Vol. XVII & Issue No. 12 December - 2024
INDUSTRIAL ENGINEERING JOURNAL

HEART DISEASE PREDICTION MODEL BASED ON MACHINE LEARNING

Deepti Kulkarni
Research scholar, Department of Computer Science, Oriental University, Indore, India
Email: deeptikulkarni018@gmail.com

Rashmi Soni
Research Supervisor & Associate Professor CSE department Oriental University ,Indore, India
Email: drrashmicseofficial@gmail.com

Abstract

Machine learning is a versatile field in the current scenario. This research paper analyzes various machine learning algorithms and implement a heart disease prediction model through machine learning. Through this paper, it can be analyzed that a high heartbeat can affect various organs and indicates symptoms of the disease. The objective is achieved by various algorithms like Decision trees, logistics, and KNN. The algorithm is selected as per the accuracy. This paper explained how the different factors of dataset: smoking, BMI, and cholesterol can affect heart rate.

Keywords: K-nearest neighbor (KNN), Body mass index (BMI), Cholesterol (totchol), Support Vector machine (SVM), Machine learning (ML), heart rate (HR), Deep Neural network (DNN).

INTRODUCTION

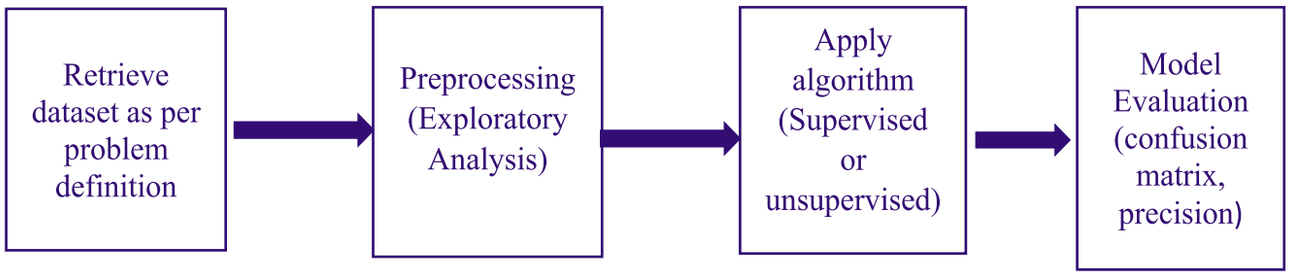
In the present scenario, it is observed that food habits and stress may cause severe health problems. Through this research paper, we proposed a model for the heart disease prediction system. Which easily identifies the abnormal behavior of heart rate. The machine learning model helps us to make decisions. Various algorithms are defined in machine learning to implement the task such as linear regression, logistic regression, K means cluster, random forest, and decision tree. In this research paper logistics, decision tree, and KNN algorithms are applied for implementing the heart disease prediction model. According to a comparative analysis by AUC & ROC curve, the algorithm which is giving higher accuracy is selected for creating a model. This research paper is categorized into three categories: section I: Literature review which includes various research theories worked in this direction Section II: machine learning algorithm description and Section III: proposed methodology, which includes proposed algorithms for preparing the model.

SECTION I: LITERATURE REVIEW

Malcolm Arnold et al. [1], and Palatini P et al. [2] explained that a high heart rate causes cardiovascular risk. The author [1] also describes if the heart resting period is high then there is a strong possibility of cardiovascular disease. The author [1] implement the model through machine learning algorithm: Logistic, Random Forest, DNN, Astral. Deep neural network (DNN) accuracy is more than other machine learning algorithm. Konstantinos et al. [3] high heart rate may cause various diseases like authorities, hypertension, organ damage (kidney, brain), stiff vessel, and aging. The author explains about heart rate is corelated with life expectancy through factors like metabolic rate and genetic influances. N. Peer et al. [4] describe how heart rate is associated with cardiovascular diseases. The author creates a model on parameters like age, gender, urbanization, alcohol use, smoking habits, glucose, blood pressure, and cholesterol. The author has applied a linear regression algorithm. The author defines a heart rate comparative chart for men and women where the graph indicates that between the age of 25-34 HR is 60-65 for men and 75-72 for women. At the age of 35-44 HR is 65-70 for men and 72-70 for women. At the age of 45-54 HR is 66-70 for men and 68-70 for women. For 55-64 age HR is 68-70 for men and 70-72 for women. Scott Reule et al. [5] explain that a high heart rate increases blood pressure. Blood pressure impacts hypertension which causes cardiovascular disease. The parameters are depending on each other and occur long-term loss. Yamaha Sandhya [6] explains heart diseases through a support vector machine. The author also explains about what the factors or symptoms of heart disease. This model helps to detect heart disease. Joon Nyung Heo et al. [7] implemented a machine Learning-Based Model for Acute Stroke. The author creates a machine learning model through a Deep neural network, random forest, and logistic regression. The author also did a comparative analysis of these three algorithms. A deep neural network will give appropriate results. Batta Mahesh [8] explains supervised learning and unsupervised learning algorithms. A supervised learning algorithm should be applied if the data is in fewer amounts and clearly labelled data. Unsupervised learning is applied to a huge amount of data for better performance. The author gives a review on supervised & unsupervised learning. Ibrahim et al. [9] explained the ML algorithms are applied for predicting and diagnose the disease. The author explained comparative analysis of various ML algorithm with their pros

and cons, also explained which ML algorithm has been applied for which disease (like covid, skin-cancer, thyroid, lung-cancer, heart) with their accuracy rate. Akinsola [10] explained various supervised learning algorithms compared with their features, precision and accuracy rate. These algorithm was applied on the diabetes dataset. SVM algorithm was performed very well. shahadat et al. [11] compare different supervised ML algorithms for disease detection. The study was focused on the performance evaluation of algorithms on the basis of key points for comparison to detect disease. Machine learning is an approach that provides various algorithms defined to create a model. In machine learning, various steps are involved to create a model. The problem statement contains some independent and dependent variables. One can validate and analyze the data through a dataset. which can be preprocessed prior then an appropriate algorithm will be selected. The data can be analyzed through graphical representation.

Figure 1 Work-flow for building model



SECTION II: CASE STUDY FOR HOW TO SELECT APPROPRIATE MACHINE LEARNING ALGORITHM

I) Linear regression model is defining the relationship between independent and dependent variables. Linear regression is a type of supervised learning which is divided into two categories: regression and classification. Regression type of problems predicts the continuous values. Classification problems predict the discrete values. In-house prediction, for example, the price will be dependent upon the area in square ft. We can draw the best fit line that can better estimate house prices. In linear regression, the best fit line will be selected as per minimum distances at each point. The whole concept is based on the equation y=mx+c. The value of y is dependent upon the x, m, and c values. If the value of these variables will be increased, then the value of y also will be increased. If outliers exist in the linear regression problem, then the line will try to fit every point. In this case, there is a chance of a false prediction. In a Linear regression-based problem one variable is independent and another is dependent. In that case, this algorithm defines the best results for the house prediction type of problem. In heartbeat prediction, multiple columns are defined. Heart rate normal and abnormal condition is based on multiple column values. In this case, we are not considering a Linear regression algorithm for the heart prediction problem.

II) Random forest is supervised learning used for classification & regression-based problems. Random forest is an ensemble type of technique that can combine multiple models. Ensemble uses two methods Bagging and Boosting. A random forest classifier is a set of decision trees from a randomly selected subset of the training set. It aggregates the nodes from different decision trees to decide the final class of the output. A random forest algorithm is used for credit card fraud detection, customer segmentation, and breast cancer detection. A random forest with one tree will overfit the problem. Another reason for overfitting is if a small data set is having various attributes. Due to overfit issue in small dataset it is not used for heartrate dataset.

III) Support Vector Machine is a supervised learning algorithm used for classified and regression analysis types of problems. The objective of SVM is to define the hyperplane for n-dimensional space that classifies the data points. The dimension of the hyperplane will be analyzed by several features. The selection of the best hyperplane will be the largest separation of lines between the two classes. If there is an outlier of one class exists in another class, then the SVM algorithm ignores the outlier and finds the maximum margin for the hyperplane. This algorithm can be best suitable for face detection, image classification, and handwriting detection.

Comparative analysis of machine learning algorithms

Table 1. Comparison of ML algorithms

Algorithm	Usage	Advantages	disadvantage
Linear regression	Simple algorithm	It is used for regression-based problems.	Cannot be used for classification-based problems.
Logistic Regression	Simple algorithm. Easy to implement.	It is used for classification-based problems.	It is not suitable for Multiple class-based problems. [11] Accuracy is not good for input variables that have a complex relationship.

Knn	Easy to implement	Gives appropriate results for Classification with multiple class-based problems and regression.	It does not work appropriately for large datasets.
Decision tree	Tree-like structure, fewer chances of error because the condition is checked at every level.	Gives appropriate results for Classification with multiple class-based problems and regression. It performed well for [10] discrete, continuous, binary attributes.	It is not suitable for large datasets. Overfitting may occur.
Random Forest	It is used for complicated problems like credit card detection or health care disease detection.	Gives appropriate results for Classification with multiple class-based problems and regression. It solves the overfitting problem of the decision tree.	A large no of trees can make the algorithm slow, and complexity may increase. It needs more time for training.
SVM	It predicts the [9] face detection type of problem very accurately.	Gives appropriate results for Classification with multiple class-based problems. The Probability of overfitting is less	SVM is not suitable for large datasets and [11] noisy datasets.

Data should be split into test and training data. Model accuracy will be analyzed by confusion matrix, accuracy score, and precision. If the model accuracy range is 85% to 95% or above, it indicates that the model is working with accuracy. The dataset should be taken from GitHub. In the dataset, there are 15 variables. Heart rate values can be dependent on various attributes like diabetes, prevalent stroke, Total chol, and current stroke. HR depends upon numerous factors. In figure 4.5 HR is much affected by totchol and glucose. The data mining algorithm will be used for data analysis. In this research paper three methodology logistics, k-nearest neighbor, and decision tree will be used for data analysis. Through this, it can be easily identified how HR can be affected by numerous factors.

SECTION III: PROPOSED METHODOLOGY

I) Logistic regression is a classification-based problem. Which is used to predict binary outcomes either 0 or 1. The values cannot be beyond 0 and 1. It is defined as a curve-like figure which is displayed as "S". In logistic regression some threshold value is defined. If the value is above the threshold value, it tends to 1 otherwise if the value is less than the threshold value, it tends to 0. A confusion matrix is a table that defines the performance of classification-based problems. This algorithm can be used to predict the heart rate prediction because heart rate prediction is a classification-based problem that consists of two classes abnormal as (0) and normal as (1). The model is evaluated by the confusion matrix which defines how many cases the model predicted correctly or wrongly.

Steps for the algorithm:

1) Linear function $y=b_0x_0+b_1x_1+\ldots.b_rx_r$
2) Apply sigmoid function $p(x)=1/1-e^{-y}$
3) If $p(x)>0.5$ then $y=1$ else $y=0$ (binary class either 0 or 1)
4) Repeat the steps till the sigmoid function cover all the points.

II) K-Nearest algorithm is the simplest algorithm used to solve classification-based problems. The data is assigned to the class which has the nearest neighbors. If we increase the no of neighbors, then the accuracy may increase. To apply the algorithm k-nearest number will be given as a parameter. The point will be assigned to the class as per the Euclidian distance-nearest neighbor is a classification and regression-based problem.

Steps for the algorithm:

1) Select the integer number for the nearest data points.
2) Calculate the Euclidian distance (ED) between the testing data and training data.
3) Based on the distance arrange in ascending order.
4) Assign these data points to the specified class.

III) A Decision Tree is a supervised Machine learning algorithm. It is used for both classification and regression problems. It is mostly used for classification tasks. The goal is to build a model

to predict the value of the target label by using simple decision rules inferred from data. The root node contains the whole data which will split further into sub-nodes. These sub-nodes are known as decision nodes. A decision tree split the nodes on all the available variables and selects the one which results in more homogenous nodes. This algorithm is also applied for HR prediction.

Steps for the algorithm:

1) Root node contains the whole training data set.
2) Calculate the Gini for root node through formula Gini $=1-\sum_{i=1}^{n}(pi)^2$.
3) Select the appropriate attribute for splitting.
4) Calculate the entropy by splitting it into two attributes A,B through the formula $E(s) = -\sum_{i=1}^{n} pi\ log2(pi)$.
5) Calculate the Gini GA and GB for the sub-node.
6) If GA>GB then A is the splitting attribute, else B is the splitting attribute.
7) Calculate the Gini for each node.
8) Repeat the splitting steps till the maximum depth is mentioned in the algorithm.

Accuracy level of 0 & 1 class through confusion matrix.

Table 2. confusion matrix for decision tree

	Precision	Recall	F1-score	support
0	0.96	0.90	0.93	721
1	0.81	0.91	0.86	339
accuracy			0.90	1060

A confusion matrix is a technique to evaluate the performance of the algorithm of class 0,1 Precision=TP/TP+FP where TP indicates True positive, False positive (FP) precision is 1(high) if FP is 0. Recall=TP/TP+FN where FN indicates False negative recall is 1(high)if FN is 0.F1Score=2* Precision* Recall/Precision-Recall. In figure predicting 0 and 1 class precision is 0.96,0.81, recall is 0.90,0.91, f1 score is 0.93,0.86. This indicates the prediction of 0 and 1 class accuracy level is 90%.

Figure 2. age, cigarettes per day Max heart rate graph

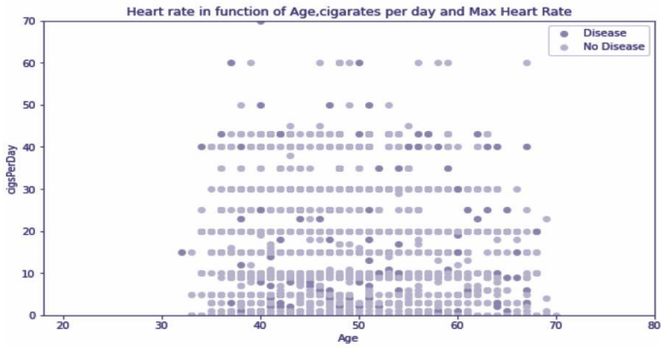


figure 2. indicates that if a person smokes per day affects the HR. It shows the HR reading corresponding with cigarettes per day. Smoking is more affected in the age range 40-65. In [9] smoking is one of the factors for abnormal heart rhythm which occurs in slow or fast heartbeats, skipped heartbeat, and chest pain.

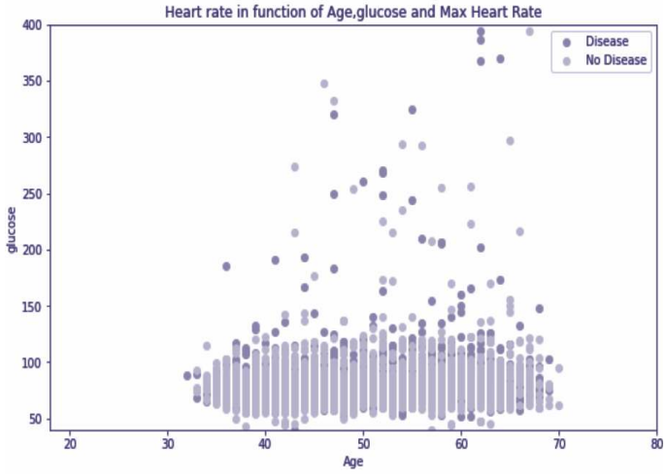
Figure 3. Age, Body mass index, heart rate graph

In Figure 3. BMI range for a normal person is 18.5 to 24.9. if the person's BMI is less than 18 or within 18-20 then the corresponding HR is not in the normal range in the 40-60 age group.

Figure 4. Age, total cholesterol, heart rate graph

In Figure 4. cholesterol range for a normal person is 200-239. In Figure 4. if the cholesterol value is more than 239 it shows a red dot as abnormality or symptoms of disease in the 40-70 age group.

Figure 5. age, glucose, heart rate graph



In Figure 5. glucose range for a normal person is 40-394. Glucose is affecting heart rate in the age of 40-70 range. In Figure 5. if the glucose range is above 100-150 it shows more abnormality in heart rate.

Figure 6. Age, BP, heart rate graph

In Figure 6. HR is more affected if BP is above 100 reading and below 60. The people are more affected by high BP 40-70 age group. In Figure 6., some people are also affected by low BP.

RESULT AND DISCUSSION

In the past studies [2][3] heart rate abnormality is affected by the person who is having smoking & drug habits. After the clinical trial author advised to do exercise or aerobics on regular basis and reduce caffeine & alcohol consumption. Smoking factor affect the heart rate which is displayed in figure 2. In this dataset, here three algorithm logistics, decision tree, and KNN were applied. The logistic algorithm is showing 81% accuracy whereas the logistic algorithm is showing 83 % accuracy. Decision tree will take less time than (logistic, knn) algorithms for small data set. A decision tree-based algorithm is a rule-based algorithm. Every step condition is crosschecked then there is no possibility of false prediction. The Decision tree algorithm is showing 86% accuracy. Comparative analysis also shows the performance and accuracy of the decision tree algorithm. According to the accuracy and performance here, for heart disease prediction decision tree algorithm is selected for creating a model. Due to elevated heart rate following symptoms may occur light-headedness, fainting, chest pain, could not get sufficient oxygen. According to graphical representation, figure 2,3,4,5,6 HR is most affected by smoking, BMI, cholesterol, BP, and glucose.

Comparative Analysis by ROC Curve

Figure 7. ROC Curve

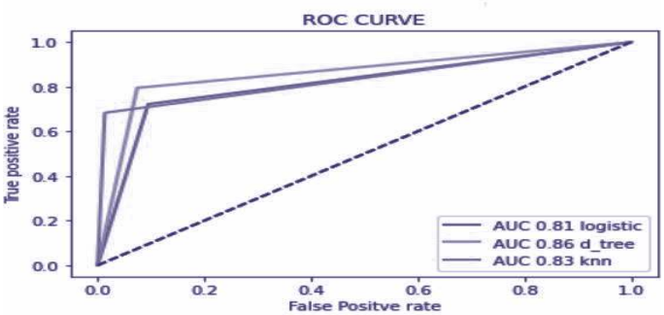
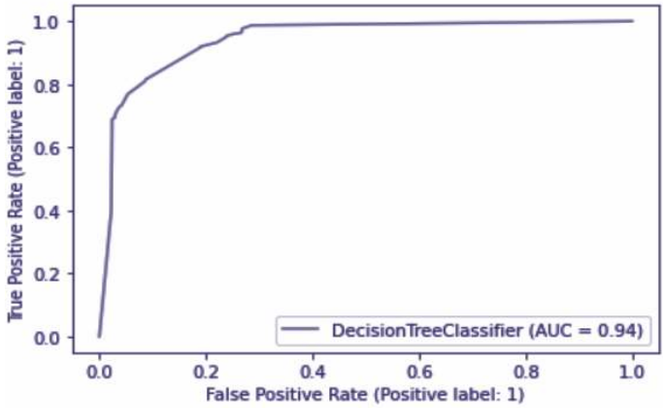


Figure 7. Roc curve is used for comparative analysis of various algorithms. This is the best way to compare the results of the algorithm in which one can easily predict the true positive rate. If the true positive rate is more & the false positive rate percentage is less, then the model predicts accurate results. AUC curve has used the accuracy of the individual algorithm.

AUC curve for Decision tree Algorithm

Figure 8. AUC curve



CONCLUSION

In this paper, three methods are applied to the HR data set. Comparative analysis of algorithm has been done by the ROC curve. This paper includes the comparative analysis of machine learning algorithm also defines the criteria to select appropriate algorithm according to dataset. The algorithm which is predicting the class (0,1) accurately. The above model is working with 86% accuracy with the decision tree algorithm. If the HR value is greater than the threshold value, then it detects danger otherwise heart rate is normal. HR is most affected by smoking, BMI, cholesterol, BP, and glucose.

Future Scope

In the future, A model can incorporate more features for patient monitoring. Analyse multiple classification-related problems or complex problems or huge dataset deep learning algorithm will be selected for better results based on the problem statement.

Compliance with ethical standards

i) Conflict of interest

We Confirm that this manuscript is original and not been published anywhere. First author Deepti Kulkarni and co-author Dr. Rashmi Soni have approved the manuscript and agree with submission to "Industrial Engineering Journal".

ii) Ethical Approval

Not Applicable.

iii) Consent Approval

Not Applicable.

iv) Funding

Not applicable

v) Data Availability Statement

We confirm that heart rate dataset is available on https://github.com/GauravPadawe/Framingham-Heart-Study from github. The analysis on data and result generation has been done by author and co-author.

REFERENCES

1. *Arnold J ,Fitchett D , Howlett J, Lonn E, Tardif J (2008), "Resting heart rate: A modifiable prognostic indicator of cardiovascular risk and outcomes", Canadian Journal of Cardiology , Vol 24 ,2008, pp 3A-15A.*
2. *Palatini P, Julius S (2004), "Elevated heart rate: a major risk factor for cardiovascular disease", Clin Exp Hypertens ,Vol 26, 2004,pp 637-644.*
3. *Boudoulas K.D, Borer J, Boudoulas S (2015), "Heart Rate, Life Expectancy, and the Cardiovascular System Therapeutic Considerations" , Cardiology, Vol 132 no. 4, 2015, pp 199 212.*
4. *Peer N, Lombard C, Steyn K , Levitt N (2020), "Elevated resting heart rate is associated with several cardiovascular disease risk factors in urban-dwelling black South Africans", Scientific Reports, Vol 10(1),2020, pp 1-8.*
5. *Reule S, & Drawz P. E (2012), "Heart rate and blood pressure: any implications for management of hypertension?", Current hypertension reports, Vol 14 ,2012, pp 478-484.*
6. *Yamala S (2020), "Prediction of Heart Diseases using Support Vector Machine" , International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 8, 2020.*
7. *Heo J, Yoon J. G, Park H, KimY. D, Nam H. S, & Heo J. H (2019), "Machine learning–based model for prediction of outcomes in acute stroke", Stroke 50, Vol 5, 2019, pp 1263 1265.*
8. *Batta M. (2020), "Machine learning algorithms-a review" , International Journal of Science and Research (IJSR), Vol 9, 2020, pp 381-386.*
9. *Ibrahim, Mahmood I & Abdulazeez M. Adnan (2021), "The Role of Machine Learning Algorithms for Diagnosing Diseases", Journal of Applied Science and Technology Trends , Vol 4 no 5, 2021, pp 6.*
10. *Akinsola J. (2017), "Supervised Machine Learning Algorithms: Classification and Comparison", International Journal of Computer Trends and Technology (IJCTT) , Vol 48 no. 3,2017, pp128-138.*
11. *Uddin, Shahadat, Khan A, Hossain Md. E, Moni Md. A (2019), "Comparing different supervised machine learning algorithms for disease prediction", BMC medical informatics and decision making, Vol 19 no. 1, 2019, pp 1-16.*