Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 4 August 2022

Prediction of Heart Disease using Machine Learning Algorithms

Dr.R V Durga , KayalaPreethi , Divya Regalla ,
HarshavardhanPolagouni
Department of ECE
Geethanjali College of Engineering & Technology
Hyderabad, India

***Abstract*—Heart disease causes significant mortality rates worldwide and poses a health threat to many. Early prediction of disease outbreaks is a primary missiontoday. Machine learning can be used in medical aidto detect illnesses timely and meticulously.The goal is to construct a coronary heart disease prophesy model primarily derived from theassociated parameters. The dataset used consists of 14 parameters that help in the prediction of the disease. The model has been built using Support Vector Machine, K-NN, and Random Forest Machine Learning algorithms, and a comparative analysis has been driven where itexhibits that the highest accuracy is achieved throughRandom Forest. This version can be useful for clinical interpreters at their sanatorium as a choice guide machine.**

Keywords—Heart Disease, Machine Learning, SVM, K-NN, Random Forest,Accuracy, Attributes.

I. INTRODUCTION

According to World Health Organization, every individual has a fundamental right to good health. Every year, millions of people are getting affected by heart disease, leading to their demise. This is due to a lack of early prediction of the disease[12]. The risk of death can be reduced by predicting the disease at its early stages.

In the modern era, Machine Learning has been given major priority in many of the sectors, where prediction is one of the concerned areas[3]. The prediction of the disease on the real-time data is done by the Machine Learning algorithms by learning the historical data. The prediction is done by building the model using Machine Learning algorithms. The algorithm with the highest accuracy can be preferably used for the foretelling of heart disease.

Authors have already put their efforts into predicting heart disease [4-6] using Machine Learning algorithms. The work done can be considered as an additional effort in predicting heart disease, where the popular Machine Learning algorithms are compared to get the most accurate algorithm.

II. LITERATURE WORK

Much research has been done to evaluate the classification accuracy of various machine learning algorithms using the Cleveland Heart Disease Database.Authors of [2] have obtained an 81% of accuracy measure for the random forest algorithm on this dataset.The authors' research, in which data mining was implemented since it produces the greatest results, aided in the disclosure of information from publicly available data.

Yu-Xuan Wang and colleagues investigated a variety of applications that highlighted the importance of machine learning approaches in a variety of fields[16].It was observed that between 90% and 95% of the apps used both unsupervised and supervised machine learning algorithms.As a result, it was depicted that Machine Learning approaches are critical in the planning of many unique applications in sectors such as medical services and industry.

III. MACHINE LEARNING ALGORITHMS

A. Support Vector Machine

Support Vector Machine or SVM, a machine learning algorithm[9], helps in solving classification issues besides the regression queries. It is best suited for classification issues. It endeavors to intelligibly organize the data points in an N-dimensional space and also specifies the hyperplane.

The model is better when the edge or segregation between the classes is intense. Support vectors are the points that are situated on the boundary. In this, the mapping of the training dataset and the method called kernel is performed. The

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 4 August 2022

advantage of using the kernel trick is that it converts lower-dimensional input space to higher-dimension. The kernel selection is one of the most crucial part to exclude the snag of over fitting and under fittings.

B. K-Nearest Neighbour

K-Nearest Neighbors or K-NN understands the similarity between raw data and ancient data, then arranges the raw data accordingly in a similar category. It is widely used for classification problems rather than regression issues. It is termed as a Lazy Learning algorithm, as it does not learn from the training dataset immediately. It can be used for identifying the class or category of a particular dataset.

In this model, the elbow method is used to calculate the k values. Euclidean distance is determined to know how close the target is to the training dataset. Assign the k-nearest neighbours into rows after finding them and repeat the process for the rowsexceedingthetarg dataset. The highest value of k is selected and then a similar model is built on the k values and then the accuracy is

C. Random Forest

Random Forest is based on the decision trees. The theme of this algorithm is to implement the decision tree algorithm on the dataset by altering the training dataset each time.Decision tree is derived from a tree-like structure where the tree is incorporated with a parent node, branches, and child nodes. The data is examined from the parent node to the child node. Random Forest averages the results of various decision trees employed to discrete subdivisions of a dataset to upgrade the dataset's projected accuracy. It compiles the predictions from individual trees and foretells the result depending on the predominance of votes. The denser the forest, the most accurate is the model, and the risk of overfitting is reduced.

IV. LIBRARIES

The libraries used for the development of the models are as follows:

A. Numerical Python Library

The Numerical Python library, also known as the NumPy library, consists of multidimensional array objects as well as methods for manipulating them. This library can be used for executing mathematical and logical processes.

B. Python Data Analysis Library

Python Data Analysis or Pandas, a non-proprietary Python toolkit, leverages robust data structures to furnish with high-performance data manipulation and interpretation. Independent of the data's origin, data can be loaded, prepared, modified, modeled, and analyzed with Pandas.

C. Matplotlib

A Python package that permits to generate static, animated, and interactive visualizations is termed as Matplotlib. It makes the seemingly impossible possible.

D. Seaborn

Seaborn is a Python data visualization package rooted from the matplotlib library. It includes a high-level interface for building aesthetically appealing and educational data visualizations.

E. Sci-kit Learn

Sci-kit Learn or Sklearnis a Python package for machine learning. It offers a variety of methods for classification, regression, clustering, and dimensionality reduction, and it supports both supervised and unsupervised machine learning. Many libraries such as NumPy and SciPy, are used in the development of this library. Other libraries, such as Pandas and Seaborn, are also compatible with it.

V. METHODOLOGY

The below-listed steps show the method using which the heart disease foretelling model has been built.

A. Data Collection

The Cleveland dataset is used in this analysis. There are 303 cases in total in this collection, with 14 different attributes.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 2, Paper 4 August 2022

TABLE I. ATTRIBUTES

S.No	Attribute	Description
1	age	Age in number
2	sex	Gender of the person
3	cp	Type of chest pain
4	trestbps	Resting blood pressure in mm of mercury
5	chol	Serum cholestrol in milligram/deciliter
6	fbs	Fasting blood sugar
7	restecg	Electrocardiographic results
8	thalach	Maximum heart rate obtained
9	exang	Exercise-induced angina
10	oldpeak	Stress Test depression-induced through angina
11	slope	Slope of Stress Test segment
12	ca	Number of major vessels ranging from 0 - 3 color by fluoroscopy
13	thal	Thallium stress results
14	target	Output class

If the class value is 1, then it is depicted as the person is examined to be positive for heart disease, and if the class value is 0, then it is interpreted as the person is examined to be negative for heart disease. From the dataset, 70% of the data has been contemplated as training data and the rest 30% data as the testing data.

TABLE II. SAMPLE TRAINING DATA

51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1
60	1	2	140	185	0	0	155	0	3	1	0	2	0
67	0	0	106	223	0	1	142	0	0.3	2	2	2	1
45	1	0	104	208	0	0	148	1	3	1	0	2	1
63	0	2	135	252	0	0	172	0	0	2	0	2	1
42	0	2	120	209	0	1	173	0	0	1	0	2	1
61	0	0	145	307	0	0	146	1	1	1	0	3	0

TABLE III. SAMPLE TESTING DATA

58	0	3	150	283	1	0	162	0	1	2	0	2	1
57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
38	1	2	138	175	0	1	173	0	0	2	4	2	1
49	1	2	120	188	0	1	139	0	2	1	3	3	0
55	1	0	140	217	0	1	111	1	5.6	0	0	3	0
55	1	0	140	217	0	1	111	1	5.6	0	0	3	0
56	1	3	120	193	0	0	162	0	1.9	1	0	3	1
48	1	1	130	245	0	0	180	0	0.2	1	0	2	1
67	1	2	152	212	0	0	150	0	0.8	1	0	3	0
57	1	1	154	232	0	0	164	0	0	2	1	2	0

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 4 August 2022

B. Data Preprocessing

The data initialization is done through Exploratory Data Analysis or EDA steps. The attributes in the dataset are identified as continuous and categorical variables. It is then followed by the univariate analysis and bivariate analysis, where the relation between the attributes is analyzed. The missing value treatment is then performed to overcome the imprecise result and reduction of the accuracy of the model. The outlier treatment step is used to upgrade the accuracy in predicting. To make the dataset compatible with Machine Learning algorithms, the values have been changed from numerical to nominal.

C. Building the Model

The models are developed by using Machine Learning algorithms and ensembling techniques. The comparative analysis of the developed models has been done by using the following accuracy measures:

- *Accuracy*
  Accuracy is used to identify the model that is better in recognizing correlations and patterns amongst variables in a dataset. The calculation of accuracy is as shown:

  Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

- *Recall*
  The recall is defined as thequantity of correctly classified positive cases over the quantity of positive examples. If the recall is strong, the class is appropriately identified.

  Recall = $\frac{TP}{TP+FN}$

- *Precision*
  The precision value is the quantity of correctly categorized positive cases over the quantity of expected positive cases. A high precision shows that a positive example is, in fact, positive.

  Precision = $\frac{TP}{TP+FP}$

- *F1 score*
  The F1 Score is the weighted arithmetic mean of Precision and Recall. As a result, the false negatives and false positives are taken into consideration in this score.

  F1 Score = $\frac{2*(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$

- *ROC_AUC*
  Receiver Operating Characteristic Curve or ROC is a curve that envisions the bartering between true-positive rate (TPR) and false-positive rate (FPR). TPR and FPR are calculated and plotted on a single curve for each threshold. The area occupied by the ROC curve, or AUC, is a measurement of how well something works.

Where,
• True Positive (TP) = Positive perception, and the expected outcome is also positive.
• False Negative (FN) = Positive perception, and the expected outcome is also negative.
• True Negative (TN) = Negative perception, and the expected outcome is also negative.
• False Positive (FP) = Negative perception, and the expected outcome is also positive.

D. Measurement of Model Accuracy

The implementation of the models is derived through the performance parameters measures in percentages. The comparative interpretation of the models is described in Table IV. The model with the highest ROC_AUC is considered to be the best of all.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume 5- Issue 2, Paper 4 August 2022

VI. RESULTS

The comparative analysis of the models is as shown:

TABLE IV. PERFORMANCE MEASURES OF MODELS

Model	Accuracy	Recall	Precision	F1	ROC
Support Vector Machine	86.67%	86%	86%	86%	87%
K-Nearest Neighbour	76.67%	86%	71%	77%	77%
Random Forest	90%	86%	92%	89%	90%

From the above analysis, it is clear that the Random Forest model has the highest performance measures of all. Themodel leads the others by having an accuracy of 90%.

VII. CONCLUSION

One of the most common life-threatening disorders encountered around the world is heart disease. The deterioration of health is exacerbated by changing lifestyles and a lack of physical activity. In the medical field, there are a variety of diagnostic procedures. Machine learning, on the other hand, is best optedfor in terms of accuracy. As analyzed,Random Forest is the highest accuracy model. Random Forest is the most preferable model as it creates N decision trees and returns the average of all decision tree outputs as a class.As a result, early-stage prediction accuracy is effectively accomplished. The employment of medical history, specifically data connected to the heart, will aid in the early identification of heart illness or abnormal heart conditions, resulting in the avoidance of eternal fatalities.If a patient or user is unable to contact a doctor, one can use this application to anticipate disease by just entering the report information.

VIII. REFERENCES

[1] Vijeta Sharma, Shrinkhala Yadav, Manjari Gupta, “Heart Disease Prediction using Machine Learning Techniques”2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020, IEEE
[2] M. Kavitha, G. Gnaneswar;R. Dinesh, Y. Rohith Sai, R. Sai Suraj, “Heart Disease Prediction using Hybrid machine Learning Model”, 6th International Conference on Inventive Computation Technologies (ICICT), 2021, IEEE
[3] M. Snehith Raja, M. Anurag, Ch. Prachetan Reddy, Nageswara Rao Sirisala, “Machine Learning Based Heart Disease Prediction System” International Conference on Computer Communication and Informatics (ICCCI), 2021, IEEE
[4] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.
[5] S. Dhar, K. Roy, T. Dey, P. Datta and A. Biswas, "A Hybrid Machine Learning Approach for Prediction of Heart Diseases," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-6, doi: 10.1109/CCAA.2018.8777531.
[6] C. Raju, E. Philipsy, S. Chacko, L. Padma Suresh and S. Deepa Rajan, "A Survey on Predicting Heart Disease using Data Mining Techniques," 2018 Conference on Emerging Devices and Smart Systems (ICEDSS), Tiruchengode, 2018, pp. 253-255, doi: 10.1109/ICEDSS.2018.8544333
[7] https://en.wikipedia.org/wiki/Decision_tree_learning
[8] https://en.wikipedia.org/wiki/Support_vector_machine
[9] K Arunaggiri Pandian, T S Sai Kumar, Sagar P. Dhandare, S Thabasum Aara, “Development and Deployment of a Machine Learning Model for Automatic Heart Failure Prediction” Asian Conference on Innovation in Technology (ASIANCON), 2021, IEEE
[10] Aviral Chanchal, Ajay Shanker Singh, K Anandhan, “A Modern Comparison of ML Algorithms for Cardiovascular Disease Prediction”, 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, IEEE
[11] Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y. K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", © IJRAT Special Issue National Conference “NCPC-2016”, pp. 104-106, 19 March 2016.
[12] https://seaborn.pydata.org/
[13] https://matplotlib.org/
[14] https://www.tutorialspoint.com/python_pandas/python_pandas_introduction.htm
[15] https://www.tutorialspoint.com/numpy/index.htm
[16] Yu-Xuan Wang, QiHui Sun, Ting-Ying Chien, Po-Chun Huang, “Using Data Mining and Machine Learning Techniques for System Design Space Exploration and Automatized Optimization”, Proceedings of the 2017 IEEE International Conference on Applied System Innovation, vol. 15, pp. 1079-1082, 2017.