"Mining the Heart: Leveraging Data Science for Deeper Understanding of Cardiovascular Diseases"

Abstract:

The abstract addresses the global prevalence of cardiac disease, including heart attacks, and emphasizes the need for improved early diagnosis and management techniques because of the substantial impact these conditions have on public health. It chronicles the development of coronary care units throughout history and highlights the necessity of automated detection techniques to deal with the difficulties of providing prompt aid during emergencies. The abstract also discusses how modern lifestyle choices, such as stress, aging, sugar, and cholesterol levels, increase the risk of heart disease. It emphasizes how important research is to improving our knowledge and care of heart attacks, establishing the study as a useful tool for the medical field. The abstract ends with a summary of the results of the analysis of the classification models, highlighting the superior performance of the Random Forest and Logistic Regression models over the others, and recommending that classification models be carefully chosen by job requirements. Lastly, it highlights the importance of trustworthy datasets and the potential of machine learning to enhance heart attack disease management and prevention. It also describes the thorough methodology used in the study, including statistical analysis, data exploration, preprocessing, and model development.

Introduction:

Heart disease continues to be the primary cause of death worldwide, necessitating ongoing improvements in diagnostic and prognostic techniques. The significant influence that heart disease, and especially heart attacks, have on public health highlights the need for improved early identification and management techniques [1]. When coronary care units were established in the 1960s, the inpatient mortality rates for myocardial infarction (heart attack) dropped dramatically from approximately 23–40% to 16–18%. A major factor in this decrease was improved identification and treatment of major rhythms [2]. According to the World Health Organization's (WHO) most recent figures, cardiovascular disease accounts for more than 37% of global deaths. Cardiovascular disease accounts for over 10% of global mortality in the 20th century [3][4]. The number of senior individuals living independently is rising, along with the aging population, which increases the likelihood that they won't get enough help in an emergency like a heart attack. The authors point out that because it is difficult to manually request assistance during such incidents, automatic detection methods are critically needed [5]. In the modern world of today. Our way of living has evolved from the past. Health, especially heart disease, was one area where changes were made. These are growing, which raises the risk to people. Every person has their own stress level, cholesterol, and heart rate. The heart beats at a regular pace of 72. The main issues are stress, aging, sugar, and cholesterol [6]. This is a potentially fatal disorder that requires medical attention right now. The body attempts to heal the damage caused by the death of the heart muscle, which increases the risk of problems and irregular heartbeat. An in-depth examination of the subject and a demonstration of the value of research in the quest for a better understanding and cure for heart attacks make this article a valuable resource for anyone in need of medical attention [7][8].

Background: Heart attacks can have a variety of causes. The development of plaque in the coronary arteries that is high in cholesterol is the common cause. Atherosclerosis, or the hardening of the arteries and the accumulation of deposits of a fatty nature, is the cause of plaque. Elevated amounts of fat and cholesterol in the circulation cause damage to the artery walls. A cholesterol level that is higher than the 75th percentile for a given age and sex is considered high blood cholesterol. Heart attacks are frequently caused by elevated blood pressure as well [9]. Large increases in government and public financing for studies on heart attacks and heart disease occurred between the 1960s and the 1980s. This ultimately resulted in the National Heart, Lung, and Blood Institute (NHLBI) being founded. The NHLBI is a primary funding source for investigator-initiated research concerning heart attacks and heart disease. A significant rise in published research on heart attacks and heart disease from the 1970s to the new century is indicative of this increasing financing.[10][11][12].

Literature Review:

Models	**Naïve Bayes, Decision Tree, KNN.**
Strength	Its comprehensive study of methods like Decision Trees, Naïve Bayes, and Neural Networks ensures that their eligibility for use in medical diagnostics is extensively investigated, which is one of its key benefits. An important breakthrough is the application of genetic algorithms to enhance feature selection and boost model predictive accuracy using real-world data from the Cleveland Heart Disease database.
Weakness	The complex nature of the methods and the depth of statistical analysis may make medical professionals without technological expertise less accessible.
Main Contribution	Its thorough comparative research and real-world use of genetic algorithms to maximize data mining for heart disease prediction constitute its main contributions. These discoveries could improve patient outcomes by enabling more accurate early detection, which would be a major step forward in the application of data science to healthcare.
Result	The idea that better hospital management contributed to the lower community-wide mortality rates from heart disease was supported by the result showing that the greater usage of these treatments was correlated with a drop in inpatient mortality rates.

Reference [13].

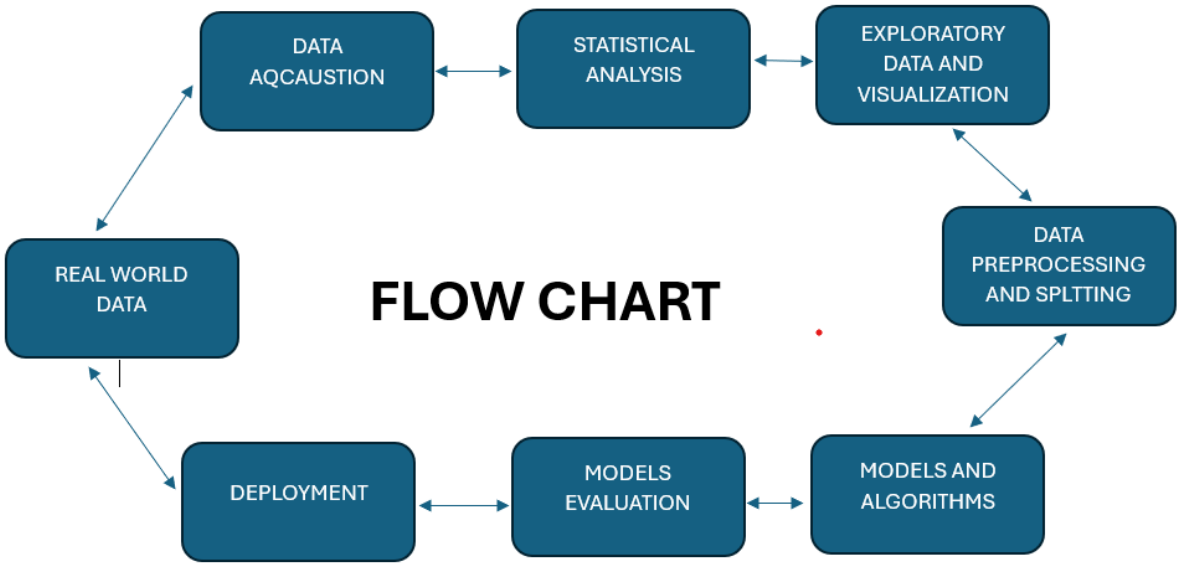
Models	**KNN, Logistic Regression, Random Forest Classifier.**
Strength	Methods using machine learning to increase the precision of heart disease prediction, such as Random Forest, KNN, and Logistic Regression. With KNN being the most accurate at 88.52%, it obtains a high predicted accuracy of 87.5%. This suggests that the patient screening process for heart disease is working well. By precisely estimating heart disease risks using past medical data, the method also saves money by minimizing the necessity for pointless medical testing.
Weakness	The study employs a dataset from the UCI repository, which may not be as thorough as it may be because it does not include all significant factors or fully represent the global population. Concerns have also been raised over the models' applicability to other population populations not included in the UCI dataset that have distinct demographic and genetic origins.
Main Contribution	The main contribution is the use of numerous machine learning algorithms to forecast cardiac disease, which is an improvement above prior systems that employed fewer techniques. It gives medical staff members a useful tool that promotes early interventions and improved patient care. Furthermore, by comparing different algorithms, the study adds significant knowledge about the best prediction methods, enhancing the field of medical data analysis.
Result	The results showed that combining many machine learning algorithms can greatly improve the prediction accuracy of cardiac disease. With the best accuracy scores, KNN and Logistic Regression were noted as being very successful. This underscores the promise of machine learning in clinical settings to help with early detection and management of cardiac disease.

Reference [14].

Methodology:

The (Figure 1) streamlined workflow is depicted in the flow chart. To better comprehend the data, real-world data collection is the first step, which is then followed by statistical analysis and exploratory data visualization. After that, data is split into training and test sets and pre-processed before different models and algorithms are applied. Following an evaluation of the models, the top-performing model is implemented for practical usage.

Figure 1



DATA AQCAUSTION:

The dataset was acquired from the well-known data science competition and dataset website Kaggle (https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data) [15]. It consists of clinical records, including age, blood pressure, and other health indicators, of people who may be at risk of heart attacks. After downloading, the data's reliability and accuracy were checked to make sure it was suitable for analysis. After that first step is to upload the dataset in the Python (Jupyter Notebook) environment using the Pandas library.

DATA DESCRIPTION:

Age: A continuous variable indicating the patients' ages.
Anemia: A binary variable that shows whether anemia is present or not.
Creatinine phosphokinase (CPK): A continuous variable that indicates the blood's concentration of the CPK enzyme (mcg/L).
Diabetes: A true-false variable indicating the presence of diabetes in the patient.
Ejection_Fraction: A continuous variable that shows how much blood the heart pumps out with each beat.
High_Blood_Pressure: A binary variable that signifies the presence of hypertension in the patient.
Platelets: A continuous variable that indicates how many platelets there are per milliliter of blood (kilo platelets/mL).
Serum_Creatinine: The blood's serum creatinine level (mg/dL) is indicated by this continuous variable.
Serum_Sodium: The continuous variable indicates the blood's sodium concentration (mEq/L).
Sex: A binary variable with 0 denoting female and 1 representing male.

Smoking: A binary variable that represents the patient's smoking status.
Time: A constant in days for the duration of the follow-up.
DEATH_EVENT: A binary target variable that indicates if the patient passed away while being monitored [15].

STATISTICAL ANALYSIS:

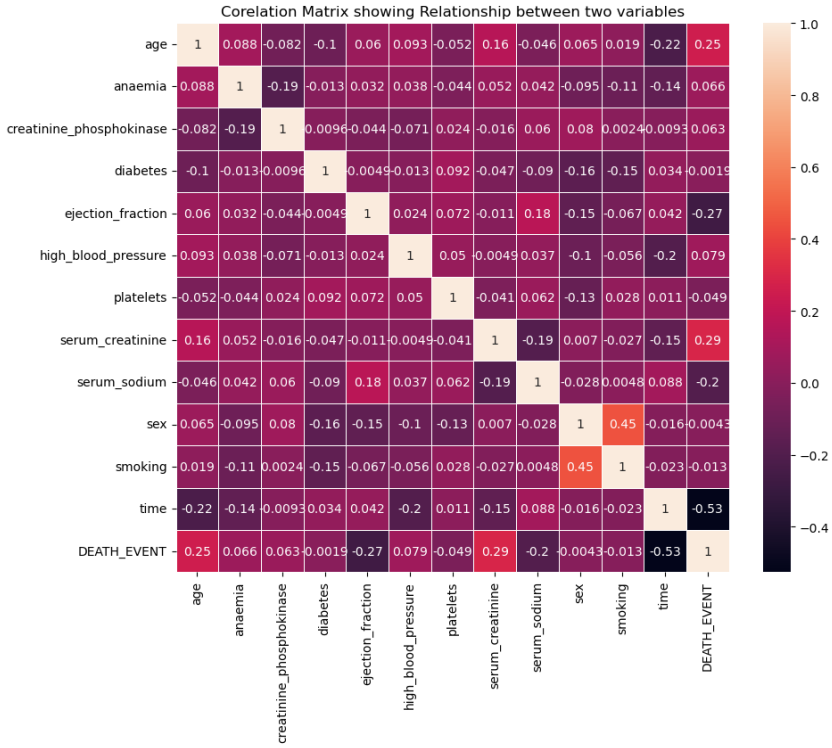
Finding and assessing the significance of correlations between different variables was the aim of the statistical analysis carried out on the heart attack dataset. Descriptive statistics were used to analyze the distribution, central tendency, variability, and general form of the data by utilizing important statistical measures including mean, median, standard deviation, minimum, maximum, and quartiles. This study serves as the foundation for further investigation and possible creation of intervention methods aimed at enhancing patient outcomes. In jupter notebook use the code .describe().**Figure 2:**

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
count	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000
mean	60.833893	581.839465	38.083612	263358.029264	1.39388	136.625418	130.260870
std	11.894809	970.287881	11.834841	97804.236869	1.03451	4.412477	77.614208
min	40.000000	23.000000	14.000000	25100.000000	0.50000	113.000000	4.000000
25%	51.000000	116.500000	30.000000	212500.000000	0.90000	134.000000	73.000000
50%	60.000000	250.000000	38.000000	262000.000000	1.10000	137.000000	115.000000
75%	70.000000	582.000000	45.000000	303500.000000	1.40000	140.000000	203.000000
max	95.000000	7861.000000	80.000000	850000.000000	9.40000	148.000000	285.000000

	anaemia	diabetes	high_blood_pressure	sex	smoking
count	299.000000	299.000000	299.000000	299.000000	299.00000
mean	0.431438	0.418060	0.351171	0.648829	0.32107
std	0.496107	0.494067	0.478136	0.478136	0.46767
min	0.000000	0.000000	0.000000	0.000000	0.00000
25%	0.000000	0.000000	0.000000	0.000000	0.00000
50%	0.000000	0.000000	0.000000	1.000000	0.00000
75%	1.000000	1.000000	1.000000	1.000000	1.00000
max	1.000000	1.000000	1.000000	1.000000	1.00000

Correlation matrix: With .corr() we can know the correlation between two variables and using a (Figure 3) correlation matrix heatmap, one can see the direction and intensity of connections between different variables in a dataset on heart attacks. Higher correlation values are indicated by stronger colors, a perfect positive correlation is represented by a value of 1, and a perfect negative correlation by a value of -1. Significantly, 'time' and 'DEATH_EVENT' exhibit a strong negative correlation (-0.53), implying that an extended follow-up period is connected with reduced mortality. Conversely, 'serum_creatinine' and 'DEATH_EVENT' demonstrate a positive correlation (0.29), implying that elevated creatinine levels may suggest a higher risk of mortality.

Figure 3:



EXPLORATORY DATA ANALYSIS AND VISUALIZATION:

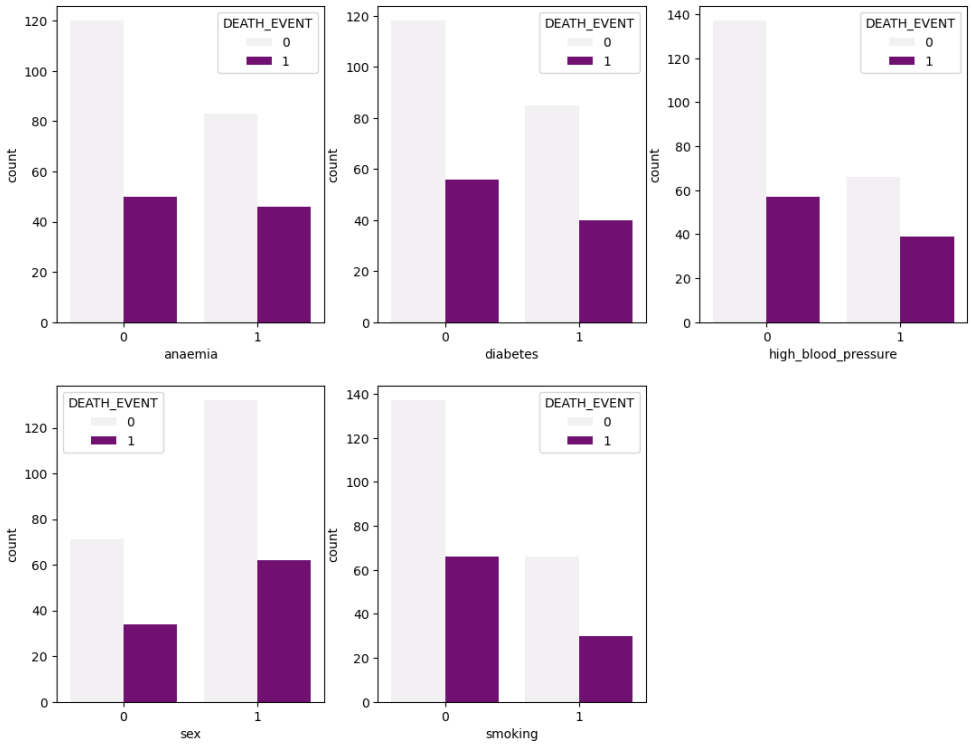
Missing Values: There was no need to perform data cleaning on this dataset because we had previously obtained clean data, without null and absent values, when we downloaded it from the website as shown in (Figure 4).

Figure 4:

```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   diabetes                  299 non-null    int64
 4   ejection_fraction         299 non-null    int64
 5   high_blood_pressure       299 non-null    int64
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64
 9   sex                       299 non-null    int64
 10  smoking                   299 non-null    int64
 11  time                      299 non-null    int64
 12  DEATH_EVENT               299 non-null    int64
dtypes: float64(3), int64(10)
```

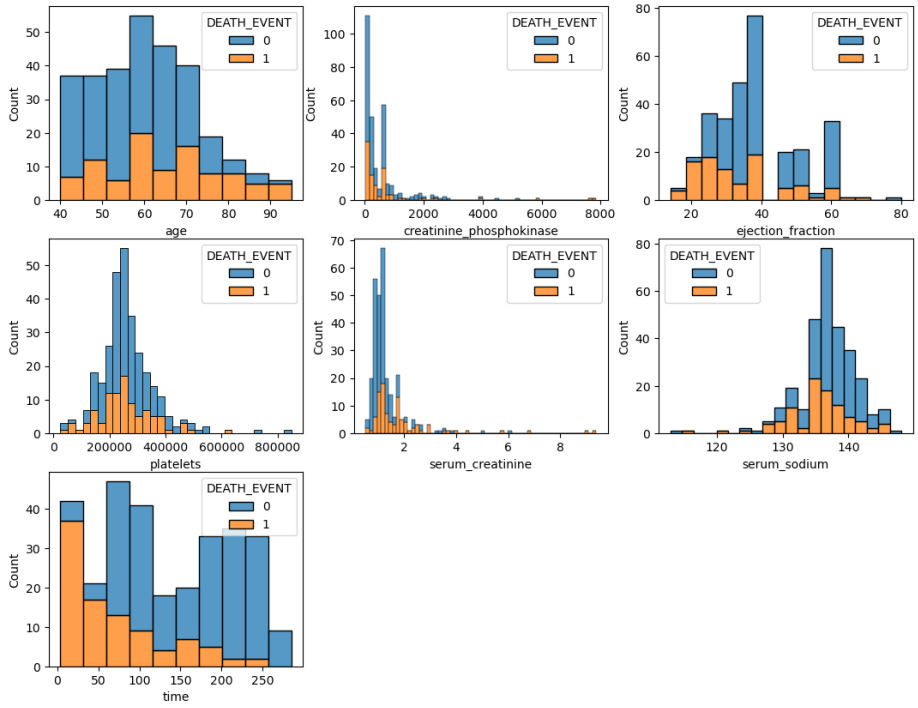
Visualization: In (Figure 5) the count plots are bar charts that display the proportion of people who died (shown in purple) and those who did not (shown in gray) for various conditions, including gender, smoking status, anemia, diabetes, and high blood pressure. A visual comparison of each condition's or characteristic's existence or absence of the incidence of death events in the dataset is provided by each pair of bars.

Figure 5: Count Plot.



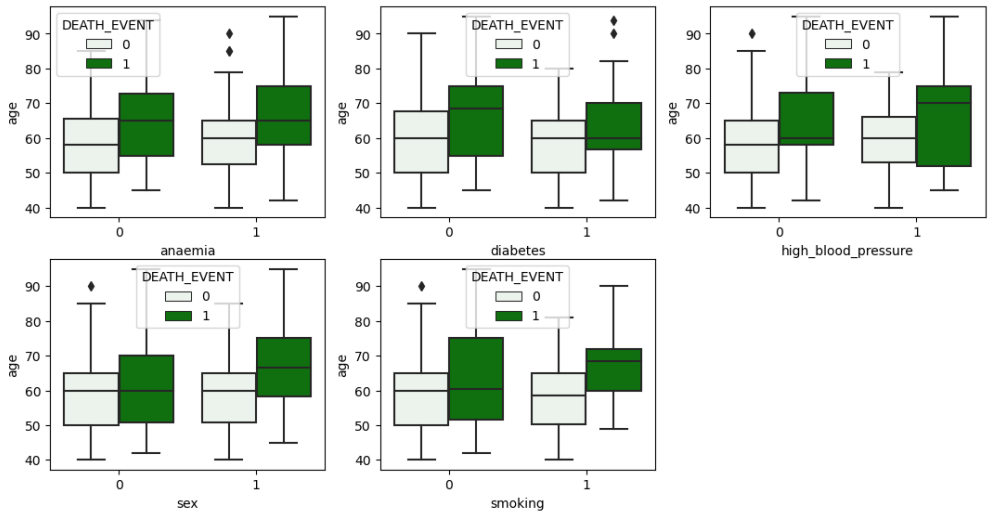
In (Figure 6) the distribution of several health-related variables against the frequency of mortality events is shown by the histograms. Age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and time are the variables that each histogram reflects. Blue bars indicate no death event (0) and orange bars indicate a death event (1). The variable's range of values is displayed on the x-axis, and the number of people falling within each range is displayed on the y-axis. The distribution of the variable for people who experienced a death event in comparison to people who did not is depicted by the overlap of the bars.

Figure 6: Histogram.



In (Figure 7) these box plots compare death events to the age distribution of people with and without specific illnesses (anemia, diabetes, high blood pressure), by sex, and smoking status. The ages of people who did not experience a death event are represented by the white boxes (zero), while those who did are represented by the green boxes (1). The interquartile range (IQR, the box itself), the median age (the line within the box), and any outliers (diamond shapes) are displayed in each box plot. The range of the data, excluding outliers, is shown by the "whiskers" (lines that extend from the boxes). Comparing age distributions linked to death events across several categories is made possible by these charts.

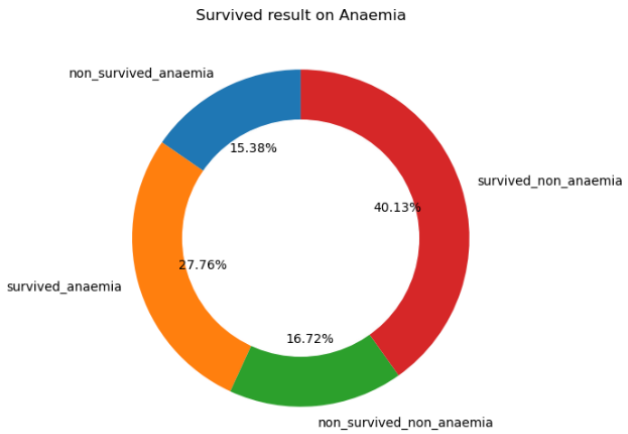
Figure 7: Boxplot.



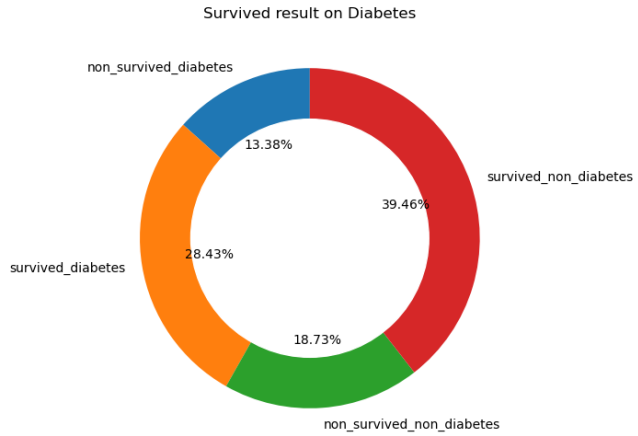
In (Figures 8,9,10,11, and 12) the survival results for people with and without different conditions are represented visually by the donut charts. Each chart has four sections that show whether an individual has the condition or not, as well as whether they survived or not (death events).

Each chart has two sections, which include those who have survived (divided into those who have the ailment and those who have not), and those who have not survived (also divided into those who have and have not). Each segment's percentages provide a brief overview of the correlations between each condition and trait and survival.

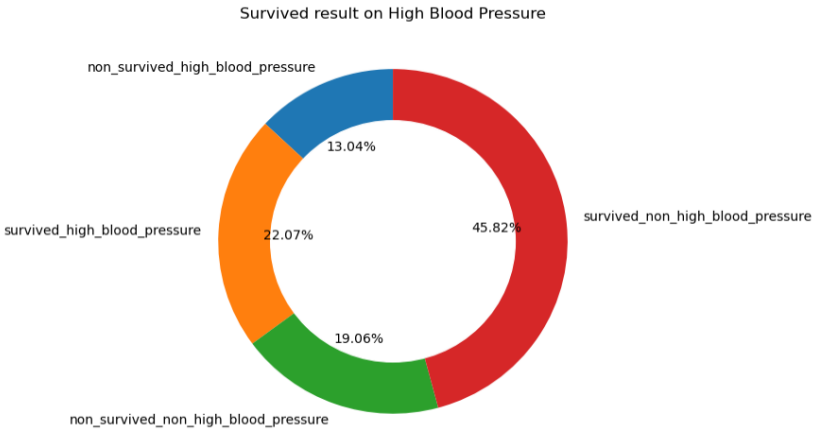
Figures 8,9,10,11 and 12: Dount Chart



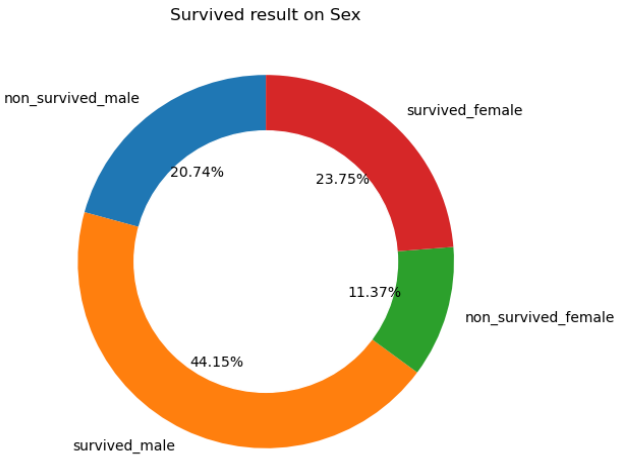
Anemia Chart: Indicates the percentage of individuals with and without anemia who are survivors and non-survivors. (Figure 8)



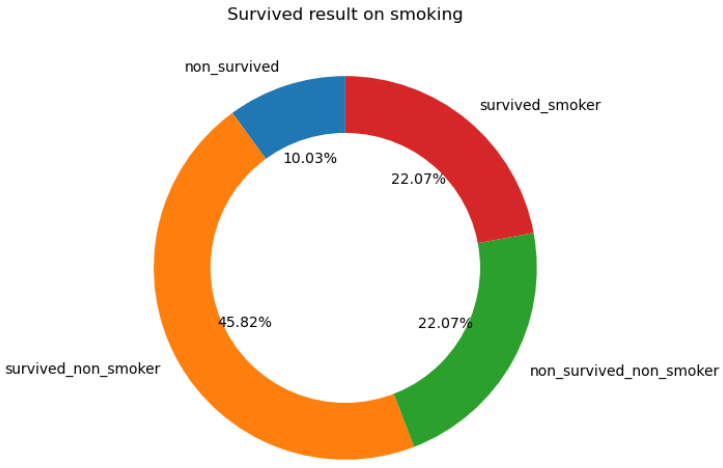
The diabetes chart presents the survival statistics of individuals with and without the disease. (Figure 9)



High Blood Pressure Chart: This illustrates the survival rates of those with high blood pressure in comparison to those without it. (Figure 10)



Sex Chart: By breaking down the data by gender, this chart shows the survival rates for both male and female individuals. (Figure 11)



The Smoking Chart displays the average lifespan of smokers and non-smokers. (Figure 12)

DATA PREPROCESSING AND SPLITTING:

The data will be split into X and Y using feature splitting. One dependent feature in Y will be the DEATH_EVENT, and all other independent variables will be kept in X. The features are then normalized using StandardScaler to guarantee that their mean is 0 and their standard deviation is 1. It is important to prevent features with bigger scales from outweighing features with smaller scales, as this might result in poor results for many machine learning methods. Then, the dataset is split into 70% and 30% training (X_train, y_train) and testing (X_test, y_test) data. Using the training data, the model is trained, and its ability to predict the outputs is tested.

MODELS AND ALGORITHMS:

Model Selection: We can take into account the following model selection since the dataset is utilized to predict binary outcomes from a set of heart disease predictors.

1. Decision Tree.
2. Logistic Regression.
3. Support Vector Machine.
4. Random Forest
5. K-Nearest Neighbors.
6. Naïve Bayes.

Model Training:

1. Decision Tree: The first step is to import the DecisionTreeClassifier class from sklearn.Tree. To train the models, use the instancesfit () method and supply the training data (X_train, y_train).
2. Importing the LogisticRegression class from sklearn.linear_model is the primary step in performing logistic regression. To train the model, use the fit () method with the training data (X_train, y_train). It then learns the weight that lowers the target's prediction mistakes.
3. Support Vector Machine, or SVM: Classification tasks is the Support Vector Classifier (SVC) class, which is imported from sklearn.svm. Use the fit () method with training data to train the SVM model. Finding the dataset's hyperplane that best divides the classes is the aim of this step.
4. Random Forest: Import the RandomForestClassifier from sklearn.ensemble to create a model instance. Using the training data as an argument, use the fit function on your model instance. This method trains the Random Forest model by building many decision trees on various dataset subsamples and averaging their predictions.
5. Import KNeighborClassifier from sklearn.neighbors to begin using KNN (K-Nearest Neighbors). Give the KNN instance fit () function to the training data. This method modifies the data to get the model ready for prediction.
6. The Naive Bayes classifier should be imported from scikit-learn. Next, create a Naive Bayes classifier instance. Gaussian Naive Bayes is a popular option in scikit-learn for classification problems; to train the Naive Bayes classifier, use the fit() method. Give this function your training data, which consists of the features and associated labels.

Model Evaluation:

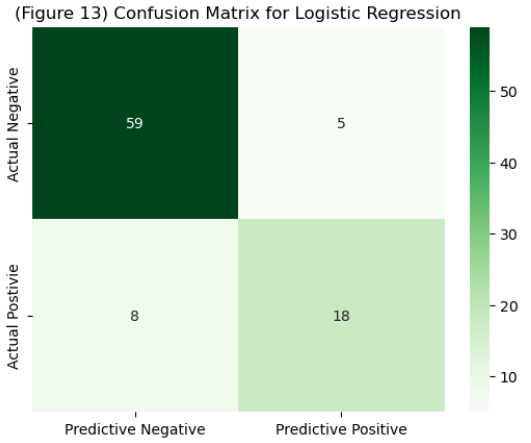
The models that were assessed show that Logistic Regression and Random Forest are the most accurate, with scores of 0.86 and 0.87, respectively, showing that they are good at classifying the data. With accuracies of 0.81, SVM and KNN models likewise function rather

well; however, when compared to Random Forest and Logistic Regression, their ROC scores indicate somewhat lower discriminating powers. With accuracies of 0.83 and 0.79, respectively, the Decision Tree and Naive Bayes models perform less robustly than the other models. Based on their superior accuracy scores and ROC performance, Random Forest and Logistic Regression seem to be the best models overall for this classification assignment.

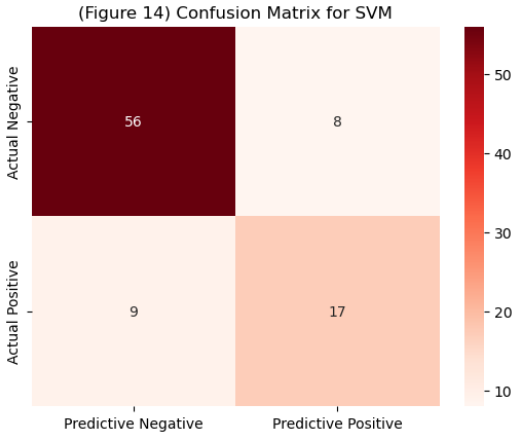
Confusion Matrix:

To determine the accuracy score of the model, confusion matrix measures must be performed. This confusion matrix can be used to see how well a model is performing. The amount of Actual Positive, Actual Negative, Predicted Positive and Predicted Negative is shown in (Figures 13, 14, 15, 16, 17, and 18). This separation is critical for comprehending the model behavior in medical diagnostics, where a Predicted Negative could have severe repercussions.

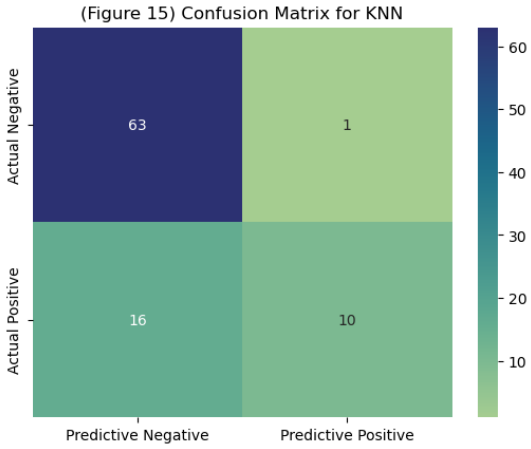
(Figure 13) Confusion Matrix for Logistic Regression



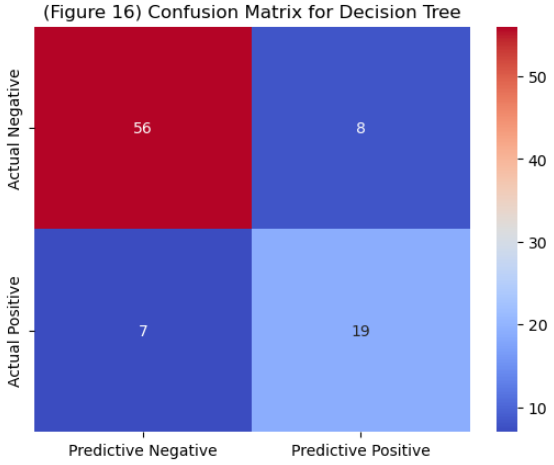
(Figure 14) Confusion Matrix for SVM



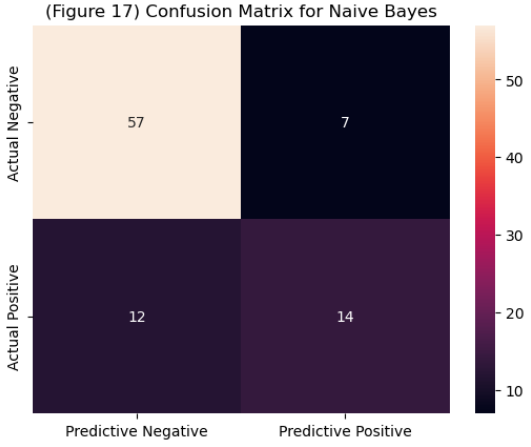
(Figure 15) Confusion Matrix for KNN



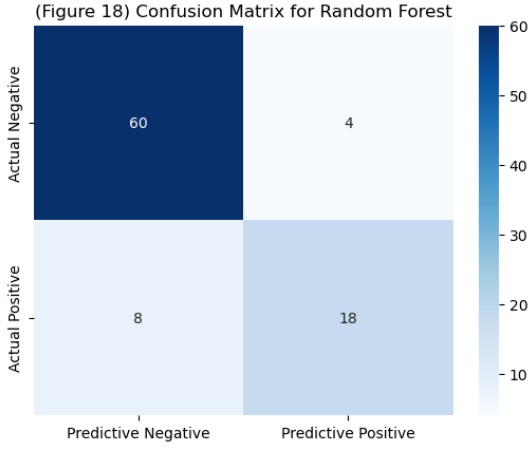
(Figure 16) Confusion Matrix for Decision Tree



(Figure 17) Confusion Matrix for Naive Bayes



(Figure 18) Confusion Matrix for Random Forest

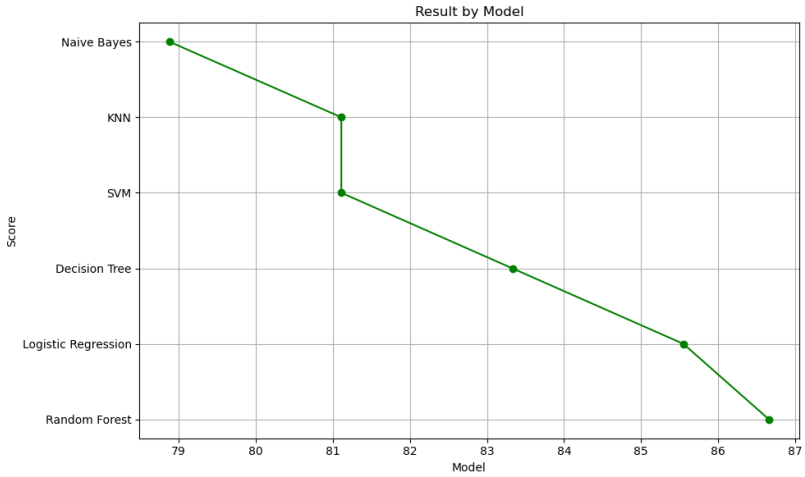


RESULT AND DISCUSSION:

An overview of the performance scores of the different classification models is given by the table and graph. With scores between 86 and 87%, Random Forest and Logistic Regression are the best methods constantly. Decision Tree comes in second, with an accuracy of 83–84%, which is moderate. SVM and KNN both perform at about the same level, with scores of about 81%. Naive Bayes frequently performs less accurately, with a score of less than 79%. All things considered, Random Forest and Logistic Regression perform better than the other models, while Naive Bayes is never very accurate.

Figures 19 and 20:

	Models	Scores
0	Random Forest	86.66
1	Logistic Regression	85.55
2	Decision Tree	83.33
3	SVM	81.11
4	KNN	81.11
5	Naive Bayes	78.88



In comparison to other models such as Decision Tree, SVM, KNN, and Naive Bayes, the results indicate that Random Forest and Logistic Regression are the most dependable models for classification tasks, constantly reaching greater accuracy. While SVM and KNN perform similarly but with somewhat lesser accuracy, the Decision Tree displays moderate accuracy. In terms of accuracy, Naive Bayes regularly performs worse than other methods. These results emphasize how crucial it is to choose the right classification models depending on the particular needs of the task.

CONCLUSION:

The conclusions of our extensive research, which included statistical analysis, data exploration, visualization, preprocessing, and model construction, are presented in this report. Our results highlight the usefulness of machine learning approaches in estimating the risk of a heart attack and provide insight into the components linked to heart attack disease. We draw attention to machine learning's potential as an aid in medical diagnostic decision-making. Our model's accuracy in classifying patients highlights how technology may help with early identification and management, which is essential for controlling chronic illnesses like heart attack disease. Furthermore, our report emphasizes how important it is to have trustworthy databases. It is essential to have sufficient and high-quality data when building trustworthy predictive models. Our study establishes the foundation for future research initiatives targeted at improving heart attack disease management and preventive strategies by adding to the growing field of healthcare analytics.

REFERENCES:

1: Patil, S.B. and Kumaraswamy, Y.S., 2009. Extraction of significant patterns from heart disease warehouses for heart attack prediction. *IJCSNS*, *9*(2), pp.228-235.

2. Brown, N., Young, T., Gray, D., Skene, A.M. and Hampton, J.R., 1997. Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register. *BMJ*, *315*(7101), pp.159-164.

3. Rajamhoana, S.P., Devi, C.A., Umamaheswari, K., Kiruba, R., Karunya, K. and Deepika, R., 2018, July. Analysis of neural networks based heart disease prediction system. In *2018 11th international conference on human system interaction (HSI)* (pp. 233-239). IEEE.

4. Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), pp.43-48.

5.Rojas-Albarracin, G., Chaves, M.Á., Fernandez-Caballero, A. and Lopez, M.T., 2019. Heart attack detection in colour images using convolutional neural networks. *Applied Sciences*, *9*(23), p.5065.

6. Gunji, J., MohamadMahaboob, S., Devarakonda, A., Muppavarupu, L.V., Inunganbi, S.C. and Bulla, S., 2023. Detection of Heart Attack Symptoms using DeepLearning.

7. Clerico, A., Zaninotto, M., Passino, C., Aspromonte, N., Piepoli, M.F., Migliardi, M., Perrone, M., Fortunato, A., Padoan, A., Testa, A. and Dellarole, F., 2021. Evidence on clinical relevance of cardiovascular risk evaluation in the general population using cardio-specific biomarkers. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *59*(1), pp.79-90.

8. Indrakumari, R., Poongodi, T. and Jena, S.R., 2020. Heart disease prediction using exploratory data analysis. *Procedia Computer Science*, *173*, pp.130-139.

9. Cenko, Edina, Lina Badimon, Raffaele Bugiardini, Marc J. Claeys, Giuseppe De Luca, Cor de Wit, Geneviève Derumeaux et al. "Cardiovascular disease and COVID-19: a consensus paper from the ESC working group on coronary pathophysiology & microcirculation, ESC working group on thrombosis and the association for acute CardioVascular care (ACVC), in collaboration with the European heart rhythm association (EHRA)." *Cardiovascular research* 117, no. 14 (2021): 2705-2729.

10. Soni, J., Ansari, U., Sharma, D. and Soni, S., 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), pp.43-48.

11. Navar, A.M., Fine, L.J., Ambrosius, W.T., Brown, A., Douglas, P.S., Johnson, K., Khera, A.V., Lloyd-Jones, D., Michos, E.D., Mujahid, M. and Muñoz, D., 2022. Earlier treatment in adults with high lifetime risk of cardiovascular diseases: What prevention trials are feasible and could change clinical practice? Report of a National Heart, Lung, and Blood Institute (NHLBI) Workshop. *American journal of preventive cardiology*, *12*, p.100430.

12. Kho, A., Daumit, G.L., Truesdale, K.P., Brown, A., Kilbourne, A.M., Ladapo, J., Wali, S., Cicutto, L., Matthews, A.K., Smith, J.D. and Davis, P.D., 2022. The National Heart Lung and Blood Institute disparities elimination through coordinated interventions to prevent and control heart and lung disease alliance. *Health services research*, *57*, pp.20-31.

13. Gooding, H.C., Gidding, S.S., Moran, A.E., Redmond, N., Allen, N.B., Bacha, F., Burns, T.L., Catov, J.M., Grandner, M.A., Harris, K.M. and Johnson, H.M., 2020. Challenges and opportunities for the prevention and treatment of cardiovascular disease among young adults: report from a National Heart, Lung, and Blood Institute Working Group. *Journal of the American Heart Association*, *9*(19), p.e016115.

14. Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P., 2021. Heart disease prediction using machine learning algorithms. In *IOP conference series: materials science and engineering* (Vol. 1022, No. 1, p. 012072). IOP Publishing.

15. Data Collection from: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data