An Effective Analytics and Performance Measurement of Different Machine Learning Algorithms for Predicting Heart Disease

S. M. Hasan Sazzad Iqbal, Nasrin Jahan, Afroja Sultana Moni, Mst. Masuma Khatun
Department of Computer Science and Engineering
Pabna University of Science and Technology
Pabna, Bangladesh

***Abstract*—This Heart disease means any condition that affects to directly heart. Globally, Heart disease is the main reason for death. According to a survey, approximately 17.9 million people died from heart disease in 2019 (representing 32 percent of global deaths). The number of people dying is increasing at an alarming rate every day. So it is necessary to detect and prevent heart disease as soon as possible. Medical experts who work inside the field of coronary heart sickness can predict the rate of coronary heart disorder up to 69%, which is not so useful. Because of the invention of various machine learning techniques, intelligent machines can predict the chance of heart disease up to 84%, which will be helpful to prevent heart disease earlier. In this paper, for picking essential characteristics among all features in the dataset, the univariate feature selection approach was employed. One-of-a-kind machine learning algorithms like K-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine were used to assess the performance of these algorithms and forecast which one performs best. These machine learning approaches require less time to predict disease with more precision, resulting in the loss of valued lives all around the world.**

***Keywords*—Machine learning; heart disease prediction; KNN; Naive Bayes; decision tree; random forest; support vector machine**

I. Introduction

Health is considered as a whole state of physical, mental, and social well-being where there's an absence of disease and infirmity. Health may be evolved through doing several activities like a physical workout, adequate sleep, and using utilizing employing through averting unhealthful sports like smoking or immoderate stress. Because of carelessness, health is being affected by various kinds of diseases. As it is known that the human heart is one of the most essential organs in the human body [12]. The average human heart beats 72 beats in step with a minute and pumps about 2000 gallons of blood to each and each part of the human body. But somehow, if the heart is affected by several diseases, then it'll be harmful to the human body, and sometimes it'll cause death also. Nowadays, heart disease is increasing at an alarming rate. Medical professionals can't get an accurate result of heart disease prediction by following a custom. With the assistance of machine learning algorithms, the prediction can be increased and many people can get alert about their disease and can also take preventive actions before it is too late. With the help of machine learning strategies, it's far feasible to collect information from a massive quantity of information and by training the dataset, the machine can predict the result. So it reduces the extra burden on medical professionals. As in the modern world, it can't be imagined daily lives without technology, machine learning has made life easier by predicting and providing proper guidelines about disease. By using machine learning techniques, millions of lives can be saved by predicting disease quickly and providing quicker service to the patients.

A. Problem Statement

From previous research, it came to know that they examined different machine learning techniques. These studies concentrated on a specific impact of machine learning techniques rather than on their optimization. Some researchers experimented with hybrid optimization techniques. The initial stage in this effort is to apply a correlation-based feature selection method. Among all the attributes of the dataset, only the correlated datasets are segmented and this is called the feature selection method. It is a preprocessing method of machine learning which eliminates irrelevant data and increases learning accuracy. To increase classification accuracy, the best subset of features is chosen from all of them. Different machine learning methods are applied to the entire dataset after it has been divided into train and test datasets. After the comparison of different algorithms, identify the algorithm which performs best for predicting heart disease.

II. Related Work

Lots of work has been done in the field of predicting heart disease in previous years. They have attained different levels of accuracy by applying different machine learning techniques. Some of them are given below:

The identity of coronary heart sickness, diabetes with the assistance of neural networks turned into brought by Niti Guru [1]. Experiments had been finished on a sampled dataset of affected person's records. The neural network changed into educated and tested with 13 input capabilities. The supervised set of rules changed into used for the diagnosis of heart sickness. The backpropagation algorithm was used for training data. Whenever any unfamiliar data was inserted, the process identified the unknown data as compared to training data and produced a probability of heart disease.

Another prediction was introduced by M.Sultana, A.Haider, and Mohammad Shorif Uddin [2]. They have illustrated that datasets that are available for heart disease are in the form of the raw datasets and are inconsistent. They extracted the crucial features from the dataset. By using this method, the time complexity and work of the training algorithm were reduced and the accuracy of the proposed model increased. They have worked with Bayes Net and SMO classifiers which are more optimal than NLP and KStar. They have collected datasets from WEKA software and measured performance by running algorithms (Bayes Net and SMO). Then compared the result with predictive accuracy, ROC Curve, and ROC Value.

An optimization of function was performed to acquire better accuracy the usage of Decision Tree by M.A.Jabbar, B.L. Deekshatu, and P. Chandra [3]. It turned into a method for early detection of coronary heart disease.

K.C.Tan, E.J. Teoh proposed a hybrid method of machine learning where Support Vector Machine and Genetic Algorithm were combined [4]. The LIBSVM and WEKA facts mining tools have been used to investigate the result. They've used five datasets for this project. After applying the SVM and Genetic algorithm, they obtained an accuracy of 84.07%.

G. Parthiban and S.K. Srivastava diagnosed coronary heart sickness in diabetic patients using Naive Bayes and SVM algorithms [5]. A record of 500 patients was used to make a prediction. After applying both the algorithms, Naive Bayes provided an accuracy of 74%, and SVM provided an accuracy of 94.60%.

V. Chaurasia and S. Pal experimented on diverse records mining techniques to come across coronary heart disease [6]. The WEKA data mining tools were used right here. They used Naive Bayes, J48, and bagging for this cause. They took a dataset of 76 attributes. Among them, they selected only 11 attributes to make a prediction. Of the three classification algorithms, bagging provided better accuracy to make a prediction.

Fahd Saleh Alotaibi proposed a version comparing among 5 one of a kind machine learning algorithms [7]. In this research, he used Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM algorithms. Among them, the Decision Tree gave the best accuracy.

After reviewing the above papers, the main concept in the back of the proposed gadget is to make a prediction of heart disease primarily based on the given entered facts. For this purpose, we've used KNN, Naive Bayes, Decision Tree, Random Forest, and SVM set of rules primarily based on their accuracy.

III. Methodology

Exploring classification techniques and performing performance analysis, our suggested model predicts heart disease. Our primary goal is to determine whether or not a patient has heart disease. The flow chart depicts the full procedure, Fig. 1.

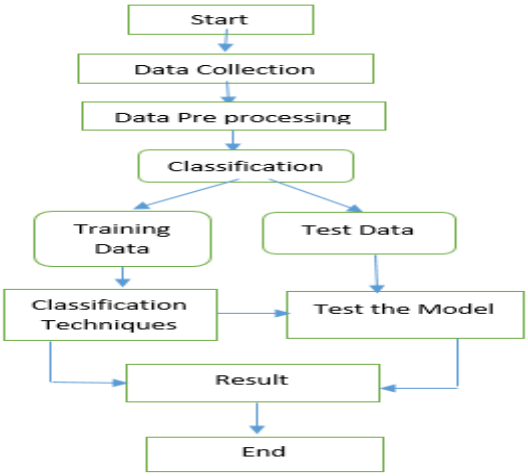

Fig. 1. A Model Predicting Heart Disease.

A. Data Collection and Preprocessing

The dataset we utilized to predict heart disease is a 16-attribute dataset containing 4241 patient records. After gathering data, dimensionality reduction is used. It entails feature extraction and feature selection. The data we've collected has several features or dimensions, but not all of them are necessary or important in terms of the model's output. A huge number of attributes may have an impact on the computational complexity, resulting in a poor outcome.

B. Feature Selection

This approach selects a subset of the original features. The best features are chosen using a univariate feature selection method. The main features are chosen using the chi-square statistics test. In our project, we first standardized our data and then selected the important features. After performing the task, the selected attributes decreased from 16 to 6.

IV. Algorithms and Techniques Used

Because heart disease prediction is a classification or clustering problem, it can be reduced to a single classification with a small number of attributes. It will be easier to identify the correct class as a consequence of this classification challenge, and the result will be more accurate than clustering. We've spoken about the theoretical context that we employed during the experiment in this part. To forecast the best result, we applied five different machine learning methods. Based on their popularity, the KNN, Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine techniques are utilized.

A. K-Nearest Neighbor

K-Nearest Neighbor rule was introduced by Hodges et al. in the year 1951 [8]. He introduced it as a "non-parametric technique for pattern classification". Which is popularly known as K-Nearest Neighbor Algorithm. It is a completely effective and popularly used classification algorithm. The KNN technique may be used for each regression and classification, but it is usually applied for classification tasks. When there's less or no prior knowledge approximately information distribution, it is then used. This set of rules reveals the k-

nearest statistics points inside the training set to the information point for which a goal value is unavailable. Then it assigns the common fee that it has found information factors to the new predicted factor. K-NN is certainly a non-parametric algorithm, which means that it makes no guesses about the underlying records. It's referred to as a lazy learner set of rules because it doesn't learn from the training set at once; rather, it saves the dataset and performs a motion on it when it comes time to categorize it. During the training segment, the KNN algorithm just shops the dataset, and while it accepts new information, it classifies it into a class that's quite similar to the new information [9]. For instance, there are two classes, Category A and Category B, and we get some other new records factor x1. Which of those categories will this information factor suit? A KNN set of rules is needed to solve this sort of trouble. We can readily find out the category or class of a dataset with the assistance of KNN.

B. Naive Bayes

Naive Bayes is an easy classification algorithm that is based totally on Bayes' theorem. It is a collection of classification algorithms and thus it is called a family of algorithms. Naive Bayes assumes independence between the features of the data. It is useful for very large datasets and easy to build. It mainly works depending on probability [10]. It provides calculation posterior probability P(c|x) from P(c), P(x) and P(x|c). The equation that is used to calculate the probability is given below:

$$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$$

Or, it can be represented as:

$$\text{Posterior Probability}=\frac{\text{Likelihood}*\text{class prior probability}}{\text{Predictor Prior Probability}}$$

$$P(c|x)= P(x1|c)\times P(x2|c)\times \ldots\ldots\ldots \times P(xn|c) \times P(c)$$

Here,

- P(c|x) is the posterior probability of the target class.
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor of the given class.
- P(x) is the prior probability of the predictor.

Naive Bayes is easy and fast to estimate the test information set's elegance. It's also proper at multi-class prediction. About to with concerning numerical input variables, it scores well with specific input variables. When the expectancy of independence is met, a Naive Bayes classifier outperforms different fashions.

C. Decision Tree

A Decision Tree is a supervised learning algorithm that is normally used in classification troubles. Here, data is constantly cut up according to a parameter. It is a tree-structured classifier where the decision tree is represented as decision nodes and leaves. The internal nodes of the decision tree denote the features of the dataset, the leaf node denotes the output and each branch denotes decision rules. The decision tree solves the problems by representing them graphically to get all the possible solutions. A decision tree simply continues its splitting process by asking a question whether the answer is positive or negative. Based on its answer, it cut up the trees into subtrees. The basic structure of a decision tree is depicted in the Fig. 2.

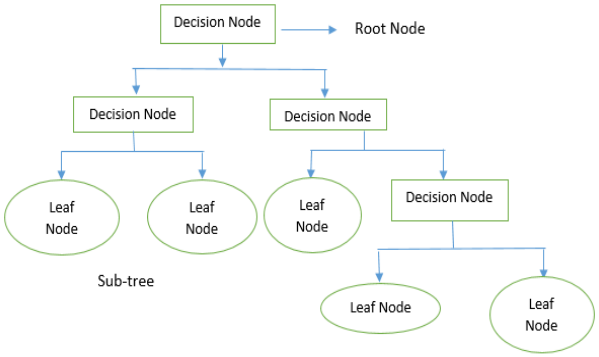

Fig. 2. Decision Tree.

When making any decision, decision trees usually mimic human thinking ability, which is easy to understand. It also uses a tree-like structure for making a decision [11]. The entropy of each characteristic is initially calculated by the decision tree algorithm. It calculates a feature's information and it is known as information gain. Based on information gain, we split the node to make a decision tree. Which characteristic has the highest value is split first, according to the value of information gain. The formula which it follows is:

Information gain= Entropy(S) – [(Weighted avg)*Entropy (each feature)]

Here, entropy is the measurement of impurity in a given attribute. Entropy can be measured as:

Entropy(S) = -P(yes)log 2 P(yes) – P(no)log 2 p(no)

Where,

S= Total no of samples

P(yes) = Probability of yes

P(no) = Probability of no

Decision trees can generate rules that are simple to understand. Classification is achieved via decision trees, which do not require a whole lot of computing. Decision trees indicate which fields are most relevant for prediction.

D. Random Forest

Random Forest may be a widely used supervised machine learning technique. It's a type of ensemble learning that's used to solve classification and regression problems. It generates multiple decision trees and analyzes them for making a decision. It is mainly used for classification problems. Random Forest algorithm can handle datasets containing continuous and categorical variables both. For regression problems, continuous variables are employed, while categorical variables are used for classification problems. Because it integrates several models, Random Forest is an ensemble algorithm. Here, a collection of

models is used rather than an individual model to make prediction easier. In the Random Forest algorithm, at first, some attributes are selected randomly from the data set. Then decision trees are constructed separately for each attribute. Each decision tree generates individual outputs using the decision tree method. And finally based on the maximum value of the decision tree, it generates the final output, Fig. 3.

As in comparison to other algorithms, it takes less time to train the dataset. When a massive quantity of the facts is missing, it can nevertheless preserve accuracy. It can predict output with proper accuracy, and it runs successfully despite a large dataset.

E. Support Vector Machine

Support Vector Machine is a supervised machine learning technique for categorization that analyzes data. It was developed by Vladimir Vapnik with colleagues in years 1992, 1993, 1995, 1997 [13]. Support Vector Machines (SVM) is a rapid and reliable classification method that works nicely with little amount records. Support Vector Machine is different from different classification algorithms because it chooses the decision boundary which maximizes the space from the nearest data factors. It is suitable for classification problems. The SVM algorithm's challenge is to discover a hyperplane in an N-dimensional space that exactly classifies the data factors. In the feature space, the SVM reveals the hyperplane which differentiates between the lessons [14]. For an SVM model, the data points which are trained, are segregated by a margin that belongs to a separate class. Then test data points are mapped into the same region to determine which side of the margin they will land on, Fig. 4.

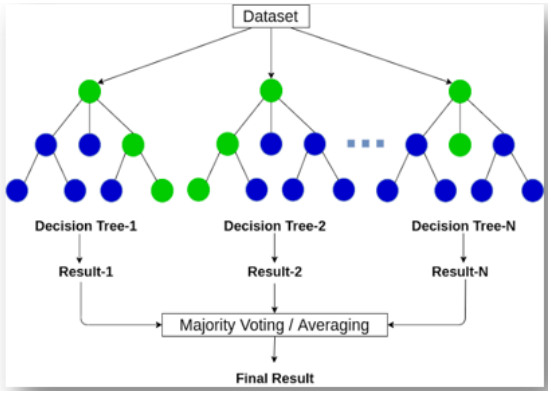

Fig. 3. Random Forest.

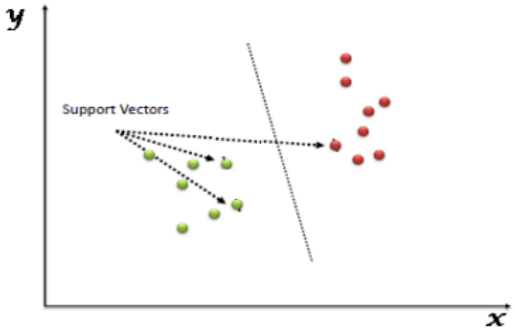

Fig. 4. Support Vector Machine.

In situations with a lot of dimensions, SVM works well. For the decision functions, several kernel functions can be given, as well as unique kernels. It saves memory by using a subset of training factors named support vectors in the selection feature.

V. Result and Evaluation

We employed five distinct types of classification algorithms to predict heart disease in this procedure. After preparing the data, we ran it through various categorization algorithms to see how well it performed. For heart disease prediction, we employed the K-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine algorithms. We evaluated each algorithm's performance using accuracy, precision, recall, and f1 score values. We learned from our experiment that Naive Bayes performed the best of all algorithms, with an accuracy of 83.96 percent. Support Vector Machine performed admirably, with an accuracy of 84.08 percent, practically identical to Naive Bayes. Although it has higher accuracy than Naive Bayes, it performs poorly in terms of precision, recall, and F1 scores. The following "Table I" is a table of our experiment's performance measure:

TABLE I. Performance of Various Algorithms

Algorithms	Accuracy	Precision	Recall	F1- Score
KNN	81.13%	76%	81%	78%
Naive Bayes	83.96%	81%	84%	81%
Decision Tree	73.70%	74%	74%	74%
Random Forest	83.02%	76%	83%	78%
SVM	84.08%	71%	84%	77%

The performance of different algorithms is also represented through a bar chart, which is given in Fig. 5.

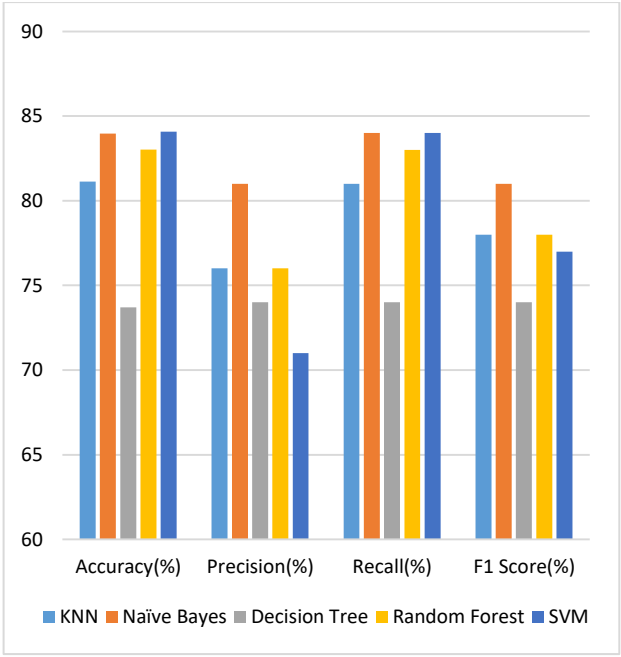

Fig. 5. Bar Chart of Performance for different Algorithms.

VI. Conclusion and Future Work

Machine learning can properly predict and guide treatment for heart disease, but models that include social determinants of health capture risk and outcomes for a wider range of people. A correct prediction of heart disease can save a person's life, while an incorrect prediction can be fatal. As can be seen from the preceding assertions, there is a lot of potential for applying various machine learning algorithms to predict cardiac disease. Our research aims to assess the performance of various machine learning algorithms and forecast which algorithm would perform better in this scenario. We've collected raw datasets, pre-processed them, and tested them for making a prediction. Some algorithms performed best or some performed worst in some cases. Naive Bayes performed the best accuracy for our dataset. Support Vector Machine algorithm also performed well but in comparison to Naive Bayes algorithm, its outcome was poor. Here, the Decision tree performed poorly in some cases. Random Forest also fared well since it used many Decision Trees to overcome the problem of overfitting. For our dataset, Naive Bayes performed well which can be used for predicting heart disease. By using these techniques for detecting heart disease, millions of lives can be saved.

The systems we employed in this work performed well in terms of predicting cardiac disease but still, there are some limitations in our research including limitation of processing power, time limit available for this research. Future research is needed to deal with high-dimensional data and overfitting. This document can serve as a starting point for learning how to anticipate cardiac disease, and it can be expanded to a more advanced level.

Acknowledgment

All thanks are due to Allah SWT, who created us and elevated us to the highest rank among his creations. First of all, I would like to admit my thanks to Allah SWT for enabling me to perform this thesis successfully. I'd like to thank my respected supervisor from the bottom of my heart, S. M. Hasan Sazzad Iqbal, Assistant Professor, Department of Computer Science and Engineering (CSE), Pabna University of Science Technology (PUST), for his scholastic supervision, valuable guidance, adequate encouragement and helpful discussion throughout the progress of this work. I owe him a debt of gratitude for enabling me to do this research under his supervision.

Finally, I owe a great deal to my family members, particularly my parents, as well as all of my friends and well-wishers for their encouragement and support.

References

[1] Rajpal, N. Decision Support System for Heart Di Decision Support System for Heart Disease Diagnosis Using Neural Network Niti Guru* Anil Dahiya.

[2] Sultana, M., Haider, A., & Uddin, M. S. (2016, September). Analysis of data mining techniques for heart disease prediction. In 2016 3rd international conference on electrical engineering and information communication technology (ICEEICT) (pp. 1-5). IEEE.

[3] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, 85-94.

[4] Tan, K. C., Teoh, E. J., Yu, Q., & Goh, K. C. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. Expert Systems with Applications, 36(4), 8616-8630.

[5] Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. International Journal of Applied Information Systems (IJAIS), 3(7), 25-30.

[6] Chaurasia, V., & Pal, S. (2014). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol, 2, 56-66.

[7] Alotaibi, F. S. (2019). Implementation of a machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 10(6), 261-268.

[8] Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.

[9] Rajathi, S., & Radhamani, G. (2016, March). Prediction and analysis of Rheumatic heart disease using kNN classification with ACO. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 68-73). IEEE.

[10] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017, July). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In 2017 IEEE Symposium on Computers and Communications (ISCC) (pp. 204-207). IEEE.

[11] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology, 7(2.8), 684-687.

[12] Nahiduzzaman, M., Nayeem, M. J., Ahmed, M. T., & Zaman, M. S. U. (2019, December). Prediction of heart disease using multi-layer perceptron neural network and support vector machine. In 2019 4th International conference on electrical information and communication technology (EICT) (pp. 1-6). IEEE.

[13] Kannan, R., & Vasanthi, V. (2019). Machine learning algorithms with ROC curves for predicting and diagnosing heart disease. In Soft Computing and medical bioinformatics (pp. 63-72). Springer, Singapore.

[14] Schuldt, C., Laptev, I., & Caputo, B. (2004, August). Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 3, pp. 32-36). IEEE.