© MAR 2023 | IRE Journals | Volume 6 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV6I9-1719790

Heart Disease Prediction and Analysis using various Machine Learning Algorithms

SOUMYA TRIPATHI[1], TANYA VERMA[2], SUNNY[3], HARSH[4]
[1]*A. P, Dept. CSE (AI&ML), Panipat Institute of Engineering & Technology, Panipat, India*
[2, 3, 4]*Dept. CSE (AI&ML), Panipat Institute of Engineering & Technology, Panipat, India*

Abstract- Despite numerous advancements in modern medicine, cardiovascular diseases still rank among the top mortality factors all around the world, with 18 million annual deaths caused by heart diseases according to the World Health Organization estimates. Identification of individuals with elevated risk of heart diseases is important in order to reduce the mortality rate especially in developing countries with limited access to medical facilities. Thanks to recent progress made in Artificial Intelligence and Machine Learning, it became possible to create intelligent diagnostic systems that can help doctors diagnose patients at an early stage. This paper aims at evaluating the performance of five supervised machine learning algorithms for heart disease prediction, namely, Random Forest, Support Vector Machine, Logistic Regression, Decision Tree, and K-Nearest Neighbors. Based on experimental evaluation, the Random Forest algorithm shows the best performance providing an accuracy of 90.16% while logistic regression shows accuracy of 85.25%, and the KNN classifier shows the lowest accuracy level of 67.21%. In addition to classification accuracy, the precision and recall values are also estimated based on the confusion matrix. The achieved results suggest that ensembles of learning algorithms provide better predictions while maintaining appropriate level of generalization.

Keywords: Random Forest, Decision Tree, SVM, Machine Learning, Heart Disease, KNN, Logistic Regression

I. INTRODUCTION

Cardiovascular diseases have been identified as a prominent threat to public health in the world and continue to be the leading cause of deaths around the globe. Millions of people die every year because of heart attacks and other cardiovascular diseases. Cardiovascular diseases have become more common due to the presence of an unhealthy lifestyle in the form of poor diet, lack of exercise, and stress generated from job pressures and economic problems. Delayed diagnosis of heart disease is one of the reasons why more and more people die due to heart diseases because their problems are diagnosed at a critical stage.

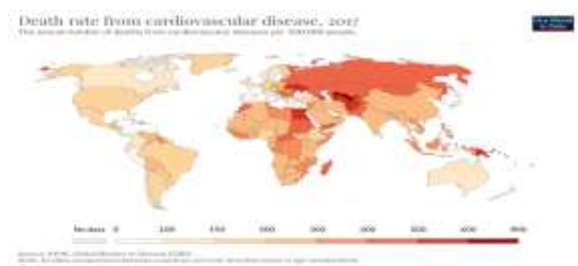

Fig.1. Demise Rate of CVD in 2017

A range of methods for heart disease prediction has been explored through the use of probabilistic classifiers, including Naïve Bayes, supplemented by Laplace smoothing to address zero-frequency issues in classification. The efficiency of the aforementioned models can be estimated based on their implementation in real-time scenarios and performance comparison using various algorithms of supervised learning. Doctors estimate the probability of occurrence of heart disease on the basis of diagnostic procedures and statistical data in order to make the right diagnosis and choose the appropriate treatment strategy. Figure 2 shows the anatomy of the human heart called cardiac organ from the Latin word cardia. The human heart is the muscular organ that consists of four cavities – two atrium and two ventricles. The atrium collects the blood coming back to the heart, and the ventricle makes the necessary pumping action to supply the blood to all organs. Oxygenated blood carried by the circulatory system provides the body tissues with the required amount of oxygen and nutrients.

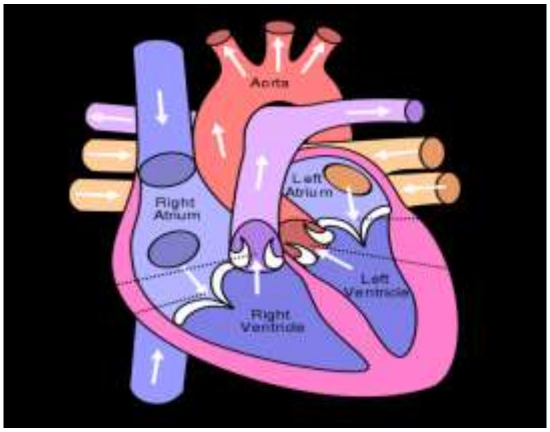

Fig.2. Anatomic Structure of Heart

In predicting heart disease, the emphasis is made on determining clinical and physiological variables that can cause cardiovascular disease. A number of pathological problems such as atherosclerosis, constriction of the blood vessels, deposition of lipids, high blood pressure, and poor blood circulation greatly raise the risk of heart disease. Moreover, various medical variables like blood pressure, cholesterol content, blood glucose levels, platelets, and other similar clinical variables play an important role in machine learning algorithms to classify diseases. Deposition of lipids inside the arteries hinders proper blood supply to the heart.

II. METHODOLOGY

The data set that was used in this study is related to heart disease and was downloaded from the Kaggle database. This data set has a number of clinical features associated with heart disease, although other data sets could have been used depending on their availability. The data were pre-processed to remove inconsistencies in the data set since low-quality data can negatively impact the performance of the model. The data set was further split into 70% training data and 30% testing data. The training data was used to develop predictive models while testing data was used to test the performance of these models. Five supervised machine learning techniques were used; they include:

Logistic Regression: Logistic Regression (LR) is a type of statistical algorithm often used in binary classification applications. LR determines the likelihood of an event occurring from one or multiple independent variables using a logistic function. Even though LR does not make predictions about class membership but rather probabilities of target outcomes, LR is a classification technique because of the use of a predetermined decision criterion where data points whose probabilities fall below the criterion are labeled as one category while those that exceed the criterion form the other class. Logistic regression is broadly classified into three types: Binary Logistic Regression, Multinomial Logistic Regression (MLR), and Ordinal Logistic Regression (OLR). MLR is applied in situations where the target variable comprises of multiple unordered classes, OLR is applied in cases where the classification categories are ordered.

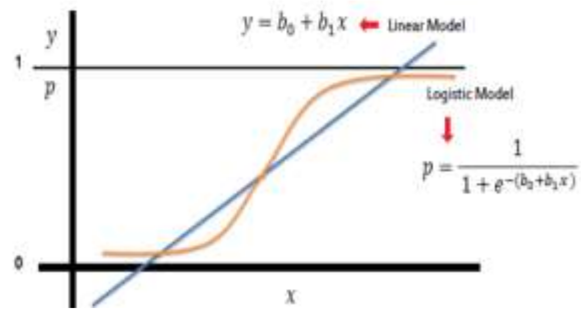

Fig.3. Logistic Regression Model

Support vector machines (SVMs) are one of the most commonly used supervised machine learning algorithms to perform classification of linear as well as non-linear data. Based on the training data set with labels, SVM computes the hyperplane which optimally separates the various classes. This hyperplane creates regions in the feature space and thereby helps to classify the input samples based on their location with respect to the decision boundary. In the two dimensional space, the hyperplane is described by a line which separates the two classes on either side of it [12]. In this study, SVM technique was utilized to detect patterns from the training dataset and construct a model for heart disease prediction.

© MAR 2023 | IRE Journals | Volume 6 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV6I9-1719790

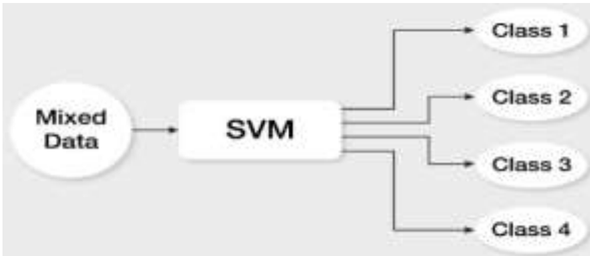

Fig.4. SVM differentiate two classes

Decision Tree: Decision Tree is one of the most commonly applied supervised machine learning algorithms in healthcare applications. Some of the basic elements of the decision tree include root node, branch, and leaf node. This algorithm repeatedly splits the dataset into small parts according to some selected feature values up to when a decision is made at the leaf nodes. An internal node denotes a decision-making criterion while each leaf node indicates the predicted class label. Decision Tree can perform both classification and regression operations and has the ability to detect the relationship between the set of features to predict outcomes. In developing a decision tree, the existing dataset is split into training and testing datasets, in which the training dataset is utilized to create the decision tree while the testing dataset is utilized to assess its predictive capability.

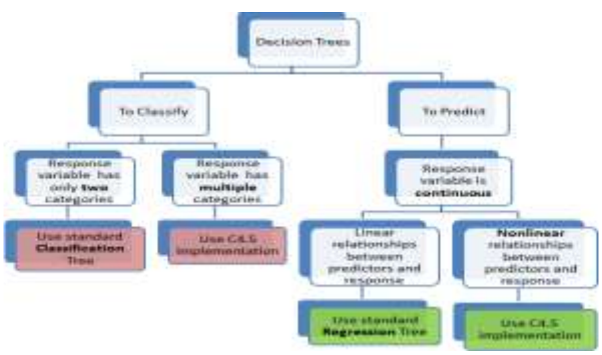

Fig.5. Decision Tree algorithm flowchart

KNN: K-Nearest Neighbour is a simple machine learning algorithm for performing classification and regression tasks. The classification is done using the similarity between the new data instance and the already available training data, and the new instance is then assigned to the class of the nearest neighbors. In contrast to other algorithms that create a model during learning, K-Nearest Neighbor does not learn any model but stores all the training data instances and makes predictions based on the similarity or distance measure. Though the algorithm can be used for classification and regression purposes, it is mostly used for classification problems. Being a lazy learner, K-NN is slower than most other supervised machine learning algorithms.

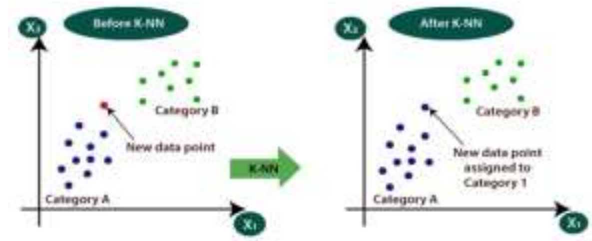

Fig. 6. KNN Machine Learning Algorithm

Random Forest:
Random Forest: Random Forest is a machine learning algorithm which uses many decision trees to increase the predictive power of the output. Each of these decision trees uses different training sets of the data and different features to predict the output class independently. The final output is calculated using the majority voting technique and hence it produces a better and more accurate model. It can be used in regression as well as classification tasks. Many decision trees are built during the training process and their overall output is considered to produce the final prediction. This technique effectively decreases the variance and captures the interactions between the input features. But one of the drawbacks of this technique is that the output model is not easily interpretable.

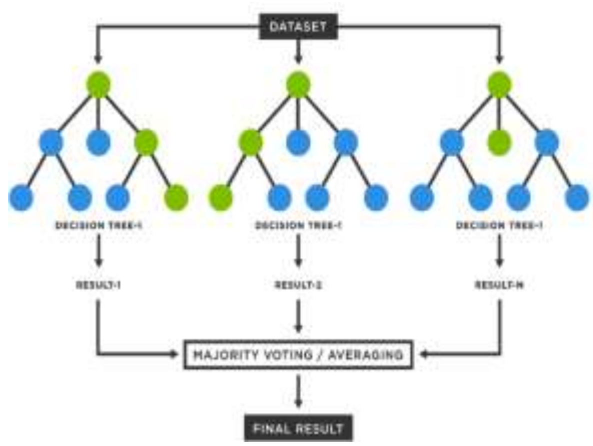

Fig. 7. Plot of Simplified Random Forest

© MAR 2023 | IRE Journals | Volume 6 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV6I9-1719790

III. RESULT AND DISCUSSION

In the current experiment, a standard heart disease dataset collected from the Kaggle repository was used, which includes various clinical features associated with human health. Prior to creating a predictive model, data pre-processing was done in order to eliminate any form of ambiguity in the data, since ambiguous data may hamper the performance of machine learning algorithms. After pre-processing, the dataset was split into 70 percent training and 30 percent testing datasets. Though, a number of different train-test splits can be used, the key purpose here is to build an efficient model that can predict with maximum accuracy. Finally, a total of five supervised machine learning algorithms, i.e., Logistic Regression, KNN, SVM, Decision Tree, and Random Forest, were developed for heart disease prediction. The entire process was done in Google Colab using Python. The results obtained from the supervised learning algorithms are presented in the next section.

Table 1 Accuracy analysis of Logistic Tree, SVM, KNN, Decision Tree and Random Forest Algorithms

Algorithms	Logistic Tree	SVM	KNN	Decision Tree	Random Forest
Accuracy	85.25	81.97	67.21	81.97	90.16

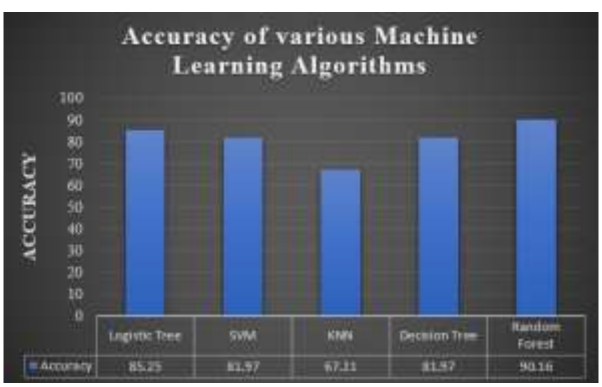

Fig. 8. Comparative analysis of accuracy among implemented algorithms

Confusion Matrix

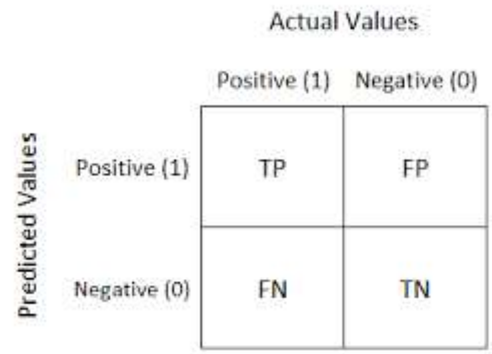

Table 4.2 Comparative analysis of different algorithm for Precision and Recall

Parameter/Algorithm	Logistic Tree	SVM	KNN	Decision Tree	Random Forest
Precision	72.88	88	70.46	85.27	91.67
Recall	94.47	91.67	75.28	89.35	92.50

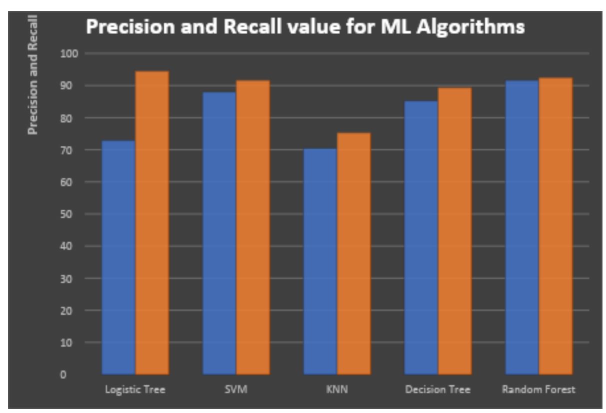

Fig.9. Precision and Recall value for ML Algorithms

IV. CONCLUSION

With the rise in cardiovascular disease cases, the need for early detection has been a key healthcare concern due to high death rate. The recent advances in Soft Computing, Machine Learning (ML) and Deep Learning (DL) fields have tremendously enhanced the ability of developing intelligent models that predict diseases with high accuracy. In this paper, a typical heart disease dataset obtained from Kaggle database was pre-processed in order to increase its data quality prior to model development. The data was then split to 70% training set and 30% test set, and five supervised ML algorithms including Logistic Regression, KNN, SVM, Decision Tree and Random Forest were developed for comparative analysis. Out of the studied models, Random Forest algorithm demonstrated the best prediction accuracy of 90.16% while KNN delivered the worst prediction accuracy of 67.21%. The accuracy of Decision Tree and SVM models was found equal to 81.97%. Along with accuracy, precision and recall were considered in this paper. In addition to this, the accuracy obtained using the Random Forest classifier was 91.67% for precision and 92.50% for recall. From this experiment, it can be concluded that the use of the Random Forest classifier is better when compared to the other supervised machine learning algorithms used here.

REFERENCES

[1]. Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath, “Heart disease prediction using machine learning algorithms”, ICCRDA 2020, IOP Conf. Series: Materials Science and Engineering, 1022 (2021) 012072 IOP Publishing DOI:10.1088/1757-899X/1022/1/012072

[2]. Pranav Motarwar, Ankita Duraphe, G Suganya, M Premalatha, “Cognitive Approach for Heart Disease Prediction using Machine Learning”, 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), IEEE, 2020, DOI: 10.1109/ic-ETITE47903.2020.242

[3]. Vijeta Sharma; Shrinkhala Yadav; Manjari Gupta, “Heart Disease Prediction using Machine Learning Techniques”, 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE, 18-19 Dec. 2020, DOI: 10.1109/ICACCCN51052.2020.9362842

[4]. Atharv Nikam; Sanket Bhandari; Aditya Mhaske; Shamla Mantri, “Cardiovascular Disease Prediction Using Machine Learning Models” 2020 IEEE Pune Section International Conference (PuneCon), IEEE, 16-18 Dec. 2020, DOI: 10.1109/PuneCon50868.2020.9362367

[5]. Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”, IEEE Access, Special Section On Smart Caching, Communications, Computing And Cybersecurity For Information-Centric Internet Of Things, DOI: 10.1109/ACCESS.2019.2923707

[6]. Dinesh Kumar G, Santhosh Kumar D, Arumugaraj K, Mareeswari V, “Prediction of Cardiovascular Disease Using Machine Learning Algorithms”, Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India

[7]. Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar, “Prediction of Heart Disease Using Machine Learning”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018), IEEE Conference, IEEE Xplore ISBN:978-1-5386-0965-1

[8]. Zhang, Shuai, Y.-L. S. A. (2017), ‘Deep learning based recommender system: a survey and new perspectives’, Journal of ACM Computing Surveys 1(1), 1–35.

[9]. AminKhatami, AbbasKhosravi, C. L. (2017), ‘Medical image analysis using wavelet transform and deep belief networks’, Journal of Expert Systems With Applications 3(4), 190–198.

[10]. AshwiniShetty, Naik, C. (2016), ‘Different data mining approaches for predicting heart disease’, International journal of innovative research in science, engineering and technology 3(2), 277–281.

[11]. Aydin, S. (2016), ‘Comparison and evaluation data mining techniques in the diagnosis of heart disease’, Indian journal of science and technology 6(1), 420–423.

[12]. Bayasi, N. and Tekeste (2016), ‘Low-power ECG-based processor for predicting ventricular arrhythmia’, Journal of IEEE transactions on very large scale integration systems 24(5), 1962–1974.

[13]. Berikol, B. and Yildiz (2016), ‘Diagnosis of acute coronary syndrome with a support vector machine’, Journal of Medical System 40(4), 11–18.

© MAR 2023 | IRE Journals | Volume 6 Issue 9 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV6I9-1719790

[14]. Zhiyong Wang, Xinfeng Liu, J. G. (2016), 'Identification of metabolic biomarkers in patients with type-2 diabetic coronary heart diseases based on metabolomic approach', 6(30), 435–439.

[15]. Singh, M. and Martins (2016), 'Building a cardiovascular disease predictive model using structural equation model and fuzzy cognitive map', Journal of fuzzy system 02(6), 1377–1382.

[16]. S.Prabhavathi (2016), 'Analysis and prediction of various heart diseases using DNFS techniques', International journal of innovations in scientific and engineering research 2(7), 678–684.

[17]. Sali, R. and Shavandi, M. (2016), 'A clinical decision support system based on support vector machine and binary particle swarm optimisation for Cardiovascular disease diagnosis', International Journal of Data mining and Bio-informatics 15(1), 312–327.

[18]. Ghadge, Prajakta, K. (2016), 'Intelligent heart attack prediction system using big data', International journal of recent research in mathematics computer science and information technology 2(2), 73–77.

[19]. G.Purusothaman and K.Krishnakumari (2015), 'A survey of data mining techniques on risk prediction: heart disease', Indian journal of science and technology 8(5), 643–651.

[20]. Richter, A., Listing, J., Schneider, M., Klopsch, T., Kapelle, A., Kaufmann, J., Zink, A. and Strangfeld, A. (2015), 'Impact of treatment with biologic dmards on the risk of sepsis or mortality after serious infection in patients with rheumatoid arthritis', Annals of the rheumatic diseases pp. 147–153.

[21]. Sairabi, Mujawar, D. (2015), 'Prediction of Hear t Disease using Modified K-means and by using Naive Bayes', International Journal of Innovative Research in Computer and Communication Engineering 3.

[22]. Vafaie, Ataei, M. (2014), 'Heart diseases prediction based on ECG signals classification using a genetic-fuzzy system', Journal of biomedical signal processing and control 14(5), 291–296.

[23]. Wang (2014), 'Study on qi-deficiency syndrome identification modes of Coronary heart disease based on metabolomic biomarkers', Journal of evidence-based complementary and alternative medicine 24(16), 192–198.

[24]. Yang, X., Li, M., Zhang, Y. and Ning, J. (2014), 'Cost-sensitive naive bayes classification of uncertain data', Journal of Scientific World 9(8), 1897–1904.

[25]. Chandna, D. (2014), 'Diagnosis of heart disease using data mining algorithm', International journal of computer science and information technologies 5(2), 1678–1680.