**CARDIOVASCULAR DISEASE PREDICTION**

**By**

**ABHISHEK TYAGI**

**HARSHWARDHAN SINGH**

**POST-GRADUATION DIPLOMA IN BIG DATA ANALYTICS**

PRN: - 230320525003

PRN: - 230320525015



**CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING (NOIDA)**

**CERTIFICATE**

I hereby certify that the project work titled “**Cardiovascular Disease Prediction Using Machine Learning**” has been satisfactorily completed by Abhishek Tyagi, Harshwardhan Singh Thakur, Khushboo Yadav, and Prankur, under the guidance of **Mrs. Priti Bhardwaj**, as part of the PG-DBDA program at CDAC Noida.

To the best of my knowledge, this report has not been submitted for any other examination and does not form a part of any other course undertaken by the candidate. I have no doubts about their excellent research potential.

I wish them all the best for their future endeavours.

Project Guide

Priti Bhardwaj

**ACKNOWLEDGEMENT**

We express our sincere gratitude to all those who contributed to the successful completion of our project, “**Cardiovascular Disease Prediction Using Machine Learning**”, under the guidance of Mrs. Priti Bhardwaj. we are immensely grateful to her for inspiring and assisting us in the project, personally reviewing our work, and providing encouragement throughout. Additionally, we extend our heartfelt thanks to all the faculty members of the PG-DBDA department for their unwavering guidance and support throughout the project, both in the challenging and smooth phases. We also appreciate the assistance provided by CDAC, Noida.

We take this moment to thank our peer for their support and appreciation, which.

we value the most.

**ABSTRACT**

In modern times, health issues are on the rise due to various factors such as lifestyle and genetics. Heart disease has become more prevalent and poses a serious threat to people's well-being. The normal values for blood pressure, cholesterol, and pulse rate differ for everyone. Medically verified standards indicate that normal blood pressure is 120/90, cholesterol should be between 100-129 mg/dL, pulse rate should be 72, fasting blood sugar level is 100 mg/dL, heart rate should be between 60-100 bpm, ECG results should be normal, and the width of major vessels should range from 25 mm (1 inch) in the aorta to only 8 mm in the capillaries. This report examines various classification techniques that are used to predict the risk level of each person based on their age, gender, blood pressure, cholesterol, and pulse rate.

The “Disease Prediction” system employs predictive modelling to determine a user’s potential illness based on the symptoms and provide input. the system evaluates the input symptoms and provides the probability of the disease. six techniques- Naïve Bayes, KNN, Decision tree, Logistic regression, Random Forest, and support vector Machine Algorithm are used for the prediction, these techniques compute the probability of the disease,

**INDEX**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Sr. No.** | **Content** | **Page No.** |
|  |  | **1.** | **INTRODUCTION** | **6-7** |
|  | **2.** | **DATA COLLECTION AND DATA VISULIZATION** | **8-11** |
|  | **3.** | **DATA PREPROCESSING AND TRAIN TEST SPLIT** | **12-16** |
|  | **4.** | **MODELING** | **17-28** |
|  | **5.** | **DATA PREDICTION** | **29** |
|  | **6.** | **USER INTERACTIVE WEB APP** | **30-31** |
|  | **7.** | **CONCLUSION AND FUTURE WORK** | **32** |
|  | **8.** | **REFERENCES** | **33** |

**1) INTRODUCTION**

Numerous daily factors have the potential to impact our cardiac health. Given the discovery of new heart conditions, it is critical to prioritize the care of this vital organ responsible for blood circulation throughout our bodies. Both personal and occupational habits, as well as inherited genetic predispositions, can influence our heart's well-being. Unfortunately, millions worldwide fall victim to cardiovascular ailments annually. Cardiovascular disease is an umbrella term encompassing various conditions that specifically impact the heart and arteries. Even those in their 20s and 30s may be susceptible to cardiac issues due to unhealthy dietary habits, insufficient sleep, stress, depression, obesity, hypertension, high blood pressure, high blood cholesterol, or smoking. Maintaining a healthy heart requires proactive measures such as regular physical activity, a balanced diet, and avoiding harmful habits.

Heart disease symptoms can vary depending on the type of discomfort experienced by an individual. Some symptoms may not be easily recognizable, but common symptoms include chest pain, breathlessness, and heart palpitations. Angina, also known as angina pectoris, is a common type of chest pain in many forms of heart disease and is caused by a lack of oxygen in a part of the heart. Physical exertion or stressful events can trigger angina, which usually lasts less than ten minutes. Sometimes, heart attack symptoms may resemble indigestion, including heartburn, stomach-ache, and a heavy feeling in the chest. Other heart attack symptoms include pain that travels through the body, such as from the chest to the arms, neck, back, abdomen, or jaw, light-headedness and dizziness, profuse sweating, nausea, and vomiting.

Data Mining is a powerful technique for extracting crucial decision-making information from records for future analysis or prediction. It can unearth hidden patterns that are otherwise unidentifiable. Medical data mining integrates classification techniques and computerized training on datasets to discover patterns in medical data sets. This aids in predicting a patient's future state and analysing their medical history, which is essential for clinical analysis.

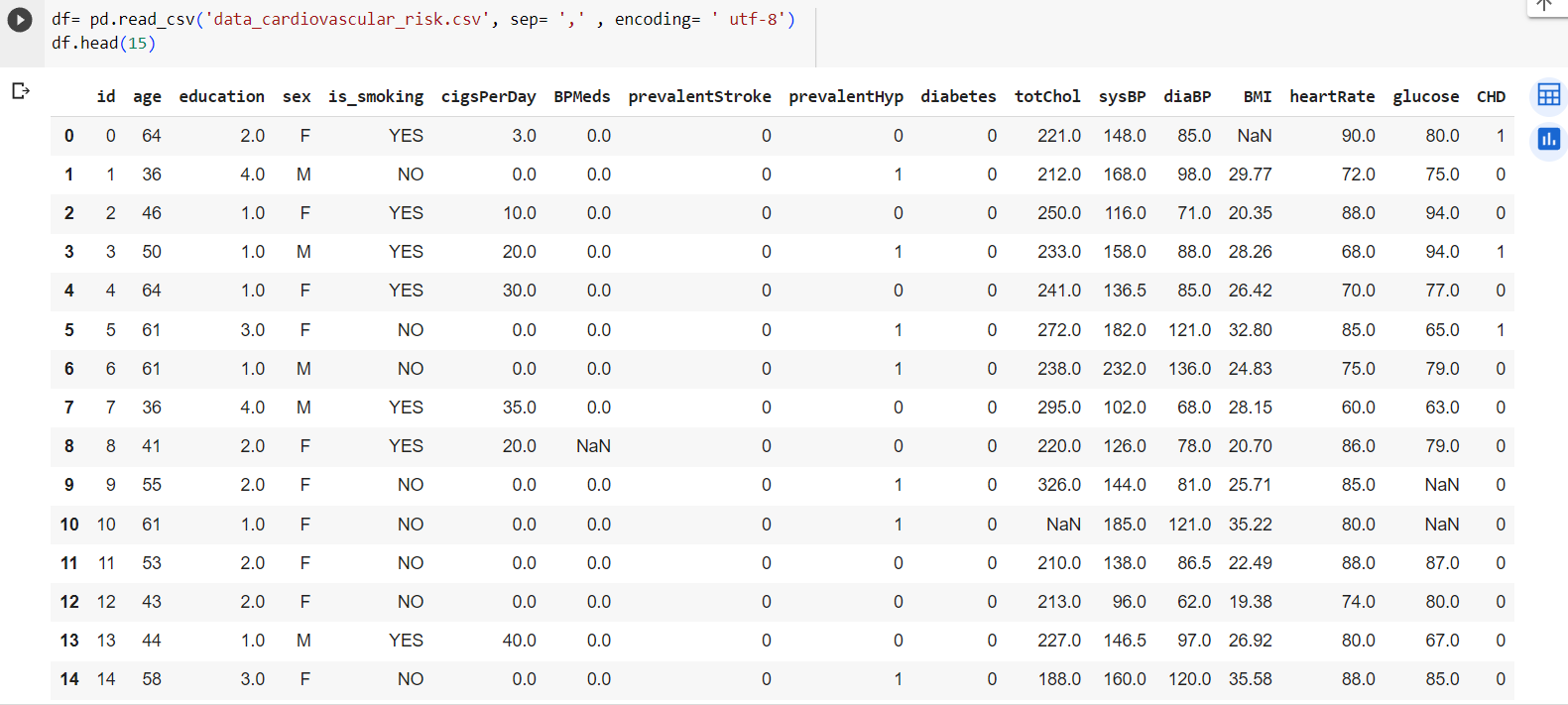
The objective of our project is to develop a robust model for predicting the occurrence of heart disease. To achieve this, a comparative analysis was conducted using six commonly used machine learning algorithms, namely Naïve Bayes, Decision Tree, KNN, Logistic Regression, Support Vector Machine, and Random Forest, XgBoost. However, since the accuracy of heart disease prediction is of utmost importance, all the algorithms were evaluated at various levels and using different evaluation strategies. By doing so, the project aims to provide valuable insights to researchers and medical practitioners, thereby enabling them to gain a better understanding of the most reliable and accurate method for predicting heart disease.

* 1. **What is heart disease?**

Heart disease is a complex medical condition that refers to a range of conditions that affect the heart and its associated blood vessels. These conditions may include coronary artery disease, arrhythmias, congenital heart defects, and other related blood vessel diseases. Cardiovascular disease is a term that is often used interchangeably with heart disease and generally refers to conditions that involve narrowed or blocked blood vessels that can lead to chest pain, heart attack, or stroke. Other heart conditions that affect the heart's muscle, valves, or rhythm are also considered forms of heart disease. Heart failure is a severe and prevalent condition that affects a significant percentage of the adult population in developed countries and is the leading cause of admission to healthcare professionals. It is a chronic and progressive condition that occurs when the heart muscle is unable to pump enough blood to meet the body's needs. Heart failure can cause a range of symptoms, including shortness of breath, fatigue, and swelling in the legs and ankles. It is associated with high medical costs, reaching up to 2% of the total health costs in developed countries.

**2) DATA COLLECTION AND DATA VISULIZATION: -**

**2.1) View of Dataset:**

****

**2.2) Describing the attribute of the dataset**

**a) Age:** This attribute represents the age of the individual in your dataset.

**b) Education:** This attribute could represent the education level of individuals.

**c) Sex:** This attribute signifies the gender of the individual.

**d) Is smoking:** This attribute indicates whether an individual is a smoker or not.

**e) BPMeds:** This attribute indicates whether an individual is taking blood pressure

medication or not.

**f) prevalentStroke:** This attribute indicates whether an individual has had a stroke or

not.

**g) prevalentHyp:** This attribute represents the prevalence of hypertension in

individual.

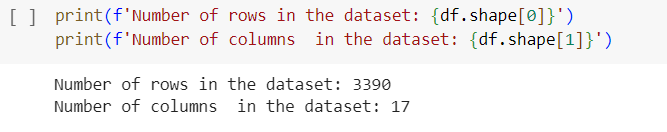
.

**h) Totchol:** This attribute represents the total cholesterol level in an individual's

blood. High levels are a risk factor for heart.

1. **sysBP and diaBP:** This attribute denotes systolic and diastolic blood pressure.
2. **BMI:** Represent an individual's body mass relative to their height. It is used to assess the impact of weight on cardiovascular issues.
3. **Heart Rate:** This attribute indicates an individual heart rate. Abnormal heart rate associated with cardiovascular disease.
4. **Glucose:** This attribute represents the blood glucose level. A high level can be indicative of diabetes.
5. **CHD:** This attribute indicates the presence or absence of coronary disease.

**2.3) SHAPE OF DATASET**

****

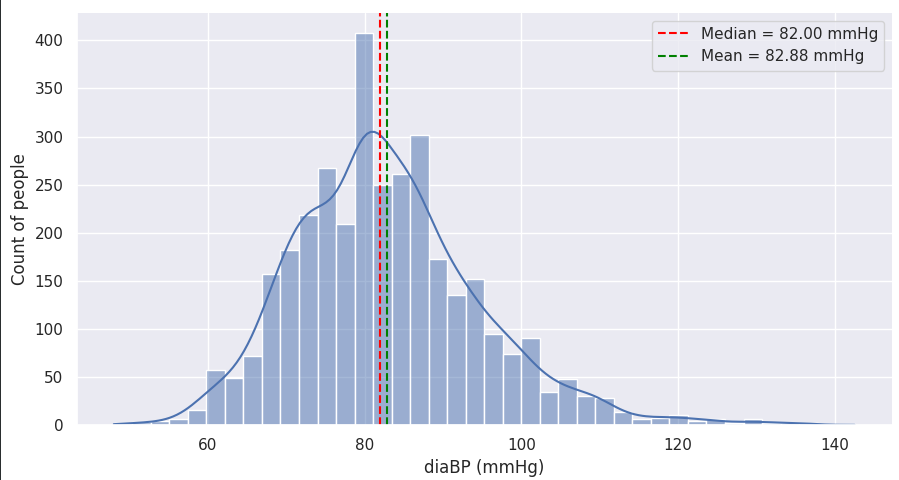
**df. shape:** It returns the number of rows and columns in the dataset.

**2.4) DATA VISULIZATION**

**A screen shot of a computer program

Description automatically generated**

**A graph of a graph

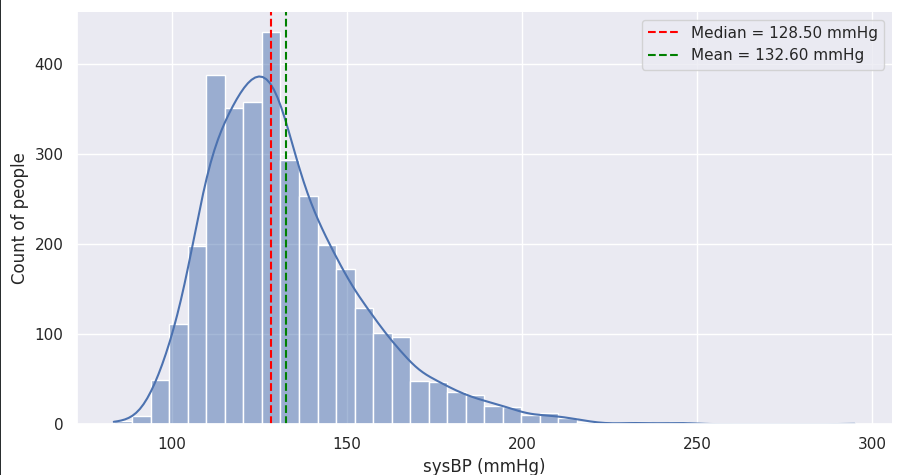
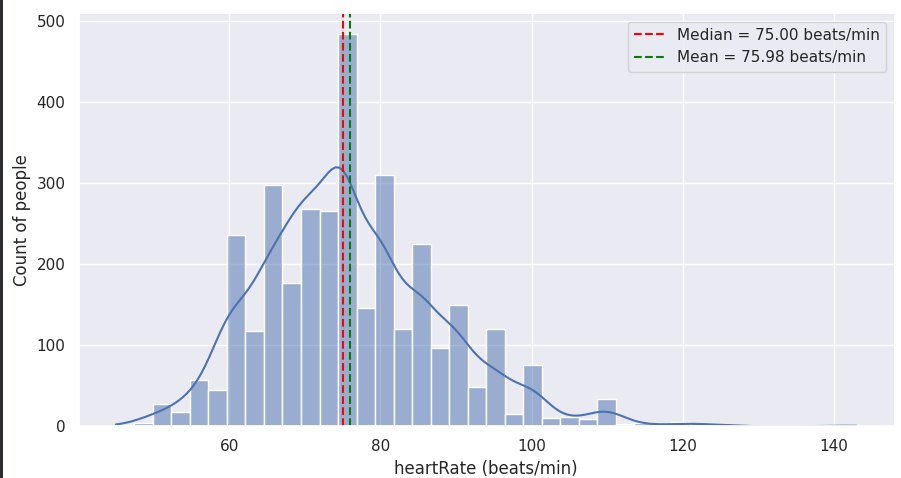
Description automatically generated**

**A graph with a green line and a red line

Description automatically generatedA graph of a graph of a graph

Description automatically generated with medium confidence**

**A graph of a normalized line

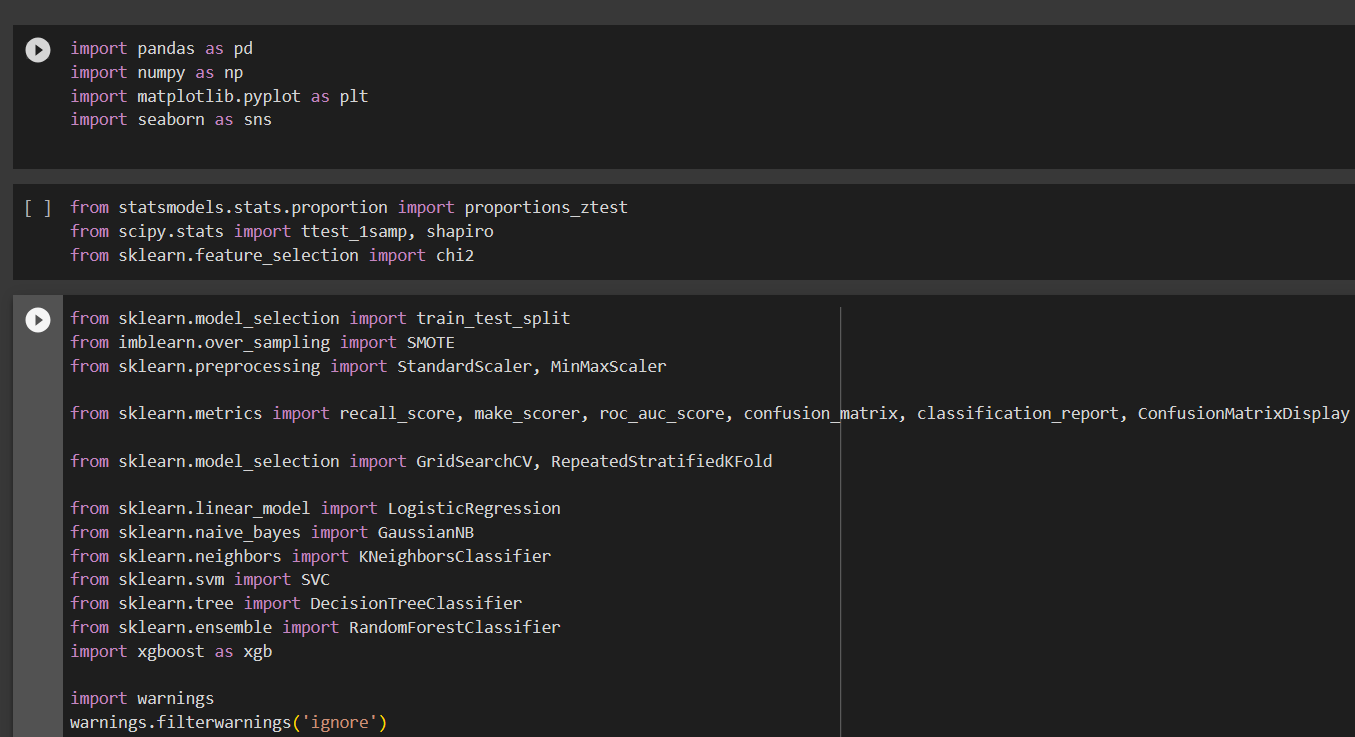
Description automatically generated with medium confidenceA graph of a normal distribution

Description automatically generated**

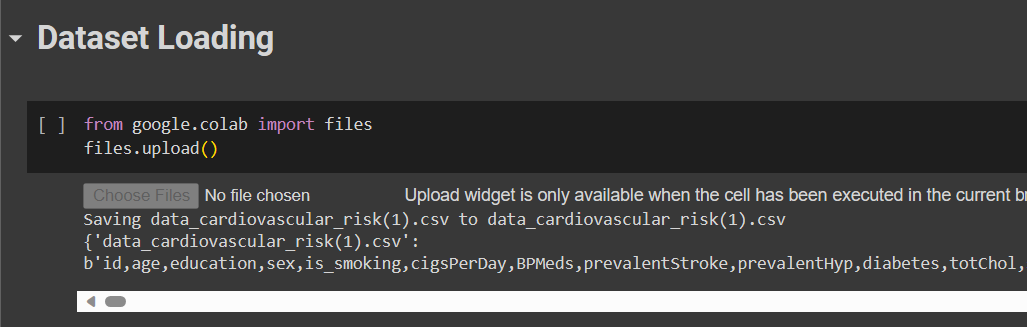
**3) Data Preprocessing and Train Test Split: -**

**Data preprocessing:** This involves cleaning, transforming, and organizing raw data into a format that is suitable for model training.

**3.1) Importing libraries**



**3.2) Load data**

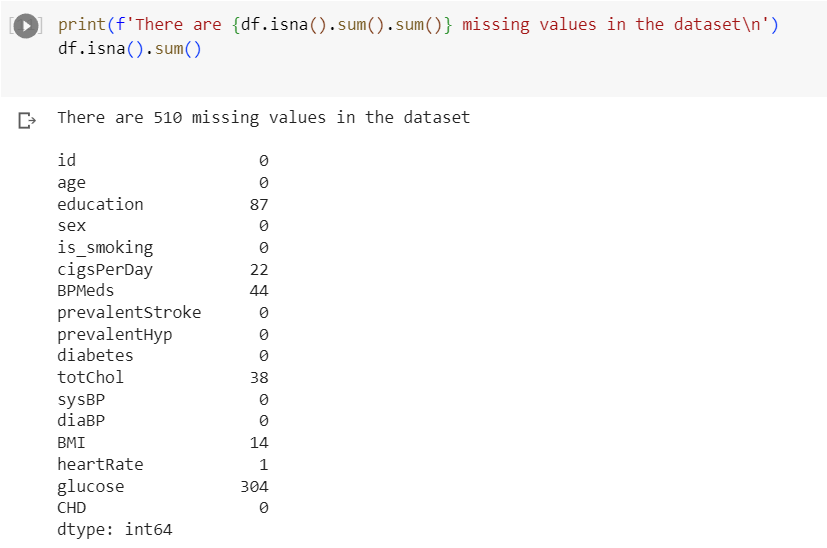


**3.3)** **Check the shape of the dataset.**

A screenshot of a computer code

Description automatically generated

**3.4) Check for the missing values.**

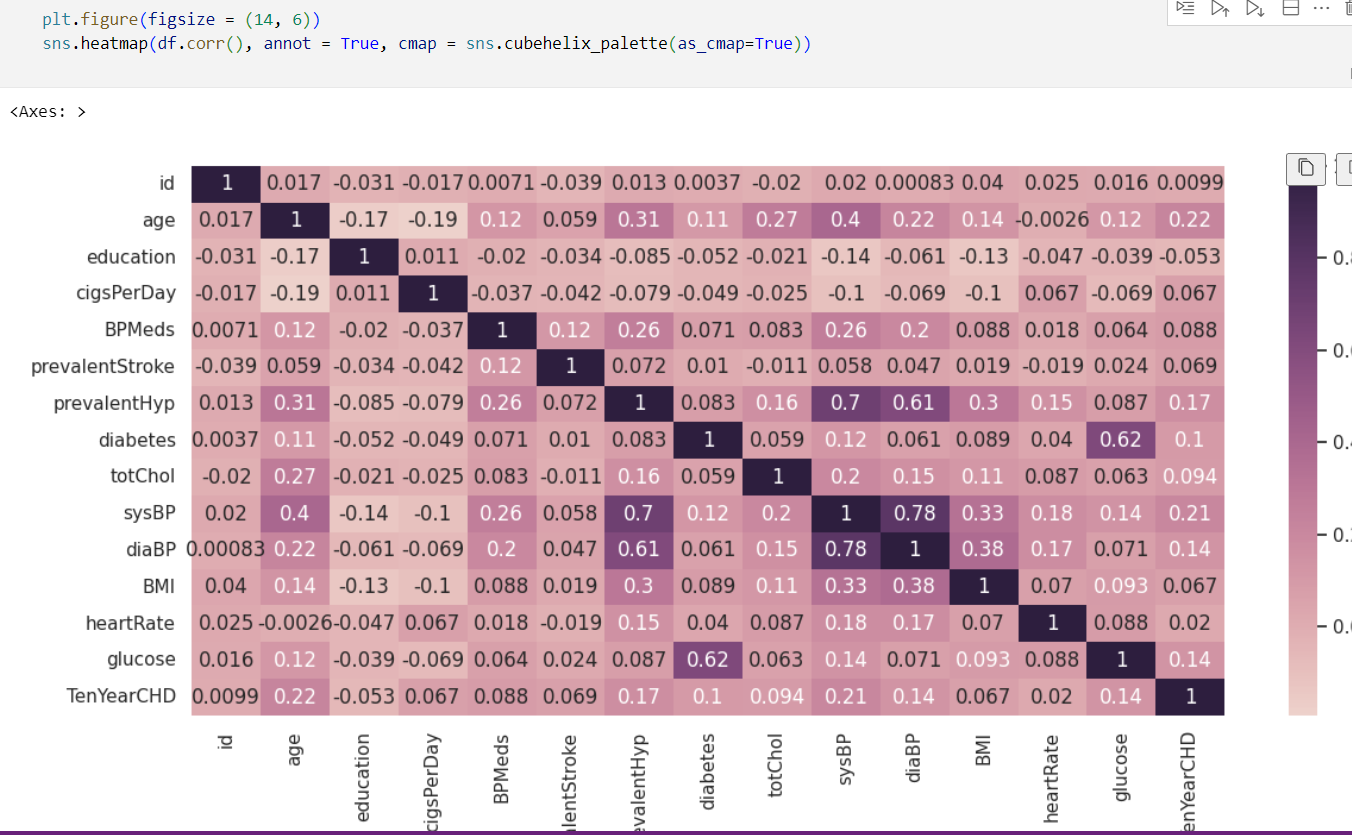


A graph with numbers and a bar

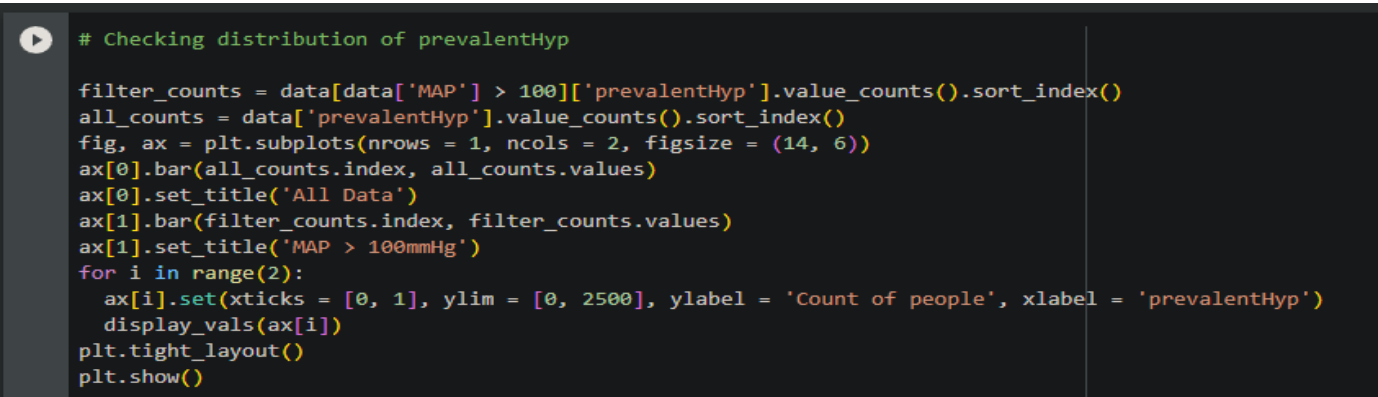
Description automatically generated

**3.4) Check the correlation.**





**3.5) Feature Engineering**



**A graph of data with numbers and a bar

Description automatically generated with medium confidenceA graph with blue squares

Description automatically generated**

**A computer screen shot of text

Description automatically generatedA screen shot of a computer program

Description automatically generated**

**3.5) Handling Imbalance Dataset:**

**SMOTE:** Smote, which stands for synthetic minority over-sampling technique. Smote helps in balancing the class distribution by creating synthetic examples for the minority class.

A close-up of a computer code

Description automatically generated

**3.6) Splitting the data:**



Splitting the data refers to dividing the dataset into two or more subsets for different purposes.

First, the training dataset is the largest portion of the dataset that is used to train the machine learning model.

Second, the test set is a separate portion of the dataset it is used to assess the final performance of the trained machine-learning model.

**4) MODELING AND PREDICTON: -**

**4.1) Logistic Regression:**

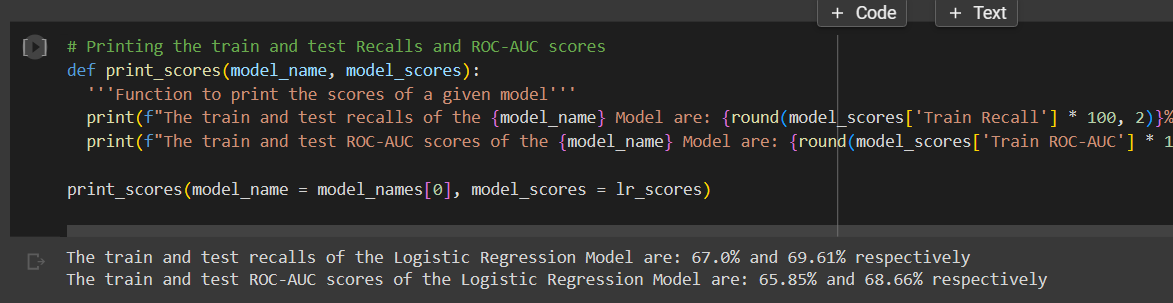
It is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated



**4.2) CONFUSION MATRIX**

A confusion matrix is a table or a matrix that is often used to describe the performance of

a classification model (or "classifier") on a set of test data for which the true values are

known.

A confusion matrix typically consists of four values:

1. **True Positives (TP):** These are cases in which the model predicted a positive class (e.g., "Yes" or "1") correctly, and the actual class is also positive.
2. **True Negatives (TN):** These are cases in which the model predicted a negative class (e.g., "No" or "0") correctly, and the actual class is also negative.
3. **False Positives (FP):** These are cases in which the model predicted a positive class incorrectly when the actual class is negative. Also known as a "Type I error."
4. **False Negatives (FN):** These are cases in which the model predicted a negative class incorrectly when the actual class is positive. Also known as a "Type II error."

**4.2.1) The confusion matrix is often used to calculate various performance metrics for a classification model, including:**

* **Accuracy:** The ratio of correctly predicted instances to the total instances. It measures the overall correctness of the model.

**Accuracy = (TP + TN) / (TP + TN + FP + FN)**

* **Precision:** Also called Positive Predictive Value, it measures the accuracy of positive predictions. It's the ratio of true positives to the sum of true positives and false positives.

**Precision = TP / (TP + FP)**

* **Recall:** Also called Sensitivity or True Positive Rate, it measures the model's ability to identify all relevant instances. It's the ratio of true positives to the sum of true positives and false negatives.

**Recall = TP / (TP + FN)**

* **F1-Score:** The harmonic mean of precision and recall. It provides a balance between precision and recall.

**F1-Score = 2 \* (Precision \* Recall) / (Precision + Recall)**

* **Specificity:** Also called True Negative Rate, it measures the model's ability to identify all negative instances.

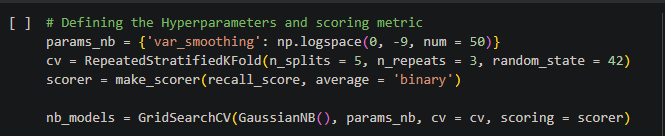
**Specificity = TN / (TN + FP)**

* **False Positive Rate (FPR):** The ratio of false positives to the sum of false positives and true negatives.

**FPR = FP / (FP + TN)**

**4.3) NAÏVE BAYES:**

Naive Bayes is a family of probabilistic machine learning algorithms used for classification and sometimes regression. It's based on Bayes' theorem, which is a fundamental theorem in probability theory.



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer program

Description automatically generated

**4.4) DECISION TREE:**

A decision tree is a popular supervised machine learning algorithm used for both classification and regression tasks. It's a tree-like model that is constructed by recursively partitioning the dataset into subsets based on the values of input features.

A computer screen with text

Description automatically generated

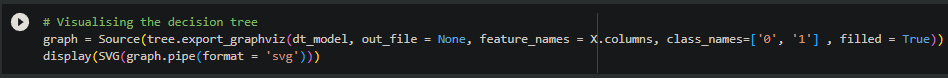
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**4.4.1) DECISION TREE VISUALISATION**



A diagram of a company

Description automatically generated with medium confidence

**4.5) K-NEAREST NEIGHBOURS:**

K-Nearest Neighbours (KNN) is a simple and widely used supervised machine learning algorithm for classification and regression tasks. It's known as a non-parametric, instance-based learning algorithm because it makes predictions based on the majority class or average value of the k-nearest data points in the training dataset.

**A black screen with white text

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screen shot of a graph

Description automatically generated**

**4.6) SUPPORT VECTOR MACHINE:**

A screenshot of a computer

Description automatically generated Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. SVM is particularly effective for classification tasks and is known for its ability to handle high-dimensional data, find complex decision boundaries, and work well in both linear and nonlinear scenarios.

A screenshot of a computer

Description automatically generated

**4.7) RANDOM FOREST:**

Random Forest is an ensemble machine learning algorithm used for both classification and regression tasks. It is a versatile and powerful algorithm known for its accuracy and robustness. Random Forest is an ensemble of decision trees, where multiple decision trees are trained on different subsets of the data and their predictions are combined to make a final prediction.

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

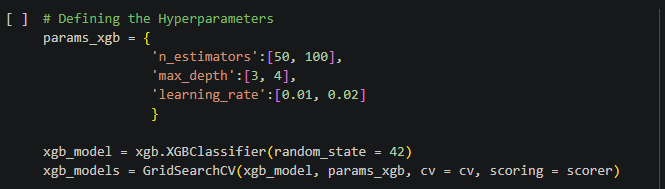
Description automatically generated

A screenshot of a computer program

Description automatically generated

**4.8) XGBOOST:**

XGBoost, short for "Extreme Gradient Boosting," is a popular and powerful machine learning algorithm that belongs to the gradient boosting family. It is known for its efficiency, speed, and high predictive accuracy, making it a popular choice for both classification and regression tasks.



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**4.9) MODEL EVALUATION:**

**A screenshot of a computer screen

Description automatically generated**

**A screen shot of a graph

Description automatically generated**

A screen shot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generated **5) MODEL PREDICTION:**

A screen shot of a computer

Description automatically generated

**6) USER INTERFACE WEB APP:**

**A screen shot of a computer program

Description automatically generatedA screen shot of a computer

Description automatically generated**

A person holding a heart with a stethoscope around their neck

Description automatically generatedA screenshot of a computer

Description automatically generated

**7) CONCLUSION AND FUTURE SCOPE:**

**7.1) CONCLUSION**

The development of a heart disease prediction model using machine learning techniques represents a significant step forward in the field of healthcare and medical diagnostics. The model's ability to analyze and interpret complex medical data to identify individuals at risk of heart disease has the potential to revolutionize early detection and intervention, ultimately saving lives and reducing healthcare costs.

In conclusion, the heart disease prediction model is a valuable tool for both patients and healthcare providers. Patients can benefit from early warnings and take proactive steps to improve their heart health, such as lifestyle modifications and regular check-ups. Healthcare providers can use the model to prioritize patients for further evaluation and tailor treatment plans to individual risk profiles.

**7.2) FUTURE SCOPE**

Looking ahead, the future scope of this model is promising. Here are some key areas for future research and development:

1. **Integration with Electronic Health Records (EHRs):** Integrating the prediction model with electronic health records can enhance its accuracy and applicability. Access to a patient's complete medical history and real-time data can lead to more precise predictions.
2. **Continuous Monitoring:** Expanding the model to support continuous monitoring of patients' health can provide ongoing risk assessments, allowing for timely interventions and personalized care plans.
3. **Multimodal Data:** Incorporating various data sources, such as genetic information, wearable device data, and patient-reported outcomes, can improve the model's predictive capabilities.
4. **Explain ability:** Enhancing the model's interpretability and providing clear explanations for predictions will be crucial for gaining trust among both patients and healthcare professionals.
5. **Global Accessibility:** Ensuring that the model is accessible and applicable to diverse populations worldwide will be essential for addressing heart disease on a global scale.
6. **Collaborative Research:** Collaboration between data scientists, medical experts, and policymakers can facilitate the development of more robust and ethical heart disease prediction models.

**8) REFERENCE**

**8.1) Datasets**

Popular sources include Kaggle, UCI Machine Learning Repository

**8.2) Books**

Classification Projects on Machine Learning for Beginners is a comprehensive reference guide for beginners.

**8.3) Papers**

Heart disease prediction using machine learning algorithms.

Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath

[*https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/?ref=lbp*](https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/?ref=lbp)

[*https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/*](https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/)

[*https://www.javatpoint.com/how-to-get-datasets-for-machine-learning*](https://www.javatpoint.com/how-to-get-datasets-for-machine-learning)

[*https://www.researchgate.net/publication/344717762\_Machine\_Learning\_Algorithms\_-A\_Review*](https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review)