Mathematical Modelling and Numerical Simulation with Applications, 2024, 4(2), 133–148

https://dergipark.org.tr/en/pub/mmnsa
ISSN Online: 2791-8564 / Open Access
https://doi.org/10.53391/mmnsa.1412304

RESEARCH PAPER

Analysis of patient data to explore cardiovascular risk factors

Jawaher Almushayqih [1,‡], **Abayomi Samuel Oke** [2,*,‡] **and Belindar Atieno Juma** [3,‡]

[1]Department of Mathematics, College of Science, Qassim University, 52571 Buraydah, Saudi Arabia,
[2]Department of Mathematics, Adekunle Ajasin University, 342111 Akungba Akoko, Nigeria,
[3]Department of Mathematics, University College London, WC1E6BT London, United Kingdom

*Corresponding Author
‡ j.j.almoshigah@qu.edu.sa (Jawaher Almushayqih); okeabayomisamuel@gmail.com (Abayomi Samuel Oke); belindar.juma.23@ucl.ac.uk (Belindar Atieno Juma)

Abstract

According to the World Health Organisation, cardiovascular diseases claim over 17.9 million lives yearly on a global scale. Hence, cardiovascular diseases are responsible for 32 percent of global deaths yearly. Furthermore, it is estimated that more than 50 percent of heart disease cases are only discovered after they have reached the critical stage of heart failure and stroke. However, early detection of these heart diseases can reduce the mortality rates of cardiovascular diseases. Scientists have suggested using machine learning algorithms to identify the risk factors. However, the unavailability of data has hindered the significant success of this approach. In this study, machine learning algorithms are used to identify the important features that should be monitored to prevent heart diseases by considering a dataset obtained from 1000 patients. The six machine learning algorithms used for this study are Logistic Regression, Support Vector Machine, k-nearest Neighbour, Decision Tree, Random Forest and Multi-layer Perception Classifier. The dataset consists of twelve features that are considered to be associated with heart disease and a target variable. The results from this study show that patients suffering from typical and atypical angina chest pain are prone to heart disease. Patients who exercise up the slope have a higher likelihood of living without heart disease. Among the six algorithms used, the MLP Multi-layer Perception Classifier outperforms all others by achieving a 99 percent accuracy. Moreover, the Random Forest algorithm follows with an accuracy of 98 percent.

Keywords: Machine learning algorithms; cardiovascular diseases; heart disease; risk factors

AMS 2020 Classification: 68T01; 92B20; 92C50

1 Introduction

According to the World Health Organisation, Cardiovascular diseases (CVDs) are the leading cause of death, responsible for 32 percent of deaths globally [1]. CVDs have also been estimated

➤ **Received:** 30.12.2023 ➤ **Revised:** 25.06.2024 ➤ **Accepted:** 30.06.2024 ➤ **Published:** 30.06.2024

to be responsible for nearly 40 percent of premature deaths due to non-communicable diseases. The fact that CVDs can be prevented by improving behavioral factors necessitates the impetus to explore patients' data to identify patterns, correlations, and potential risk factors associated with heart disease [2]. This present study analyses a dataset of 1000 patients to provide information on the behavioral activities that can prevent cardiovascular health challenges.

The human heart, located in front of the chest behind the ribcage, is responsible for blood circulation throughout the body, control of the rhythm and speed of the heart rate and the maintenance of blood pressure [3]. The heart is divided into 5 parts; the walls, the chambers, the valves, the blood vessels and the electrical conduction system [4]. The muscles that contract and relax to pump blood throughout the body are the heart walls; separated into the left and right halves by a layer of muscle called the septum. The heart wall consists of three layers; endocardium (inner layer), myocardium (middle layer made up of muscle) and epicardium (outer layer of protection). There are four chambers in the heart; right atrium, left atrium, right ventricle and left ventricle. The atrium is on the top part of the heart while the ventricles are on the lower part. The right atrium receives blood with low oxygen content and pumps the blood through the right ventricle to the lungs for oxygenation. The oxygenated blood is passed back into the heart through the left atrium and finally pumped through the left ventricle to other parts of the body. The passageways in the heart are called the heart valves [5, 6]. The valves are classified as atrioventricular valves and semilunar valves. The tricuspid valve (the valve connecting the right atrium and right ventricle) and the mitral valve (between the left atrium and the left ventricle) are the atrioventricular valves. The semilunar valves open when blood flows out of the ventricles, the aortic valve and the pulmonary valve. The blood vessels are vessels through which blood is pumped to and from other parts of the body and there are three types of blood vessels; arteries, veins and capillaries [7]. The arteries carry oxygenated blood from the heart to other body parts (except the pulmonary artery that carries deoxygenated blood to the lungs), the veins carry deoxygenated blood into the heart (except the pulmonary vein that carries oxygenated blood from the lungs to the heart) and the capillaries are small blood vessels where your body exchanges oxygen-rich and oxygen-poor blood. The electrical conduction system is responsible for the exchange of electrical signals and pulses within the heart [8, 9].

The human heart is susceptible to a myriad of conditions, with heart disease representing a prominent threat [10]. Heart diseases are diseases that affect the heart, ranging from diseases that affect the blood vessels and heartbeat rhythm to the heart muscle and heart valves. Symptoms associated with heart disease depend on the type of heart disease. A common heart disease is coronary artery disease in which blood flow to and from the heart is hindered. Symptoms include chest pains, breath shortness, neck pain, numbness and weakness in the legs and arms. Heart attack or failure, angina and stroke are usually the symptoms that bring the patients to the hospital for diagnosis. Arrhythmias, a distortion in the rhythm of the heartbeat, is another common heart disease whose symptoms include chest pain, fainting, chest fluttering breath shortness, and slow breath [11]. Heart valve diseases and other diseases often come with symptoms such as chest pains, dizziness, and breath shortness. According to Shah et al. [12], the survival rate among heart disease patients is low because the diagnosis of most cases is done after the heart disease has reached critical stages. Recognizing the significance of early detection and intervention, this study explores key attributes that may contribute to the presence or absence of heart disease.

Machine learning (ML) is an artificial intelligence technique that learns hidden patterns in a dataset, aiding more accurate prediction or classification for decision-making. Several algorithms have been developed to guide the learning process and formulate a reliable model for any given dataset but the machine learning algorithms are generally categorised as either supervised (for datasets with a target column) or unsupervised (for datasets without a target column). Supervised learning

algorithms include linear regression, logistic regression, decision trees, random forest, support vector machine, k-Nearest Neighbours, Naive Bayes and Neural Networks. Unsupervised learning algorithms include K-Means Clustering, Density-Based Spatial Clustering of Applications with Noise, Autoencoders, t-distributed Stochastic Neighbour Embedding, Association Rule Mining and Singular Value Decomposition (SVD). ML algorithms have been applied by several authors on the UCI dataset and authors have found seemingly conflicting results on the best classifier. Saboor et al. [13] used XGBoost, random forest, decision trees, support vector machine (SVM), multinomial Naïve Bayes, logistic regression, linear discriminant analysis, AdaBoost classifier, and extra trees classifier on the UCI dataset on heart disease. The performance of all the algorithms indicates that the Support vector machine outperforms all other algorithms. Ramesh et al. [14] also considered the dataset by including the use of k-Nearest Neighbour, Random Forest, Decision Tree, Logistic Regression, Naive Bayes and SVM. In their study, k-NN outperformed other models. According to Chang et al. [15], the Random Forest classifier outperforms other classifiers. Boukhatem et al. [16] also found SVM as the best classifier among others. The inconsistencies in the outcomes of the studies could be due to the few data available on the UCI dataset on heart disease patients, containing only 303 rows.

In this study, we identify the human features that point towards heart disease. This is achieved by using machine learning algorithms to classify the data and extract the significance of each feature in contributing to heart disease. This current study differs from existing literature in two ways. Firstly, studies from the literature utilised the dataset sourced from the UCI website that consists of data from 303 patients, but this study utilises the dataset from the Kaggle website that consists of data from 1000 patients. The use of a larger dataset provides us with the potential to offer a richer understanding of cardiovascular risk factors. Secondly, this study delves into the question of classifier selection, building upon the observations from prior studies that revealed a lack of consensus among authors regarding the optimal classifier. The significance of this study includes the identification of the optimal classifier for cardiovascular disease. Furthermore, this study provides a good pointer to the features that can reduce the chance of heart disease in any patient.

2 Methodology

Data source and features

The data on the cardiovascular disease dataset is downloaded from the Kaggle website (the data can be found on the link https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset). The dataset consists of 14 columns; column 1 for patients' identification number, columns 2 to 13 for the features, and column 14 for the target variable. The patient identification number is dropped from the dataset, leaving the 12 feature columns and 1 target column. The feature columns are age, gender, chestpain, restingBP, serumcholestrol, fastingbloodsugar, restingrelectro, maxheartrate, exerciseangia, oldpeak, slope and noofmajorvessels.

age: The age feature is a numeric data that represents the age of a patient.
gender: The gender is considered binary taking 1 for male and 0 for female.
chestpain: The chestpain variable is the type of chest pain experienced by the patients. The chest pains are classified into four with 0 representing typical angina, 1 representing atypical angina, 2 representing non-anginal pain and 3 representing asymptomatic pain.
restingBP: The restingBP is a numeric data type, ranging between 94 mmHg and 200 mmHg, that shows the blood pressure of the patients when they are resting.
serumcholesterol: The serumcholesterol is a numeric data type that specifies the level of cholesterol in the blood of the patients, typically ranging between 126 mg/dl and 564mg/dl.
fastingbloodsugar: The fastingbloodsugar is a numeric data type that represents blood sugar

Table 1. Descriptive statistics for patients with heart disease

	count	mean	std	min	25%	50%	75%	max
age	420	49.07	18.7	20	32	49	66	80
gender	420	0.76	0.43	0	1	1	1	1
chestpain	420	0.36	0.68	0	0	0	1	3
restingBP	420	134.77	26.56	94	122	130	142	200
serumcholestrol	420	281.06	87.56	132	230	270	345	465
fastingbloodsugar	420	0.13	0.34	0	0	0	0	1
restingrelectro	420	0.36	0.52	0	0	0	1	2
maxheartrate	420	136.31	39.26	71	103.75	134	171	202
exerciseangia	420	0.52	0.5	0	0	1	1	1
oldpeak	420	2.51	1.72	0	1.1	2.3	3.8	6.2
slope	420	0.6	0.55	0	0	1	1	2
noofmajorvessels	420	0.67	0.83	0	0	0	1	3
target	420	0	0	0	0	0	0	0

levels in the patients. It is divided into a binary mode where 0 represents blood sugar < 120 mg/dl and 1 represents blood sugar > 120 mg/dl.

restingrelectro: The restingrelectro is a nominal data type where 0 represents normal, 1 represents having ST-T wave abnormality, and 2 represents probable or definite left ventricular hypertrophy by Estes' criteria.

exerciseangia: The exerciseangia is binary data in which 0 indicates that there is no exercise-induced angina and 1 means angina is induced by exercise in the patient.

maxheartrate: The maxheartrate is a numeric data type between 71 and 202 for the maximum heart rate of the patient.

oldpeak: The oldpeak is a numeric value between 0 and 6.2 representing the ST depression induced by exercise relative to rest.

slope: The slope variable represents the type of slope on which exercises are carried out;1 for upslope, 2 for flat surface, and 3 for downslope.

noofmajorvessels: The noofmajorvessels is a nominal value; 0, 1, 2, 3.

target: The target is a numeric variable that indicates whether the patient has heart disease (value=1) or does not have heart disease (value=0).

Data description

Table 1 and Table 2 show the descriptive statistics for the patients with heart disease and without heart disease respectively. There are a total of 1000 patients out of which 420 are with heart disease and 580 are without heart disease. The mean age of patients with heart disease is 49.07 years while the mean age of the patients without heart disease is 49.37 years. This is an indication that patients who have heart disease are younger than the ones without heart disease. The mean chest pain for patients with heart disease is 0.36. This indicates that both typical and atypical angina chest pain are significant features signalling heart disease. The mean chest pain for the patients without heart disease is 1.43 and this indicates that non-angina chest pains are not an indication of heart disease. The analysis of the gender distribution is shown in Figure 1. The bar chart shows that there are more male patients with heart disease than there are females. It can also be seen that males visit the hospital more frequently to complain about heart-related problems.

Table 2. Descriptive statistics for patients without heart disease

	count	mean	std	min	25%	50%	75%	max
age	580	49.37	17.25	20	35	49	63	80
gender	580	0.77	0.42	0	1	1	1	1
chestpain	580	1.43	0.87	0	1	2	2	3
restingBP	580	164.04	26.04	94	143	168	187	200
serumcholestrol	580	333.45	153.5	0	241	351.5	456.25	602
fastingbloodsugar	580	0.41	0.49	0	0	0	1	1
restingrelectro	580	1.03	0.8	0	0	1	2	2
maxheartrate	580	152.12	28.22	96	133	152	175	202
exerciseangia	580	0.48	0.5	0	0	0	1	1
oldpeak	580	2.85	1.71	0	1.4	2.7	4.3	6.2
slope	580	2.22	0.65	1	2	2	3	3
noofmajorvessels	580	1.63	0.87	0	1	2	2	3
target	580	1	0	1	1	1	1	1

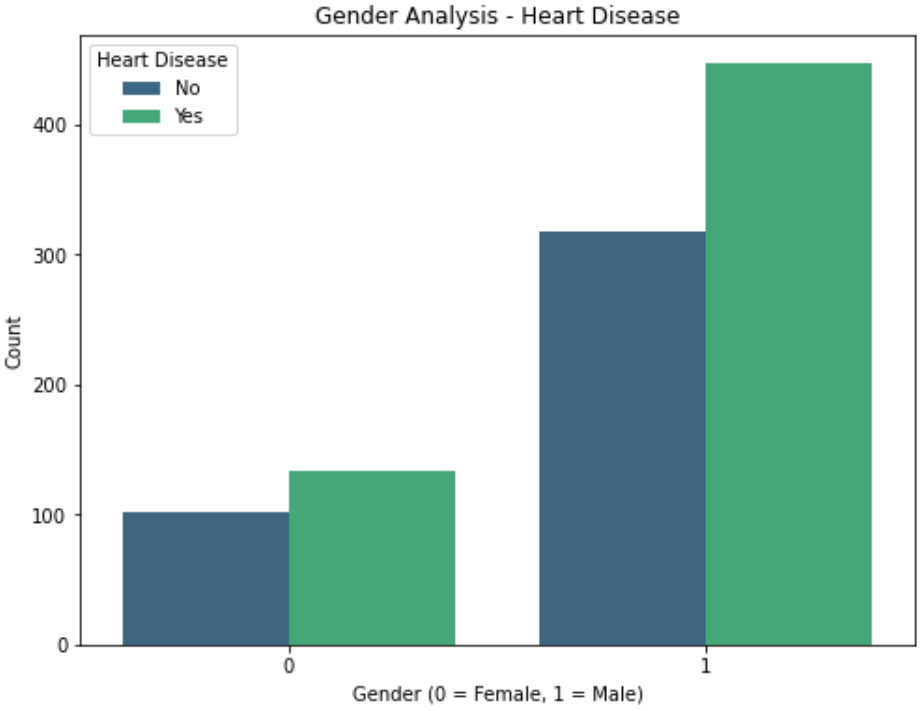


Figure 1. Gender analysis

Data exploration

The characteristics and patterns in the dataset are explored with the aid of a correlation matrix and Kernel Density Estimation (KDE) plots. It is important to note that correlation coefficients range from -1 to 1 (where -1 indicates perfect negative correlation, 0 represents no correlation, and 1 represents perfect positive correlation). Figure 2 shows the correlation coefficients and what they mean. The correlation matrix shows the correlation between all the variables in the dataset. The correlation matrix between all the data features and the target is shown in Figure 3. Figure 3 shows the matrix of the correlation coefficients for all features and target variables. The correlation coefficients for all features against the target variable are recorded in the last row (and also the column) of the correlation matrix. The slope has the strongest positive correlation coefficient of 0.80, followed by chest pain (0.55) and then the number of vessels (0.49) and resting blood

pressure (0.48) have a weak correlation with the target. The slope and chest pain show a very strong positive correlation with the target; indicating that heart disease in a patient is highly dependent on the choice of slope of the peak exercise and the nature of the chest pain.

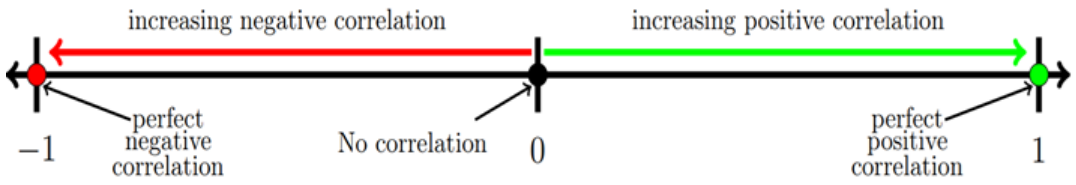


Figure 2. Correlation coefficients

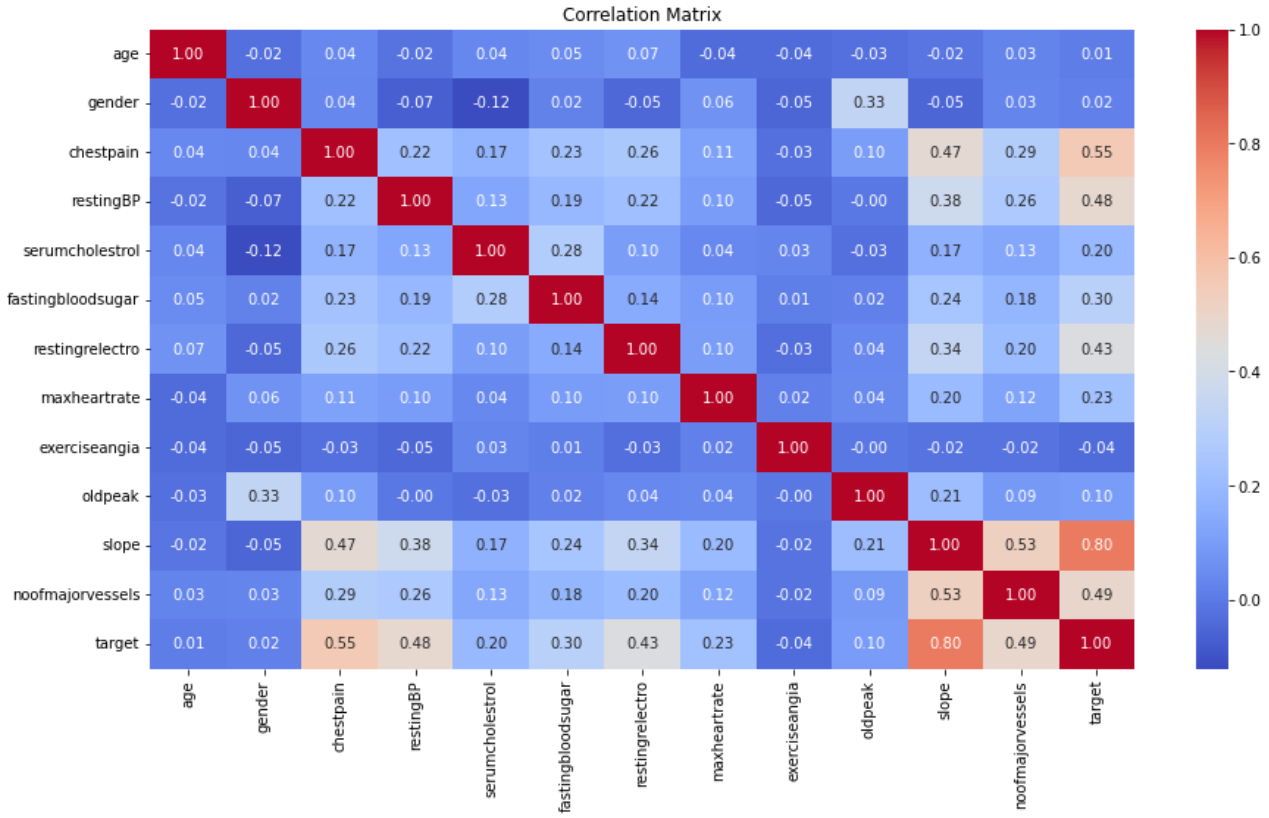


Figure 3. Correlation matrix

The KDE plot is used to estimate the probability distribution function of a dataset. By starting with the initial dataset, a smooth symmetric kernel is placed at each data point and the contribution of all kernels is summed up to create a continuous, smooth curve that estimates the probability density function. The results are normalised to ensure that the area under the curve is 1. In this case, the KDE plots for the two features that show a strong positive correlation with the target variable are displayed in Figure 4 and Figure 5. Figure 4 shows the distribution of slope with the density of the target. The slope is divided into 3 classes; 1 for upsloping, 2 for flat and 3 for downsloping. The upsloping represents the involvement of the patient in an exercise up a slope, the flat represents the involvement of the patient in an exercise on a flat surface, and The downsloping represents the involvement of the patient in an exercise down a slope. It is clear from the KDE plot that patients who are involved in exercises up the slope have a high tendency

of not developing heart diseases and the patients who are involved in exercises on a flat surface or down the slope are more likely to have heart disease than the ones who engage in exercises up a slope. Figure 5 shows the distribution of chest pain with the density of the target. For chest pain, the value 0 represents typical angina, value 1 represents atypical angina, value 2 represents nonanginal pain and value 3 represents asymptomatic chest pains. The concentration of patients with no disease is around 0 while the concentration of patients with diseases is around 2. The patients with angina chest pain are most likely not suffering from any heart disease while patients with non-anginal chest pain are most likely to suffer from heart disease.

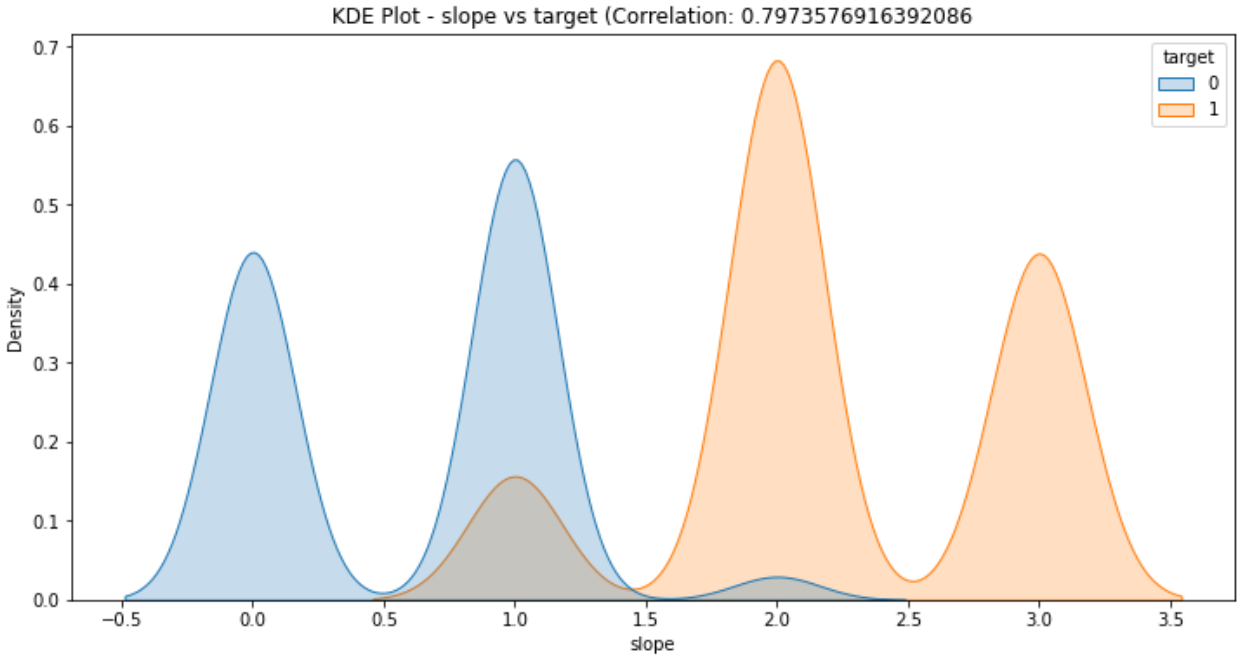


Figure 4. KDE plot for slope

Machine learning algorithms

Six machine learning algorithms used in training the data are Logistic Regression (LR), Support Vector Machine (SVM), *k*-Nearest Neighbour (*k*NN), Decision tree, Random Forest and Multi-layer Perception Classifier (MLP). The algorithms are discussed below. The Logistic Regression [17, 18] estimates the probability that an instance belongs to one of two categories, hence it is used for binary classification. A linear regression model for the target variable y is formulated as a linear function of the features x_k as

$$y = \beta_0 + \sum_{k=1}^{n} x_k. \tag{1}$$

The sigmoid function

$$P(Y = 1) = \frac{1}{1 + \exp\left(\beta_0 + \sum_{k=1}^{n} \beta_k x_k\right)}, \tag{2}$$

is adopted to ensure results stay in the interval (0,1). A decision line is used to separate the classes so that the instances above the line are classified into class 1 while the instances below are classified

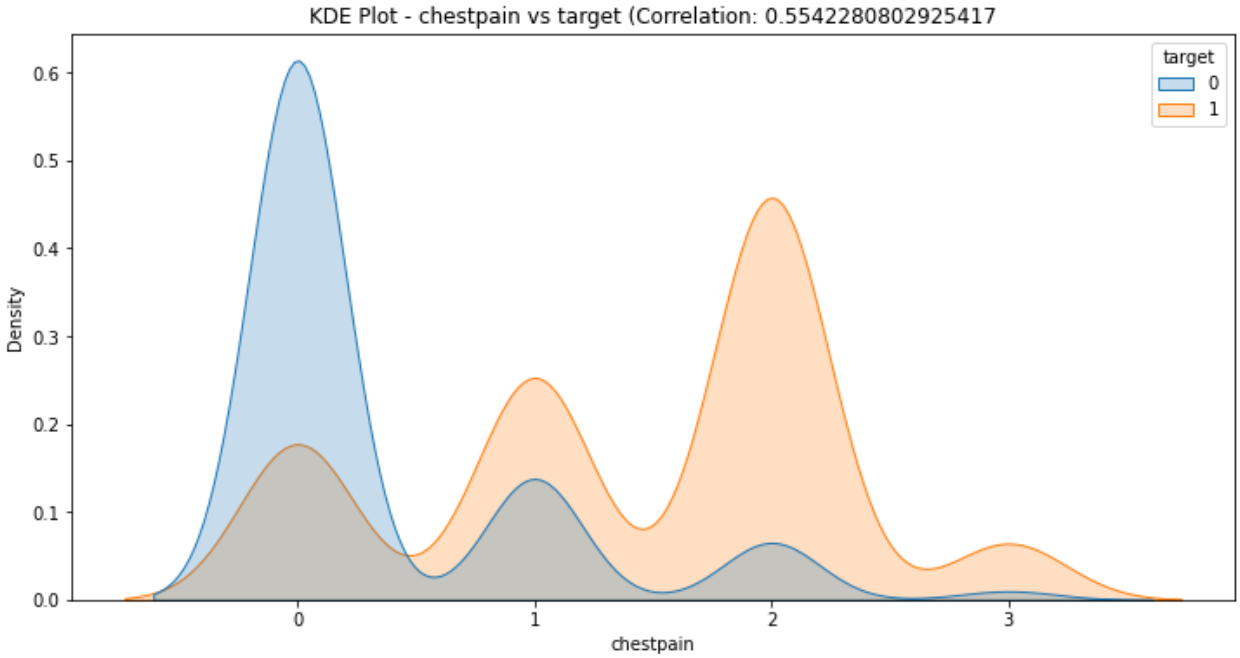


Figure 5. KDE plot for chestpain

into class 0. The SVM classifier [19] attempts to obtain the optimal hyperplane (also called the support vector) that best separates the classes in the feature space. The hyperplane is usually of dimension $n-1$ for an n-dimensional feature space, thereby serving as the decision boundary. For a binary classification, a decision function $f(x)$ is defined as

$$f(x) = \operatorname{sgn}(\boldsymbol{w} \cdot \boldsymbol{x} + b), \tag{3}$$

(where $\boldsymbol{w}, \boldsymbol{x}, b$ are the weight vector, feature vector and bias term) is used to determine the class for each instance. The k-nearest neighbour classifier [20] uses the Euclidean distance to measure the similarity between instances. The Euclidean distance is defined as

$$d_i = \left(\sum_{m=1}^{k} \|x_m - x_i\|^2 \right)^{\frac{1}{2}}. \tag{4}$$

The algorithm chooses the k-nearest neighbours to a certain instance and takes the frequency of their classes. The instance is allocated to the class with the highest frequency under the assumption that neighbouring instances have greater influences on each other. the choice of k, however, must be carefully made to avoid noise sensitivity or smoothing out local patterns. The Decision tree [21] algorithm starts by partitioning the dataset into subsets based on the significant attributes at each step. The tree starts with a root node, representing the entire dataset. The root node is partitioned into child nodes depending on the feature that provides the best separation. Next to the root node are the decision nodes, each representing a test condition on a specific feature. The predicted outcome is the leaf nodes, with each leaf corresponding to the class label for binary classification. Random Forest [22] uses a technique called bagging, which involves creating multiple subsets of the training dataset with replacement (bootstrap samples). Each subset is then used to train an individual Decision Tree. At each node of each Decision Tree, a random subset of features is considered for splitting. This randomness introduces diversity among the trees, leading to a

more robust ensemble. For classification tasks, the final prediction is determined by a majority vote among the trees. The Multi-layer Perceptron (MLP) [23] is a type of artificial neural network used for classification and regression. Its interconnected nodes are organized into an input layer, hidden layers, and an output layer. Each node in the network processes the input information using weights and biases. MLPs are trained to adjust the weights and biases in the model to minimize the difference between predicted and actual outcomes.

Performance metrics

The metrics used in evaluating the performance of the six algorithms are confusion matrix, precision, recall, F1-score and accuracy. The confusion matrix is a matrix of the form shown in Figure 6. The true positive is the number of classifications that were classified as 0 that truly belong to the class 0, false positive is the number of classifications that were classified as 0 that do not belong to class 0, false negative is the number of classifications that were classified as 1 that does not belong to class 1, true negative is the number of classifications that were classified as 1 that truly belong to class 1.

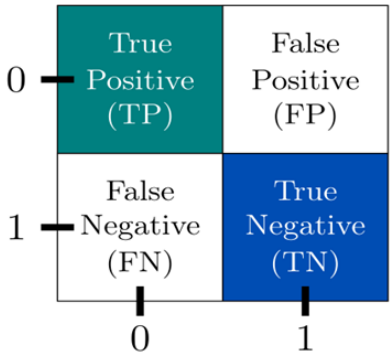


Figure 6. Confusion matrix general form

The confusion matrices provide a comprehensive view of the model's performance but precision, recall, F1-score, and support are derived metrics that offer more specific insights about the models. The precision gives an insight into the accuracy of a positive classification and is defined as the ratio of True Positives to the Total Positive Classification i.e.,

$$\text{precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}. \tag{5}$$

The recall measures the ability of the model to capture all positive instances and is defined as the ratio of true positive to the total positive

$$\text{recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}. \tag{6}$$

The F1 score gives a single metric that provides a balance between false positives and false negatives and is defined as

$$\text{F1 Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{7}$$

The accuracy of a model is the metric that measures the overall correctness of the model and is

defined as

$$\text{accuracy} = \frac{\text{True Positive} + \text{recall}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}. \tag{8}$$

In this study, all codes were executed using Spyder on Anaconda distribution. The machine used is a 64-bit operating system, x64-based processor, Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz 1.99 GHz HP laptop.

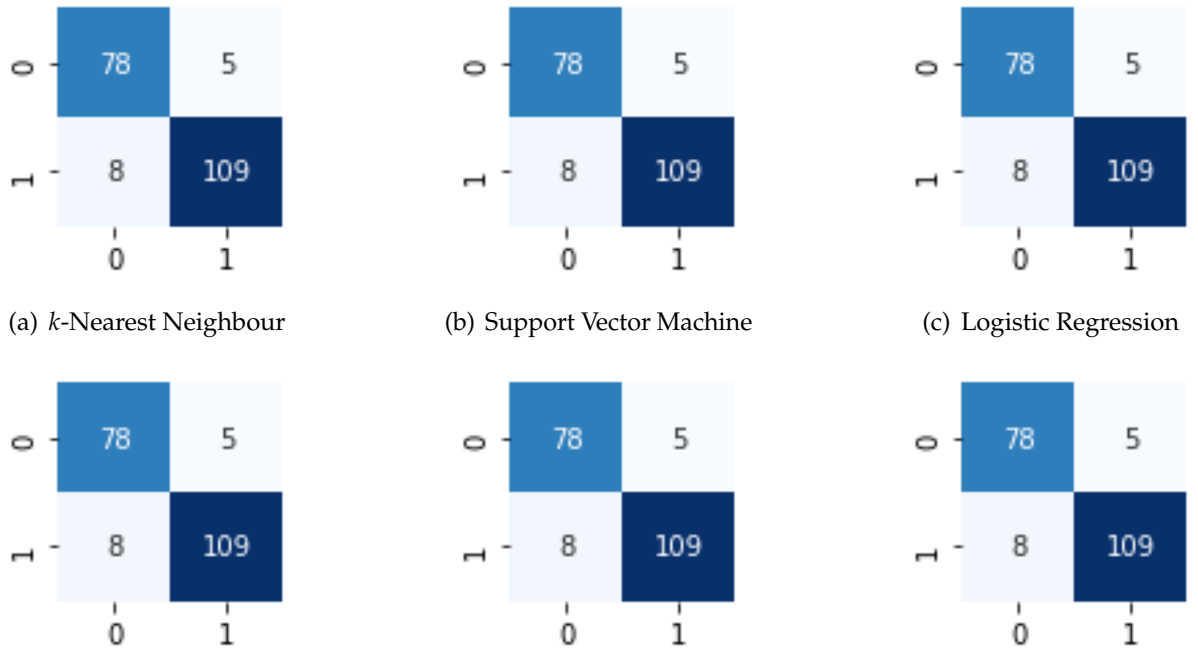
3 Analysis and discussion of results

The dataset was split into two; 80 percent for training and 20 percent for testing. The machine learning algorithms learned from the training dataset to build a model for Classification. By using the models to predict the classes of the testing dataset, the confusion matrix and other metrics can be used to evaluate each of the algorithms. The confusion matrices for all six algorithms are shown in Figure 7. All six models classify 83 instances into class 0 and 117 instances into class 1. The k-NN model classifies 78 instances accurately into class 0 and 109 instances correctly into class 1 while it wrongly classifies 5 instances into class 0 and 8 instances into class 1 (see Figure 7(a)). The SVM model classifies 79 instances accurately into class 0 and 113 instances correctly into class 1 while it wrongly classifies 4 instances into class 0 and 4 instances into class 1 (see Figure 7(b)). The LR model accurately classifies 79 instances into class 0 and 114 instances correctly into class 1 while it wrongly classifies 4 instances into class 0 and 3 instances into class 1 (see Figure 7(c)). The Decision Tree model classifies 78 instances accurately into class 0 and 115 instances correctly into class 1 while it wrongly classifies 5 instances into class 0 and 2 instances into class 1 (see Figure 7(d)). The RF model classifies 81 instances accurately into class 0 and 115 instances correctly into class 1 while it wrongly classifies 2 instances into class 0 and 2 instances into class 1 (see Figure 7(e)). The MLP model classifies 82 instances accurately into class 0 and 115 instances correctly into class 1 while it wrongly classifies 1 instance into class 0 and 2 instances into class 1 (see Figure 7(a)). Checking through the confusion matrices, MLP outperforms all the remaining 5 models while k-NN performs the least among all the models.

The superior performance of the Multi-layer Perceptron (MLP) Classifier in this study could be attributed to several factors. MLP is a type of neural network capable of learning complex, non-linear decision boundaries due to its multiple layers and non-linear activation functions. This property of MLP is particularly advantageous if the relationships in the data are not linearly separable. Also, MLP can automatically capture and model interactions between features which other algorithms (such as Logistic Regression) might require manual feature engineering to capture such interactions effectively. MLPs benefit from advanced optimization algorithms like Adam or RMSprop, which help in efficiently navigating the complex loss landscape and converging to a good solution.

Table 3 shows the precision, recall, F1-score, and accuracy of the six algorithms. MLP has 98 percent precision in classifying class 0 correctly, 99 percent precision in classifying class 1 correctly and 99 percent accuracy in any classification. RF has a precision of 98 percent in classifying into either class 0 or 1 and an accuracy of 98 percent in any classification. However, k-NN has 91 percent precision in classifying an instance into class 0, 96 percent precision in classifying an instance into class 1 and an accuracy of 94 percent in any classification. Hence, the MLP outperforms all the other algorithms and can therefore be used in this context to discuss the effects of the features on the chance of heart disease in any patient.

It is important to note that the correlation matrix indicates that the slope and chest pain are very significant in determining the chance of a patient having a heart disease. However, the outcome



(a) *k*-Nearest Neighbour (b) Support Vector Machine (c) Logistic Regression

(d) Decision tree (e) Random Forest (f) Multi-layer Perception Classifier

Figure 7. Confusion matrices for the six algorithms

Table 3. Performance metrics for the six models

		0	1	accuracy			0	1	accuracy
MLP	precision	98%	99%	99%	Random Forest	precision	98%	98%	98%
	recall	99%	98%	99%		recall	98%	98%	98%
	f1-score	98%	99%	99%		f1-score	98%	98%	98%
Decision Tree	precision	98%	96%	97%	Logistic Regression	precision	96%	97%	97%
	recall	94%	98%	97%		recall	95%	97%	97%
	f1-score	96%	97%	97%		f1-score	96%	97%	97%
SVC	precision	95%	97%	96%	KNN	precision	91%	96%	94%
	recall	95%	97%	96%		recall	94%	93%	94%
	f1-score	95%	97%	96%		f1-score	92%	94%	94%

only showed the correlation of a feature against the target without considering other features. A more detailed observation is carried out using the Random Forest to estimate the importance of each feature in determining the chance of heart disease in a patient. The Feature Importance is shown in Figure 8. The slope remains at the top of the chart, with the highest importance in determining whether a patient has heart disease or not. The resting blood pressure, old peak, serum cholesterol and chest pain are also significant in determining the heart condition of a patient.

4 Conclusion, recommendations and future research

Conclusion

Twelve features that are considered to be associated with heart disease were recorded for 1000 patients. Each patient was tested for any heart disease and the record is taken as 0 (if they have no heart disease) and 1 (if they have a heart disease). Out of the 1000 patients, 420 have heart disease and 580 do not have any heart disease. The data description shows that both typical

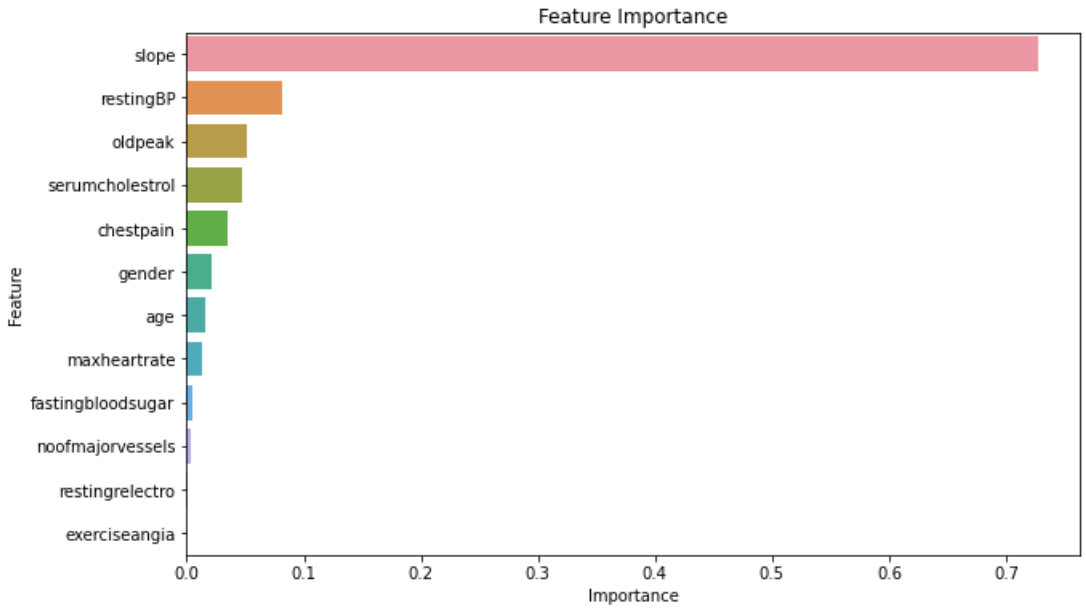


Figure 8. Feature importance

and atypical angina types of chest pain are features that are common among patients with heart diseases and hence, patients suffering from atypical angina chest pain are prone to heart disease and this agrees with the findings of Cubukcu et al. [24]. Furthermore, the male gender suffers from heart disease more than the females do. By drawing the correlation matrix, it is observed that the slope upon which a patient exercises is highly indicative of whether they will have heart disease or not. The Slope is found to have the strongest positive correlation coefficient of 0.80 with the heart disease condition of the patients, followed by chest pain (0.55). The KDE plot however showed that exercises up the slope is a good preventive measure for heart disease.

Machine learning classifiers are used to analyse the dataset to identify patterns and classify the instances. Performances of the classifiers were measured by using the confusion matrix, precision, recall, F1-score and accuracy. The results indicate the superiority of MLP over the other classifiers with an accuracy of 99 percent. The Random Forest algorithm follows with an accuracy of 98 percent. In furtherance to the analytics, the feature importance is estimated using the Random Forest. The result indicates that the slope upon which the patient carries out exercises is of the highest importance in determining whether a patient has heart disease or not. The resting blood pressure, old peak, serum cholesterol and chest pain are also significant in determining the heart condition of a patient.

Practical recommendations for patients and healthcare providers

The insights from this study highlight several key features associated with heart disease. Practical recommendations for patients and healthcare providers include:

- Regular Monitoring and Screening: The study showed that resting BP, serum cholesterol, and fasting blood sugar are significant factors in heart disease. Patients should regularly monitor their blood pressure, cholesterol levels, and blood sugar. Healthcare providers should prioritise these screenings during routine check-ups, especially for high-risk groups such as older adults.
- Chest Pain Evaluation: The study identified chest pain as a significant feature associated with heart disease. Any form of chest pain, particularly atypical angina, should be promptly evaluated by healthcare providers. Early detection and management of chest pain can prevent the progression to heart disease.

- Exercise Recommendations: The slope of exercise was found to be the most critical factor in determining heart disease. Engaging in regular physical activity, especially exercises that involve varying slopes, can be beneficial. Providers should encourage patients to incorporate such exercises into their routines as they are shown to be good preventive measures against heart disease.
- Gender-Specific Approaches: The study found that males suffer from heart disease more than females. Given that males are more prone to heart disease than females in this study, tailored intervention programs should be developed to address specific risk factors prevalent in men.

Implications for public health initiatives

The findings from the study have several implications for public health initiatives aimed at preventing cardiovascular disease:

- Targeted Screening Programs: Resting BP, serum cholesterol, and chest pain were significant factors identified. Public health campaigns should promote regular health screenings for high-risk individuals, focusing on monitoring blood pressure, cholesterol, and chest pain.
- Health Education and Awareness: The study highlighted the importance of chest pain and exercise slope. Increase public awareness about the importance of recognizing symptoms such as atypical angina and understanding their risks. Educational campaigns can inform the public about the significance of regular exercise, particularly on varied slopes, as a preventive measure.
- Promotion of Physical Activity: The slope of exercise was found to be highly indicative of heart disease presence. Encourage communities to create and maintain spaces where people can engage in physical activities that include varied terrain to promote heart health. Public health initiatives can include organized exercise programs that emphasize the benefits of slope exercises.

Key behavioural activities

The study identified a lack of regular exercise as a key risk factor for heart disease. The slope of exercise was identified as the most critical factor. The findings underscore the importance of regular physical activity, especially exercises involving varied slopes, as a preventive measure against heart disease. Inactivity or insufficient exercise can increase the risk. By addressing these key behavioural activities, both patients and healthcare providers can better manage and mitigate the risks associated with heart disease.

Future direction

The risk factors identified in this study by considering the demographic information (age and gender), clinical information (resting BP, serum cholesterol, fasting blood sugar, maximum heart rate and old peak), symptom information (chest pain and exercise angina) and diagnostic test results (resting relectro, slope and the number of major vessels). Further research is required to include the Lifestyle Factors (Diet, Physical activity level, Smoking status, and Alcohol consumption), Genetic Factors (Family history of heart disease and Genetic predispositions), Environmental Factors (Air quality exposure, Noise pollution exposure and Socioeconomic status), Psychological Factors (Stress levels, Mental health status and Sleep quality) and Healthcare Access (Frequency of medical check-ups, Accessibility to healthcare facilities and Health insurance status). Including these factors will increase the reliability of the outcomes.

Declarations

Use of AI tools

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Data availability statement

The dataset used for this study was obtained from Kaggle website:
https://www.kaggle.com/datasets/jocelyndumlao/cardiovascular-disease-dataset

Ethical approval (optional)

Not applicable

Consent for publication

Not applicable

Conflicts of interest

The authors declare that they have no conflict of interest.

Funding

No funding was received for this research.

Author's contributions

J.A.: Conceptualization, Methodology, Data Curation, Writing - Original Draft, Writing - Review & Editing. A.S.O.: Methodology, Software, Validation, Writing - Review & Editing, Visualization, Supervision. B.A.J.: Validation, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Visualization. All authors discussed the results and contributed to the final manuscript.

Acknowledgements

All authors want to show thankfulness to each contribution for accomplishing this research work.

References

[1] WHO, Cardiovascular diseases, (2023). https://www.who.int/health-topics/cardiovascular-diseases.

[2] Allen, L.A., Stevenson, L.W., Grady, K.L., Goldstein, N.E., Matlock, D.D., Arnold, R.M. et al. Decision making in advanced heart failure: a scientific statement from the American Heart Association. *Circulation*, 125(15), 1928–1952, (2012). [CrossRef]

[3] Mori, S., Tretter, J.T., Spicer, D.E., Bolender, D.L. and Anderson, R.H. What is the real cardiac anatomy? *Clinical Anatomy*, 32(3), 288–309, (2019). [CrossRef]

[4] Buijtendijk, M.F.J., Barnett, P. and Van Den Hoff, M.J.B. Development of the human heart. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 184(1), 7-22, (2020). [CrossRef]

[5] Gumpangseth, T., Mahakkanukrauh, P. and Das, S. Gross age-related changes and diseases in human heart valves. *Anatomy & Cell Biology*, 52(1), 25-33, (2019). [CrossRef]

[6] Gumpangseth, T., Lekawanvijit, S. and Mahakkanukrauh, P. Histological assessment of the

human heart valves and its relationship with age. *Anatomy & Cell Biology,* 53(3), 261–271, (2020). [CrossRef]

[7] Niklason, L.E. and Lawson, J.H. Bioengineered human blood vessels. *Science,* 370(6513), (2020). [CrossRef]

[8] Padala, S.K., Cabrera, J.A. and Ellenbogen, K.A. Anatomy of the cardiac conduction system. *Pacing and Clinical Electrophysiology,* 44(1), 15–25, (2021). [CrossRef]

[9] Hochman-Mendez, C., Mesquita, F.C.P., Morrissey, J., Da Costa, E.C., Hulsmann, J., Tang-Quan, K. et al. Restoring anatomical complexity of a left ventricle wall as a step toward bioengineering a human heart with human induced pluripotent stem cell-derived cardiac cells. *Acta Biomaterialia,* 141, 48–58, (2022). [CrossRef]

[10] Khan, M.A.B., Hashim, M.J., Mustafa, H., Baniyas, M.Y., Al Suwaidi, S.K.B.M., AlKatheeri, R. et al. Global epidemiology of ischemic heart disease: results from the global burden of disease study. *Cureus,* 12(7), (2020). [CrossRef]

[11] Khairy, P. Arrhythmias in adults with congenital heart disease: what the practicing cardiologist needs to know. *Canadian Journal of Cardiology,* 35(12), 1698–1707, (2019). [CrossRef]

[12] Shah, D., Patel, S. and Bharti, S.K. Heart disease prediction using machine learning techniques. *SN Computer Science,* 1, 345, (2020). [CrossRef]

[13] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M.F. and Ullah, N. A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems,* 2022(1), 1410169, (2022). [CrossRef]

[14] Ramesh, T.R., Lilhore, U.K., Poongodi, M., Simaiya, S., Kaur, A. and Hamdi, M. Predictive analysis of heart diseases with machine learning approaches. *Malaysian Journal of Computer Science,* 132–148, (2022). [CrossRef]

[15] Chang, V., Bhavani, V.R., Xu, A.Q. and Hossain, M.A. An artificial intelligence model for heart disease detection using machine learning algorithms. *Healthcare Analytics,* 2, 100016, (2022). [CrossRef]

[16] Boukhatem, C., Youssef, H.Y. and Nassif, A.B. Heart disease prediction using machine learning. In Proceedings, *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, pp. 1-6, Dubai, United Arab Emirates, (2022, February). [CrossRef]

[17] Ahmadini, A.A.H. A novel technique for parameter estimation in intuitionistic fuzzy logistic regression model. *Ain Shams Engineering Journal,* 13(1), 101518, (2022). [CrossRef]

[18] José R. Berrendero, Beatriz Bueno-Larraz, and Antonio Cuevas. On functional logistic regression: some conceptual issues. *Test,* 32, 321-349, (2023). [CrossRef]

[19] Joshi, A.V. Support vector machines. In *Machine Learning and Artificial Intelligence* (pp. 89–99). Cham, Switzerland: Springer International Publishing, (2023). [CrossRef]

[20] Nino-Adan, I., Landa-Torres, I., Portillo, E. and Manjarres, D. Influence of statistical feature normalisation methods on K-Nearest Neighbours and K-Means in the context of industry 4.0. *Engineering Applications of Artificial Intelligence,* 111, 104807, (2022). [CrossRef]

[21] Meng, L., Bai, B., Zhang, W., Liu, L. and Zhang, C. Research on a decision tree classification algorithm based on granular matrices. *Electronics,* 12(21), 4470, (2023). [CrossRef]

[22] Bai, J., Li, Y., Li, J., Yang, X., Jiang, Y. and Xia, S.T. Multinomial random forest. *Pattern Recognition,* 122, 108331, (2022). [CrossRef]

[23] Al Bataineh, A., Kaur, D. and Jalali, S.M.J. Multi-layer perceptron training optimization using

nature inspired computing. *IEEE Access*, 10, 36963–36977, (2022). [CrossRef]

[24] Cubukcu, A., Murray, I. and Anderson, S. What's the risk? Assessment of patients with stable chest pain. *Echo Research & Practice*, 2(2), 41–48, (2015). [CrossRef]

Mathematical Modelling and Numerical Simulation with Applications (MMNSA)
(https://dergipark.org.tr/en/pub/mmnsa)

Copyright: © 2024 by the authors. This work is licensed under a Creative Commons Attribution 4.0 (CC BY) International License. The authors retain ownership of the copyright for their article, but they allow anyone to download, reuse, reprint, modify, distribute, and/or copy articles in MMNSA, so long as the original authors and source are credited. To see the complete license contents, please visit (http://creativecommons.org/licenses/by/4.0/).

How to cite this article: Almushayqih, J., Oke, A.S. & Juma, B.A. (2024). Analysis of patient data to explore cardiovascular risk factors. *Mathematical Modelling and Numerical Simulation with Applications*, 4(2), 133-148. https://doi.org/10.53391/mmnsa.1412304