**COMPARATIVE ANALYSIS FOR HEART DISEASE PREDICTION**

Student names

Subject

Date

**Abstract**

Heart disease is the leading cause of death globally, and accurate heart disease risk prediction is essential for prevention and treatment. This research's primary aim was to analyze different data mining models for their efficacy in predicting heart disease. We systematically reviewed the literature and identified three studies that compared different methods for predicting heart disease risk. These studies formed the basis of our comparative analysis. We obtained a dataset from the UCL Machine Learning Repository to help us quantitatively analyze the performance of different models using the CRISP-DM methodology. We cleaned and preprocessed the data, then performed an EDA to understand the data better. After that, we applied three different prediction models (logistic regression, support vector machine, and k nearest neighbors) to the data and evaluated the performance of each model using a classification accuracy metric. Our results showed that the linear regression model had the best performance, with an accuracy of 80.22%. We concluded that the linear regression model is the best algorithm to predict heart disease risk. The insights from this study can help healthcare practitioners to predict heart disease risk better and take appropriate preventive measures.

***Keywords: heart disease, classification, comparative analysis***

# **INTRODUCTION**

Despite advances in treatment and prevention, the burden of heart disease continues to rise, particularly in low- and middle-income countries. Data mining has been used extensively in the healthcare sector to build predictive models for the prediction of heart diseases [1]. It has led to a better understanding of the risk factors associated with heart diseases, improving the efficiency of healthcare operations and reducing the cost of care. The use of data mining in healthcare is expected to grow in the future as more data sets become available and as the need for more effective predictive models increases [2]. These models will help healthcare providers to target their resources better and to provide more personalized care to their patients.

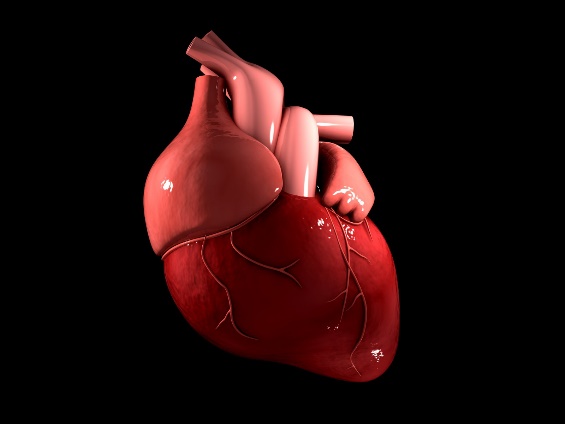


Figure 1: Heart

Heart disease is the leading cause of death globally. According to the World Health Organization (WHO), cardiovascular disease is responsible for more than 17.3 million deaths yearly, about 31% of all deaths globally [3]. Heart disease prediction involves estimating an individual's risk of developing heart disease over a specific period. This is usually done using statistical methods, and the results are typically presented as a percentage or as a number of cases per thousand people. There are many risk factors that can affect an individual's risk of developing heart disease, including family history, lifestyle choices (such as smoking and diet), and medical conditions (such as high blood pressure and diabetes)[4]. The accuracy of heart disease predictions can vary depending on the population being studied and the methods used. However, heart disease prediction can be a valuable tool in helping to identify those individuals who may benefit from lifestyle changes or medical intervention.

The number of deaths due to heart disease is expected to continue to rise, making it even more essential to find effective ways to prevent and treat this condition. A study by the WHO suggests that heart disease mortality is projected to increase to over 23.6 million between 2020 and 2030 [5]. The cost of diagnosing and treating heart disease can be prohibitive for many people. Moreover, the risk of complications and death associated with heart disease is significant; hence, more effective ways to prevent and treat this condition are needed. Data mining can be used to identify data patterns that can help predict which individuals are at risk of developing heart disease. This information can then be used to develop targeted interventions to prevent the onset of heart disease. There are various data mining techniques that can be used for the prediction of heart disease. These techniques include classification, regression, clustering, and association rule mining [6]. Each technique has its strengths and weaknesses, and the best technique for a particular problem will depend on the available data type.

There is a pressing need for better prediction models to identify those at the highest risk of heart disease so that early interventions can be targeted more effectively. This study seeks to develop and validate a suitable prediction model for heart diseases. We will use the CRISP-DM data mining framework to develop and compare a range of prediction models. We aim to find the model that best predicts heart disease risk so that it can be used to screen for the disease in high-risk populations. We will also look at the potential factors influencing the model's accuracy to attain the best possible model. The rest of this paper is organized as follows: in the next section, we will discuss the methodology used to select the best algorithm for our dataset. In Section 4, we will describe the data analysis procedure used for our experiments. Section 5 will present the results of our experiments, and in Section 6, we will discuss our conclusions and future work recommendations.

# **LITERATURE REVIEW**

Over the past few decades, there has been an increased interest in using data mining techniques for predictive modeling in medicine. Various studies have demonstrated the potential of data mining for heart disease prediction. This section will discuss some of the terms relevant to this discussion and review a selection of existing studies that have used data mining to predict heart disease

## **(i) Data Mining**

Data mining refers to the process of extracting valuable information from large data sets [7]. It is a relatively new field that has emerged in the past few years with the advent of big data. Data mining has many applications in business, medicine, science, and engineering.

## **(ii) Heart Disease**

Heart disease is a general term for various conditions that affect the heart. Heart disease is often used interchangeably with the term "cardiovascular disease." Cardiovascular disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain (angina), or stroke [8]. Other heart conditions, such as those that affect the heart's muscle, valves, or rhythm, also are considered forms of heart disease. The most common types of heart disease include Coronary artery disease, Heart failure, Arrhythmia, and heart valve disease. Heart disease is diagnosed using a series of tests. For instance, various imaging tests can be used to look for evidence of heart damage. These tests include a Cardiac CT scan, Cardiac MRI, Cardiac PET scan, Echocardiogram, and Nuclear stress test. Blood tests may also be used to look for markers of heart damage, such as troponin and CK-MB. Blood tests may help check the conditions causing heart diseases, such as high cholesterol, diabetes, and thyroid problems.

## **(iii) Classification Techniques**

Classification techniques are a crucial component of data mining. Classification is the task of correctly assigning new observations to one of a set of predefined classes. It is a supervised learning technique since the class labels of training data are used to build the classification model. Some popular classification algorithms include Naive Bayes, Decision Trees, and Support Vector Machines (SVM) [9]. The choice of the classifier will depend on the data's nature and the desired classification accuracy.

## **RELATED WORKS**

Ensemble classification is a machine learning technique that combines the predictions of multiple models to produce a more accurate prediction. Latha and Jeeva [10]developed a model using ensemble classification techniques to predict the chances of individuals contracting heart disease. An ensemble model produces a model with a higher prediction accuracy by reducing overfitting in the dataset. In their research, Latha and Jeeva used a Cleveland heart disease dataset from the UCI Machine Learning Repository to classify different algorithms. The dataset was grouped into a train and test set, with a Weka tool used to classify the dataset. They used the training dataset to train each classifier before testing the efficiency of each using the test dataset. In the model development process, the dataset is first cleaned before preprocessing it to identify missing or unwanted data. The data is then categorized using Bayes Net, PART, SVM, Multilayer Perceptron, Naive Bayes, and random forest (RF) classifiers. While classifying the algorithms, the weak classifiers are then boosted using ensemble techniques. The research used Boosting, Majority vote, Stacking, and Bagging ensemble techniques to improve the classifier outcomes. A comparison of the accuracies of the models was performed using the 10-fold cross-validation method, and the results were obtained.

From the results, it was found that weak classifiers produce better results after ensemble techniques are applied. To enhance the performance of the models, feature selection was applied, increasing the model's prediction accuracy further. Feature selection reduces overfitting by discarding the noisy and irrelevant features from the model. The study concluded that ensemble techniques effectively enhance the classifiers' performance and develop a model that can effectively predict the occurrence of cardiovascular disease. Ramalingam et al. [11] combined SVM, ANN, and KNN to build an ensemble model with higher predictive accuracy for heart disease diagnosis.

Ali et al. [12] proposed a heart disease prediction and diagnosis model using different types of supervised learning machine learning algorithms. A dataset from the Kaggle repository [13] containing 14 attributes was selected. The dataset consisted of 713 and 312 male and female records, all containing different ages. Also, from this dataset, some patients were in good health while others suffered from heart disease. The Weka tool was used for data mining, while the Python language was used to analyze the data. In the study, Ali et al. preprocessed the data to identify missing and invalid figures before applying the Interquartile Filter (IQF) filter to remove outliers. Outliers are abnormal values caused by errors in data entry or sensor malfunction. After this, the data was partitioned into six different sets where various classification algorithms were applied to compare the performance and accuracy of each. The classification algorithms comprised the k-neighbors (KNN), random forest (RF), multilayer perceptron (MP), AdaboostM1 (ABM1), logistic regression (LR), and decision tree (DT) were used to develop models, while the support vector machines algorithm (SVM) was used for feature selection. A comparison of the accuracies of the models was performed using the 10-fold cross-validation method, and the results were obtained. From the results, KNN, DT, and RF algorithms produced the highest accuracy (100%), giving the researchers the perfect model to predict and identify the risks of heart diseases.

In another study, Dangare and Apte [14] suggested a model to predict cardiovascular diseases using the classification technique based on historical heart disease occurrences. In their study, they categorized their 573 records dataset into three sets, key attribute (patient identification number), input attribute (description and values), and a predictable attribute (diagnosis) to develop a model for heart disease prediction. Three data mining techniques, comprising neural networks, decision trees, and Naive Bayes were used in the dataset analysis. Fifteen attributes associated with heart disease, including diabetes, obesity, cholesterol, and blood pressure, were included in the study as attributes to aid in heart disease prediction. Using the Weka tool, the heart disease dataset was preprocessed to replace missing values before being split into training (303 entries) and testing (270 entries) sets. To classify the dataset, they used a Multi-layer Perceptron Neural Networks (MLPNN) for their Neural Network model, a J48 decision tree algorithm, and a Naive Bayes classifier for their decision tree Naive Bayes model. Next, a confusion matrix was used to evaluate the performance of each model. The results showed that the neural networks model had the highest accuracy, followed by the Decision Tree model and Naive Bayes model. The research findings concluded that the Neural Network prediction model had the best performance in terms of accuracy for cardiovascular disease prediction.

After reviewing the existing literature on heart disease prediction, it is clear that there is still room for improvement in this area. This study seeks to build upon the work of previous researchers in order to create a more accurate and reliable prediction model. Different data mining algorithms will be applied to create a prediction model that is more accurate and reliable than existing models. This will help to improve the ability of medical professionals to identify individuals at risk for heart disease and ultimately help to save lives.

# **METHODOLOGY**

This study sought to comparatively analyze various data mining algorithms and their ability to predict heart disease. We used a qualitative approach to analyze the heart disease dataset while using the CRISP-DM data mining framework. This approach was appropriate for this type of research as it allowed for in-depth data exploration. A description of the dataset and its source was provided. The study's results were then used to access the most suitable algorithm for predicting heart disease. The step-by-step process of data acquisition, data preprocessing, model selection, and evaluation is described in the following sections.

## **Data Acquisition**

The research used a dataset consisting of demographic information and medical history from patients with heart disease. The dataset used in this study was obtained from the UCL Machine Learning Repository [15]. It consisted of 76 attributes and 303 instances, while only 14 attributes that were the most relevant for predicting heart disease were used in the study.

## **Data description**

The dataset used in this study were: age: a patient's age in years, sex:(male/female), cp:type of chest pain for a patient (1: normal angina, 2: irregular angina, 3: non-anginal pain, 4: asymptomatic type), trestbps: diastolic blood pressure for a patient at the time of hospital admission, chol:HDL/LDL cholestrol, fbs:glucose level, restecg: the heart’s electrical activity, thalach:maximum rate of heart beat, exang:chest pressure, ST depression: induced by exercise relative to rest, slope:the steepest point on the peak exercise ST segment, ca:blood vessels present, thal:heart rate at peak, and num:absence/presence of a heart disease.

## **Data Preprocessing**

Preprocessing is a critical step in data mining. It is used to clean the data, deal with missing values, and remove any noise. First, we will check for missing values from the dataset. Missing values can introduce bias and lead to inaccurate results [16]. Next, we normalized the values in the dataset to scale the range between 0 and 1. This was important because some of the attributes in the dataset (such as age) had a much larger range of values than others ( such as cholesterol level). Had the data not been normalized, the algorithms would have been biased towards attributes with a larger range of values. After cleaning and normalizing the data, the dataset was ready to be used in the data mining process. We saved it as a CSV file so that it could be easily imported into the Python programming language.

## **Exploratory Data Analysis**

EDA is essential in data mining to help us understand the data and get insights into it. In this study, we used EDA to understand the distribution of the data and the relationships between the various attributes [17]. First, we drew a heatmap to visualize the correlation between the attributes since some of the algorithms are sensitive to correlation. We also plotted density plots, bar charts and pie charts to help us understand the data and see if there were any trends that we could identify.

## **Training and Testing the Datasets**

Next, we split the dataset into two: a training and a testing set [18]. We used the sci-kit learn library in Python to split the dataset. The training set is used to train the model, while the testing set is used to assess the algorithm's performance. Separating the testing set would enable us to determine how well the model generalized to new data. If the model only performed well on the training set, it would likely be overfitting the data, whereas if it only performed well on the testing set, it would likely be underfitting the data. We split our dataset using a 70-30 ratio, where we would use 70% of the dataset for training. We would use 30% of the data for testing. This is a standard split ratio that would help us to get a reasonable estimate of the performance of the algorithms.

## **Implementation of Classification algorithms**

We then selected the three most suitable data mining algorithms for our study: support vector machines (SVMs), logistic regression (LR), and K neighbor classifiers (KNN). These algorithms are all commonly used for classification tasks and have been shown as the most effective in previous studies. Logistic regression is commonly used to predict the probability of a particular outcome. We used logistic regression to predict a patient's probability of developing heart disease. This algorithm would allow us to understand better the relationship between the various attributes and heart disease [19]. SVMs are best suited for binary classification tasks (such as our task of predicting heart disease) [20]. We used SVMs to classify whether or not a patient has heart disease since they can be more accurate than other classification algorithms. Moreover, SVMs can handle nonlinear data and are less likely to overfit the data. KNN works by finding the K nearest neighbors of a given data point and then classifying the data point based on the majority class of its neighbors [19]. This algorithm is simple to understand and implement. For logistic regression, we use the cleaned and normalized dataset to train the model. We then used the testing set to access the model’s performance. The process was then repeated for SVMs and KNN algorithms.

## **Model Evaluation**

We compared and contrasted the performances of the three models to evaluate their effectiveness by looking at the classification accuracy of each model. The classification accuracy score is the most commonly used metric for model evaluation. It simply measures the percentage of correct predictions made by the model. While it can be a useful metric, it does not give us any insight into the specific errors made by the model. A classification accuracy score of 1.0 would mean that the model correctly predicted all classes, while a score of 0.5 would mean that the model correctly predicted half of the classes [21]. The model with the highest classification accuracy score was considered the best suitable for predicting the risk of heart disease.

# **RESULTS**

The main aim of this research study was to analyze how we can use data mining techniques to predict patients with heart diseases. We used a qualitative approach to analyze the dataset. The various institutions can use the final model to predict patients likely to have or develop heart attacks. This chapter will summarize the results obtained after running the data analysis. The results are divided into several sub-sections; Exploratory data analysis, data preprocessing, modeling and evaluation.

## **Exploratory data analysis**

The data was first prepared using different methods before performing the analysis. The analysis was done using Jupyter Notebook using various libraries such as metrics, seaborn, LogisticRegression, Kmeans, SVM, Pandas, NumPy, and Matplotlib. These libraries were of great help while analyzing the dataset since they made work easier than using functions. First, the data was loaded using the Panda library. The data was in the form of a delimited dataset separated using commas. The variables in the dataset were either in categorical, continuous, or discrete forms. The dataset had 303 observations and 14 characteristics. The variables were in the form of integers and float.

The data were also checked for missing values. Fortunately, the data had no missing values. The variables were sorted into two categories; numeric variables and categorical variables. The numerical variables were first analyzed using density plots.

### **Numerical Analysis with Distplot**

**Age**

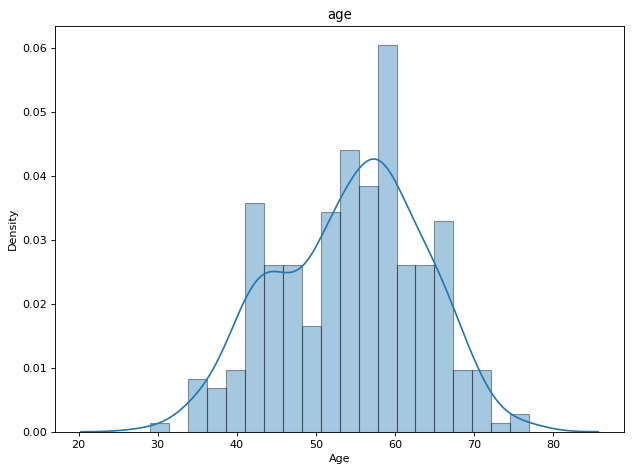


Figure 2:Age distribution among the respondents

The density plot above represents the distribution age of the respondents. From the graph above, we note that most respondents are between 50 and 60 years old.

**Blood pressure**

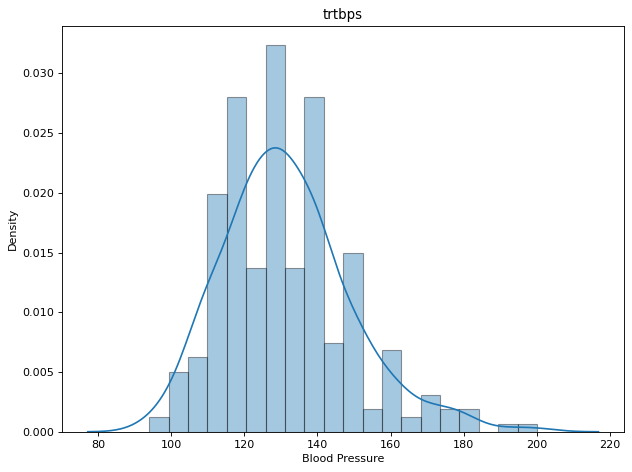


Figure 3: Blood pressure distribution among the respondents

The density plot above represents the distribution of blood pressure among the respondent. We can note that most respondent has a resting blood pressure between 110 and 140.

**Cholesterol**

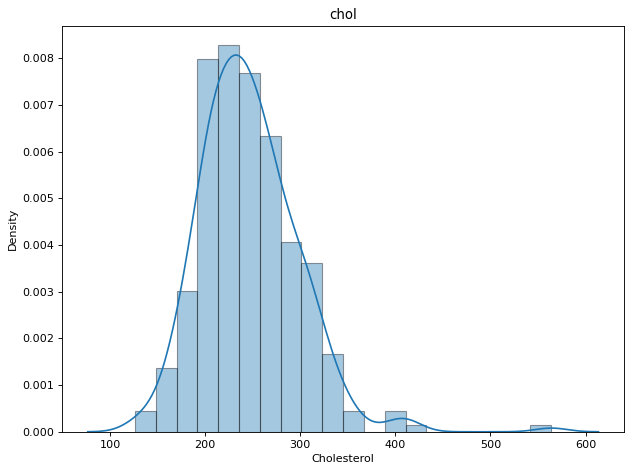


Figure 4: Cholesterol distribution among the respondents

The density plot above represents the distribution of cholesterol among the respondents. We note that most respondents have a cholesterol value of between 200 and 280.

**Maximum heart rate**

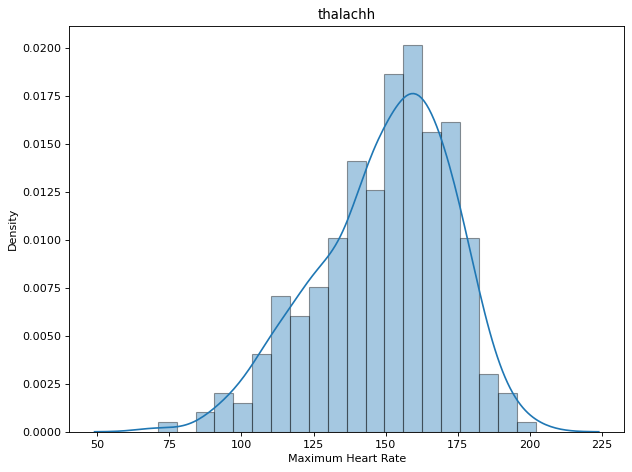


Figure 5: Maximum heart rate distribution among the respondents.

The density plot above represents a density distribution of maximum heart rate among the respondent. From the plot, we can note that most respondents achieved a maximum heart rate between 145 and 170.

**ST depression**

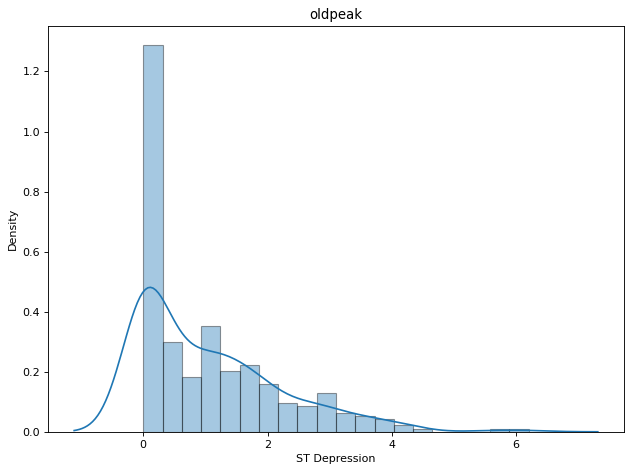


Figure 6: ST depression distribution among the respondents

The density plot above represents a density distribution of ST depression among the respondent. From the plot, we note that most respondents had an ST depression under 2.

### **Categorical Analysis with Pie Chart**

**Sex**

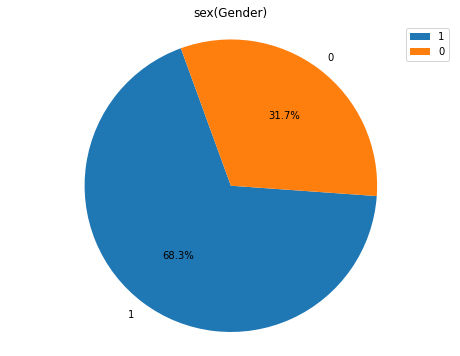


Figure 7: Gender of the respondent

The pie chart above shows the percentages of males and females among the respondent. The results show that 68.3% of the respondents were males and 31.7% were females.

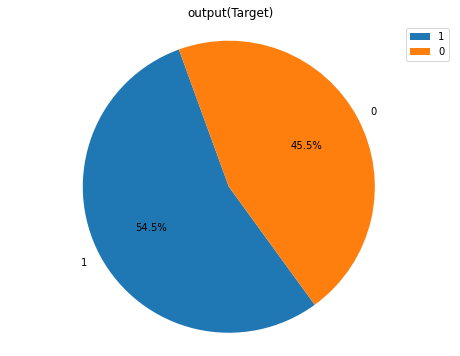


Figure 8: Respondent with heart diseases

The pie chart above shows the percentages of patients with heart disease and those without. From the chart, we can note that 54.5% have a heart attack risk and the remaining 45.5% have no heart attack risk.

**Categorical Analysis with Count Plot**

**Heart attack patients by sex**

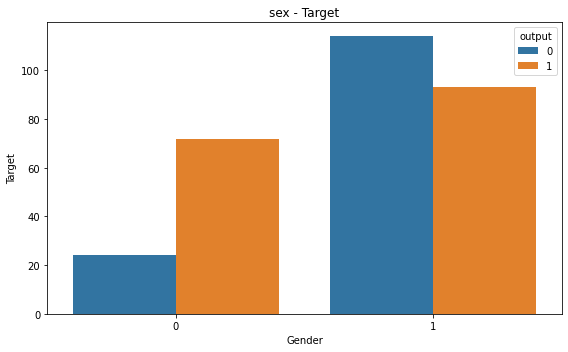


Figure 9: Heart attack patients by sex

The plot above represents the distribution of patients with and without heart attacks by gender. We note that males are more likely not to have a heart attack while females are at a higher risk.

**Heart attack patients by chest pains**

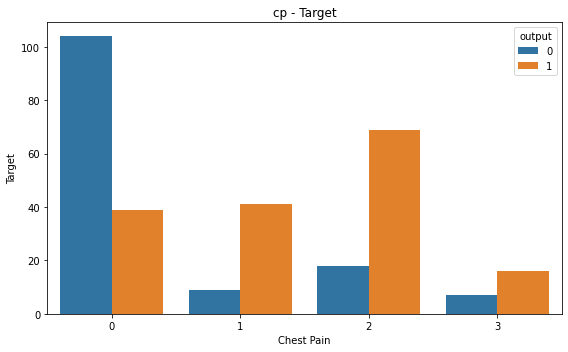


Figure 10: Heart attack patients by chest pains

From the plot above, we can note that patients without chest pains have less chance of heart attack, while patients with atypical angina-type chest pains have more chance of a heart attack.

Blood sugar

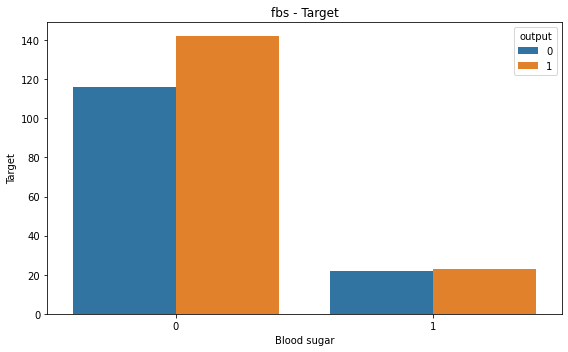


Figure 11: Blood sugar

From the above plot, we note that most patients with less than 120 mg/dl in fasting blood sugar have higher chances of a heart attack.

**Major vessels**

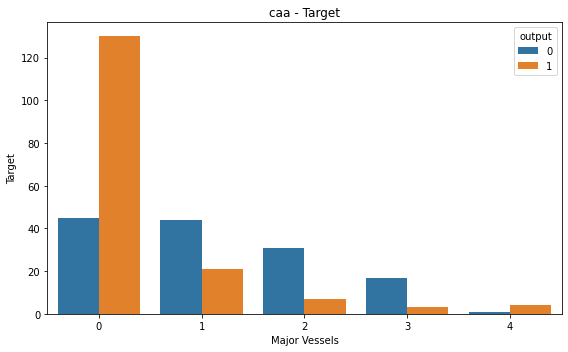


Figure 12: Major vessels

From the plot above, we note that patients with no major vessel have a higher chance of heart attack than patients with one or more major vessels.

**Maximum heart rate achieved**

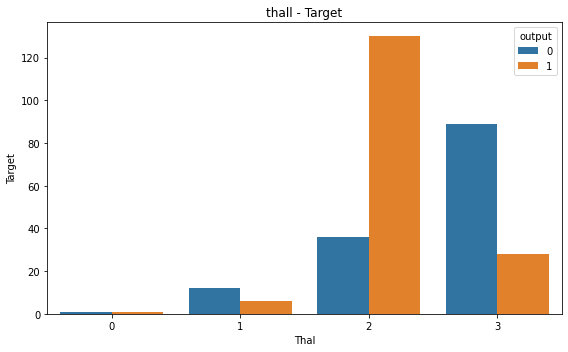


Figure 13: Maximum heart rate achieved

From the figure above, we note that patients with a value of thal 2 are three times more likely to have a heart attack than if they have not.

### **Correlation**

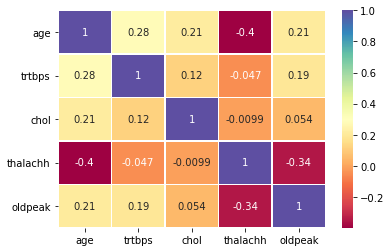


Figure 14: Correlation of the numerical variables

The heatmap above shows how the variables are related. From the results, we note that age has the highest negative correlation with “thalach”. However, no variable had a higher correlation that could affect our model.

## **Modeling**

Before splitting the dataset, it was first standardized and transformed. Secondly, the data was divided into training and testing datasets using a ratio of 70/30. We needed to predict whether a patient has a chance of having a heart attack using classification techniques. Therefore, we used three classification techniques; K-means algorithm, logistic regression, and Support vector machines (SVM). After running the algorithms, the k-means algorithm gave an accuracy of 0.7143, while logistic regression and support vector machines gave an accuracy of 0.8022 and 0.7912, respectively. Therefore, the best performing algorithm was logistic regression.

## **Results summary**

After running the analysis, we noted that most patients were between 50 and 60 years old. However, patients less than 25 years were less. We also noted that 68.3% of the respondents were males and 31.7% were females. Males are more likely not to have heart attacks, while females are at a higher risk. The analysis also revealed that patients without chest pains have less chance of heart attack, while patients with atypical angina chest pains have more chance of heart attack. Also, most patients have a cholesterol value of between 200 and 280, which can be considered high. It was also noted that patients without chest pains have less chance of heart attack, while patients with atypical angina chest pains have more chance of heart attack. In addition, most patients with less than 120 mg/dl in fasting blood sugar has higher chances of a heart attack.

# **DISCUSSION, CONCLUSION & FUTURE WORK**

This research's primary goal was to find the best model to help in heart disease prediction by comparing the performances of various data mining models. We obtained a dataset from the UCL Machine Learning Repository and used a subset of 14 attributes we found the most relevant for predicting heart disease. Next, we preprocessed the data before analyzing it to avoid any issues that could have arisen due to the differences in the distributions of the data. A series of plots were generated to observe the relationships between the different attributes and heart disease before proceeding to the modeling process. Once we obtained the results, we analyzed the models' performances and drew conclusions based on our findings.

Males dominated this research with a 68.3% majority, while females accounted for only 31.7% of the respondents. A significant number of respondents were aged between 50 to 60 years. Of these respondents, the study's findings found that females faced a higher risk of developing heart disease as they got older. In support of this, a different study aimed to determine the risk of males and females developing heart failure found that the risk was 1.71 times higher for females than males [22]. Among the different types of chest pains, patients with atypical angina-type chest pain posed more risk of developing heart diseases. This was unsurprising, given that atypical chest pain is often a symptom of coronary artery disease, which is a leading cause of heart diseases [23].

Also, most respondents obtained a maximum heart rate between 145 and 170 beats per minute, the standard heart rate for adults. If the patient achieved a thalach value of 2, they faced higher chances of getting heart disease than when they did not. Many patients' blood pressure levels were within the normal range (120/80 mmHg). However, some outliers with higher blood pressure levels were at a higher risk for heart disease [24]. Cholesterol levels ranging between 200 and 280 mg/dl were also found in most patients, although some patients had very high cholesterol levels. [25] recommends that cholesterol levels for healthy people should range below 200 mg/dl (total cholesterol), 130mg/dl (LDL), and 40mg/dl (HDL). Of the entire respondents, 54.5% suffered from heart disease. This could be attributed to high cholesterol and blood pressure, a leading cause of heart disease [26]. Patients without a major vessel also faced a higher risk of developing heart disease than those with one or more. Previous studies have found that blockages in the major vessels are often symptomatic of heart disease, and patients with blockages are more likely to have had a previous heart disease [27]. Patients with fasting blood sugar levels of less than 120mg/dl were at higher risk of developing heart diseases. In [28], it was found that patients with fasting blood sugar levels less than 100mg/dl were twice as likely to have heart disease.

We trained three classification models to predict heart disease: logistic regression, the KNN algorithm, and SVM. We then tested these models on our test set and found that the logistic regression model had the best performance, with an 80.22% accuracy.

In conclusion, we have performed a comparative analysis of different classification models to predict heart disease. We have found that logistic regression outperforms the other two models. Furthermore, we have also found that factors such as high blood pressure, cholesterol, and fasting blood sugar indicate a higher risk of developing heart disease.

There are a few limitations to our study. First, we did not have a large dataset to train our models on, which could have improved our models' performance. Second, we did not perform cross-validation, which could have helped us to avoid overfitting. Third, we only used three classification models, and there are many other models.

Future work could focus on increasing the size of the dataset and using more sophisticated models such as deep learning to improve the performance of the models. Additionally, it would be interesting to see if other factors such as lifestyle choices, family history, and diet could also help predict the risk of heart disease. Finally, we could also experiment with different classification models to see if we can improve the performance of our models.

# **References**

[1] M. Amin, Y. Chiam and K. Varathan, "Identification of significant features and data mining techniques in predicting heart disease", *Telematics and Informatics*, vol. 36, pp. 82-93, 2019. Available: 10.1016/j.tele.2018.11.007 [Accessed 18 July 2022].

[2] M. Malik, S. Abdallah and M. Ala'raj, "Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review", *Annals of Operations Research*, vol. 270, no. 1-2, pp. 287-312, 2016. Available: 10.1007/s10479-016-2393-z [Accessed 18 July 2022].

[3] D. Sharma, U. R. Saroj, A. Upadhyay and B. K. Singh, "AYURVEDIC REVIEW ON DIAGNOSIS AND MANAGEMENT OF HRIDROGA WITH THE SPECIAL REFERENCE TO CARDIOVASCULAR DISEASE", *AYUSHDHARA*, pp. 2719-2724, 2020. Available: 10.47070/ayushdhara.v7i3.570.

[4] W. Meijers and R. de Boer, "Common risk factors for heart failure and cancer", *Cardiovascular Research*, vol. 115, no. 5, pp. 844-853, 2019. Available: 10.1093/cvr/cvz035 [Accessed 18 July 2022].

[5] D. Zhou, B. Xi, M. Zhao, L. Wang and S. Veeranki, "Uncontrolled hypertension increases risk of all-cause and cardiovascular disease mortality in US adults: the NHANES III Linked Mortality Study", *Scientific Reports*, vol. 8, no. 1, 2018. Available: 10.1038/s41598-018-27377-2.

[6] A. Altamimi and M. Azzeh, "Utilizing Data Mining Classification Algorithms for Early Diagnosis of Heart Diseases", *Predicting Heart Failure*, pp. 155-170, 2022. Available: 10.1002/9781119813040.ch7 [Accessed 18 July 2022].

[7] S. Agarwal, "Data Mining: Data Mining Concepts and Techniques," 2013 International Conference on Machine Intelligence and Research Advancement, 2013, pp. 203-207, doi: 10.1109/ICMIRA.2013.45.

[8]WHO "Cardiovascular diseases (CVDs)", 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Accessed: 17- Jul- 2022].

[9] S. Gowtham and S. Karuppusamy, "Review of Data Mining Classification Techniques", *Bonfring International Journal of Software Engineering and Soft Computing*, vol. 9, no. 2, pp. 8-11, 2019. Available: 10.9756/bijsesc.9013.

[10] C. Latha and S. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019. Available: 10.1016/j.imu.2019.100203 [Accessed 17 July 2022].

[11] V. V. Ramalingam, A. Dandapath and M. Karthik Raja, "Heart disease prediction using machine learning techniques : a survey", *International Journal of Engineering &amp; Technology*, vol. 7, no. 28, p. 684, 2018. Available: 10.14419/ijet.v7i2.8.10557.

[12] M. Ali, B. Paul, K. Ahmed, F. Bui, J. Quinn and M. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison", *Computers in Biology and Medicine*, vol. 136, p. 104672, 2021. Available: 10.1016/j.compbiomed.2021.104672 [Accessed 17 July 2022].

[13] https://www.kaggle.com/johnsmith88/heart-disease-dataset [Accessed 17 July 2022].

[14]C. S.Dangare and S. S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", *International Journal of Computer Applications*, vol. 47, no. 10, pp. 44-48, 2012. Available: 10.5120/7228-0076.

[15] “UCI Machine Learning Repository: Heart Disease Data Set", *Archive.ics.uci.edu*, 2022. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease. [Accessed: 19- Jul- 2022].

[16] [1]R. Sanni and H. Guruprasad, "Analysis of performance metrics of heart failured patients using Python and machine learning algorithms", Global Transitions Proceedings, vol. 2, no. 2, pp. 234, 2021. Available: 10.1016/j.gltp.2021.08.028 [Accessed 19 July 2022].

[17] C. Wu, M. Badshah and V. Bhagwat, "Heart Disease Prediction Using Data Mining Techniques", *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, 2019. Available: 10.1145/3352411.3352413 [Accessed 19 July 2022].

[18] M. Uçar, M. Nour, H. Sindi and K. Polat, "The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets", *Mathematical Problems in Engineering*, vol. 2020, pp. 3-4, 2020. Available: 10.1155/2020/2836236 [Accessed 19 July 2022].

[19] M. T., D. Mukherji, N. Padalia and A. Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", *International Journal of Computer Applications*, vol. 68, no. 16, pp. 13, 2013. Available: 10.5120/11662-7250.

[20] M. Abdar, S. R. Niakan Kalhori, T. Sutikno, I. Much Ibnu Subroto and G. Arji, "Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases", *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 6, p. 1571, 2015. Available: 10.11591/ijece.v5i6.pp1569-1576.

[21] X. Gao, A. Amin Ali, H. Shaban Hassan and E. Anwar, "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method", *Complexity*, vol. 2021, pp. 4-5, 2021. Available: 10.1155/2021/6663455 [Accessed 19 July 2022].