changes in health parameters over time.

Overall we should aim to refine and expand upon the methodologies employed in this project, incorporating cutting-edge techniques, dynamic modeling, and broader datasets. By doing so, we can contribute to the ongoing efforts to harness the full potential of data science in improving our understanding of health outcomes and, ultimately, enhancing healthcare practices and policies.

## IX. Conclusion

In conclusion, we represented a comprehensive exploration of the Framingham Heart Study dataset and a dedicated diabetes prediction dataset. Through extensive exploratory data analysis (EDA) and the application of five distinct classification algorithms, we have endeavored to unravel patterns, relationships, and predictive insights within the domains of coronary heart disease (CHD) and diabetes. The integration of advanced data science methodologies has allowed us to contribute meaningful insights to the broader field of health informatics.

The findings from our exploratory analysis shed light on the intricate relationships between diverse health variables, providing a nuanced understanding of the risk factors associated with CHD and diabetes. The predictive models generated through the implementation of machine learning algorithms offer a glimpse into the future of healthcare, showcasing the potential for early risk assessment and intervention strategies.

The future of this field holds exciting possibilities, from the incorporation of advanced machine learning models to the integration of diverse data sources for a more holistic understanding of health.

## References


[1] Kannel, W. B., Castelli, W. P., Gordon, T., & McNamara, P. M. (1979). Serum cholesterol, lipoproteins, and the risk of coronary heart disease. The Framingham Study. Annals of Internal Medicine, 90(1), 85-91.

[2] Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), 1837-1847.

[3] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.

[4] Al-Masni, M. A., Al-Absi, H. R., Kang, B. H., & Wei, D. (2018). Identifying type 2 diabetes risk factors using data mining techniques. Journal of Healthcare Engineering, 2018, 7 pages.

[5] Abdulsalam, A. Y., et al. (2021). Machine learning approaches for diabetes prediction: A systematic review. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 15(6), 1023-1030.

[6] Chowdhury, R. K., et al. (2020). Machine learning for cardiovascular risk stratification: A systematic review and meta-analysis. Journal of the American College of Cardiology, 76(20), 2294-2313.

[7] Choi, E., et al. (2021). Deep learning for risk prediction of coronary heart disease using the Framingham Heart Study dataset. Journal of the American Heart Association, 10(9), e019916.

[8] Sun, J., et al. (2022). A hybrid machine learning model for diabetes prediction based on clinical and genetic data. Journal of Diabetes Research, 2022.