Innovating Across Boundaries: Bridging Science, Technology, and Humanities
September 2024
ISBN Number: 978-93-95305-66-2

PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS FOR HEART ATTACK PREDICTION

Paras Negi
Research Scholar,
Department of Computer Science,
S.S.J. University, Campus, Almora, Uttarakhand, India
Manoj Kumar Bisht
Assistant Professor,
Department of Computer Science,
S.S.J. University, Campus, Almora, Uttarakhand, India

Abstract: Machine learning is a branch of artificial intelligence that generates predictions without requiring explicit programming instructions. Machine learning techniques such as Artificial Neural Network (ANN), Decision tree, K-Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, Multi layer perceptron etc, are used to predict the heart disease. This study offers an analysis of the existing algorithm, provides a comprehensive overview of the previous research and evaluate the performance of the machine learning modals. In the study, patients with a low and high probability of having a heart attack were examined. The results indicate that methods Logistic Regression algorithms outperform traditional classifiers in terms of prediction accuracy and generalization.

Keywords: Machine learning (ML), Logistic Regression, Confusion matrix, K-nearest neighbors, Data mining.

1. Introduction

Artificial intelligence in the form of machine learning makes it possible for programs to predict outcomes more accurately. The goal of machine learning is to create apps or models that can identify a model's accuracy by using data. The quality of information that the system receives from the outside world is a key factor in machine learning. Learning is an additional process that converts external data into knowledge, which is then stored in a repository. Several categorization methods are employed to balance the data and forecast future results. Machine learning (ML) focuses on the study of algorithms and statistical models that enable computers to perform tasks without explicit instructions. Instead, these systems learn from data and progressively improve their capabilities. To process data and generate predictions or judgments, machine learning (ML) uses a variety of algorithms. Machine learning algorithm like linear regression, support vector machine (SVM), logistic regression, random forest, decision tree, naïve bayes, k-nearest neighbors and so many algorithms use to classify the data. The use of machine learning is to work on different models that learn from a training set and define the accuracy

Heart disease, primarily heart attacks, is one of the leading causes of death globally, causing millions of deaths each year. Lowering death rates and improving the results for patients depend on early detection and prevention of heart attacks. Traditional risk assessment methods, which generally depend on statistical methods, find it difficult to handle the nonlinearities and complexity seen in clinical data. But new developments in machine learning (ML) offer strong answers to these problems by using large datasets to identify patterns that more conventional approaches could overlook.

2. Literature Review

S. Seema et al,[1] focuses on data mining techniques that can predict chronic disease. They used Decision tree, Naïve Bayes, Support Vector Machine(SVM) and Artificial Neural Network(ANN). From this experiment, SVM gives highest accuracy, whereas for diabetes Naïve Bayes gives the highest accuracy.

Ashok Kumar Dwivedi et al, [2] used different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms Sairabi H.Mujawar et al, [3] used k-means and naïve bayes to predict heart disease. This paper contain 13 independent variable for building the system. To extract knowledge from database, data mining techniques such as classification, clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

MeghaShahi et al, [4] suggested Heart Disease Prediction System using Data Mining Techniques. They used WEKA software for automatic diagnosis of disease and to give qualities of services in healthcare centres. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Boshra Brahmi et al, [5] developed different classification techniques like J48, Decision Tree, KNN, SMO and Naïve Bayes to evaluate the prediction and diagnosis of heart disease. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated and compared. J48 and decision tree gives the best technique for heart disease prediction.

Kumar Dwivedi [6] used SVM and KNN for heart disease prediction and found that SVM obtained 82% accuracy. Similar for that R. Sharmila et al, [7] used SVM and gave 85% accuracy and M. Kumari et al. [7] used SVM with 84.12% accuracy.

Moloud Abdar, et al (2015)[9] applied and compared five data mining techniques to predict the rise of heart disease such as C5.0, Naive Bayes, Support vector Machine, Logistic Regression and Multilayer Perceptron with accuracy measures as: 93.02, 86.05, 88.37,80.23 and 85.22 using 13 attribute independent medical variables.

G.Purusothaman and P.Kirshnakumari (2015)[10] had cited various data mining prediction models namely Decision Table, Association Rule, KNN, Artificial Multilayer Perceptron, Naïve Bayes and Hybrid models with accuracies as 76%, 55%, 58%, 85%, 69%, 86% and 96%. Hybrid data mining has outperformed other data mining heart disease diagnosing techniques.

K.Aravintan and Dr. M. Vanitha (2016)[11] did a comparative study on prediction of heart disease using and they were used Naive Bayes Algorithm, J48 Algorithm, Multilayer Perceptron Algorithm on 305 instances with 14 medical attributes. The accuracy measures of Naive Bayes, J48 and Artificial Multilayer Perceptron are as 81.3021%, 80.099%, and 82.56.

N Paras et al [12] used SVM and K-NN for the prediction of heart attack and demonstrated that K-Nearest Neighbors is much suitable and efficient for predicting the likelihood of heart attack prediction.

3. Methodology

Machine learning algorithms are used to predict the categorical outcomes. They can recognize patterns in labeled data and forecast fresh, unseen data. These models are frequently applied in many different domains, including disease prediction, spam detection, real/fake prediction and many more. Logistic regression, decision trees, random forests, support vector machines, and naive Bayes are examples of common classification techniques. The selection of algorithm depends on factors like the nature of the data, the complexity of the problem, and the desired performance metrics.

I. Dataset Source

This paper will use a CSV file dataset from the online repository Kaggle. The shape of the dataset is (300, 14) that contain attribute such as age, gender, chest pain etc.

II. DATASET AND MODEL DESCRIPTION

Data Gathering and Preparation

1. Data Gathering: Gather or collect the heart attack dataset from the specified repository.

2. Data Exploration: To comprehend the variables, missing values, and structure of the dataset, perform a preliminary analysis.

3. Feature Engineering: To enhance model performance, add new features or modify current ones as

needed (e.g., interaction terms, normalization).

4. Data Splitting: To construct a model, split the dataset into training and testing sets.

Model Selection and Training

1. Algorithm Selection: Based on the parameters of the dataset and the nature of the problem, select appropriate classification algorithms. Algorithms such as support vector machines, decision trees, random forests, and logistic regression, k-nearest neighbors.

2. Model Training: Train the selected model to learn patterns and relationships within the data.

Model Evaluation

1. Performance Metrics: Evaluate model performance using confusion metrics such as accuracy, precision, recall, F1-score analysis.

2. Comparison: Compare the performance of different models to identify the best-performing one.

Results and Discussion

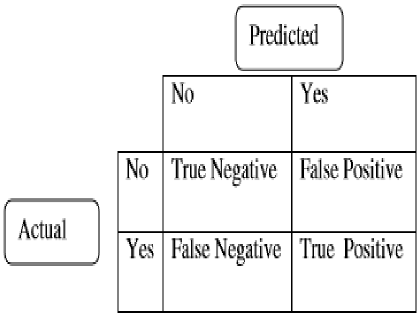
1. Interpretation: Discuss the ramifications of the findings after analyzing the data.

2. Limitations: Recognize the study's shortcomings, including those related to generalizability, model assumptions, and data quality.

3. Future Directions: Provide possible directions for further investigation, such as examining new features, enhancing model functionality, or resolving drawbacks.

4. Confusion Matrix & Measures Derived

Confusion matrix is a binary classification predicts test data sets as positive and negative, and they produce four outcomes : True positive, true negative, false positive, false negative(Fig 1).



		Predicted	
		No	Yes
Actual	No	True Negative	False Positive
	Yes	False Negative	True Positive

Fig 1: Confusion matrix [13]

Accuracy: The ratio of all accurate predictions to the entire number of true/false, positive/negative predictions is known as accuracy. One represents the best accuracy, and zero represents the lowest.

Accuracy = (True Positive + True Negative)

(True Positive + True Negative+ False Negative+ False Positive)

Sensitivity/ Recall/ True positive rate: Sensitivity is the ratio of true positive predictions and the total no. of Actual positives. The best Sensitivity is 1 and the worst is 0.

Sensitivity = True Positive

(True Positive + False Negative)

Precision: Precision is the ratio of true positive predictions and the total no.Of predictive positives**.** It is also called positive predictive value. The best precision value is 1 and the worst is 0.

Precision = True positive

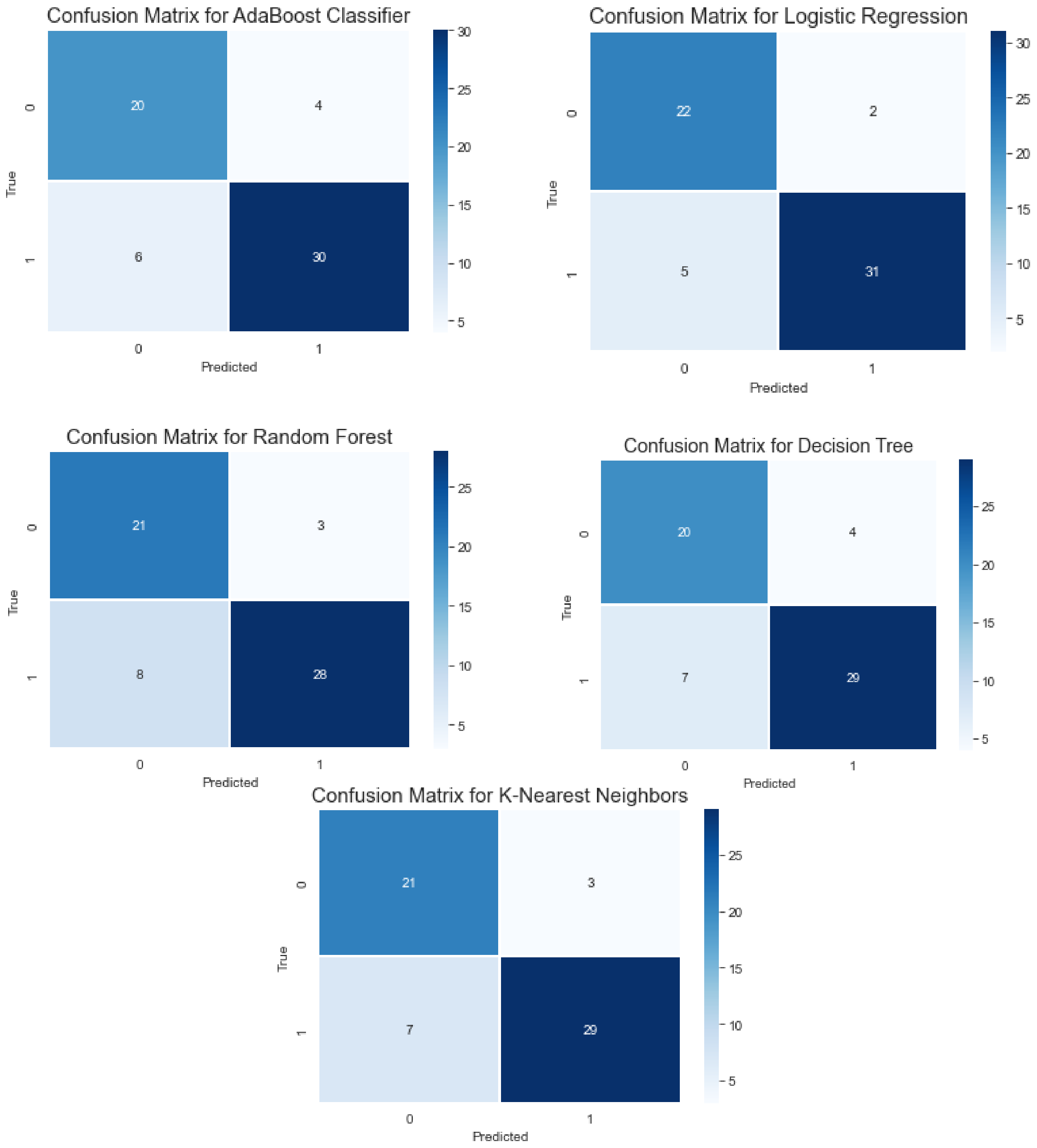
(True positive + False positive)

5. Conclusion

This paper examined different machine learning algorithms for the prediction of heart attacks. The objectives of our paper were to analyze the Heart Attack dataset by evaluating Machine Learning predictions algorithms. For this research paper, we used Decision Tree, Random Forest, Logistic Regression, K- Nearest Neighbors, Adaboost Classifier and we concluded that Logistic Regression has the highest accuracy rate for the prediction of heart attacks, with an Accuracy of 88 %. (Below result Table. 1 and Fig 2).

Algorithms	Accuracy	Precision	Recall	F1
Decision Tree	0.81	0.87	0.80	0.84
Random Forest	0.81	0.90	0.77	0.83
Logistic Regression	0.88	0.93	0.86	0.89
K- Nearest Neighbors	0.83	0.90	0.80	0.85
AdaBoost Classifier	0.83	0.88	0.83	0.85

Table 1 : Accuracy/Precision/Recall /F1 rate of different algorithms


Confusion Matrix for AdaBoost Classifier

Confusion Matrix for Logistic Regression

Confusion Matrix for Random Forest

Confusion Matrix for Decision Tree

Confusion Matrix for K-Nearest Neighbors

Fig 2: Confusion matrix of different Machine Learning Modals

6. Future Scope

- Future research focus should be enhancing these models' accuracy, transparency, and incorporation into clinical workflows should be the main goals of future research.
- Combining modals with other ML can create more robust and accurate predictive models.

- The accuracy of models can be improved by utilizing more data preprocessing techniques.
- Using More accurate training set can improve model performance.

7. References

1. Shedole, S. S., & Deepika, K. (2016). Predictive analytics to prevent and control chronic disease. Retrieved from https://www.researchgate.net/publication/316530782.
2. Dwivedi, A. K. (2016). Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation. Springer.
3. Mujawar, S. H., & Devale, P. R. (2015). Prediction of heart disease using modified K-means and Naïve Bayes. International Journal of Innovative Research in Computer and Communication, 3, pages.
4. Shahi, M., & Gurm, R. K. (2017). Heart disease prediction system using data mining techniques. Oriental Journal of Computer Science and Technology, 6, 457-466.
5. Brahmi, B., & Hosseini Shirvani, M. (2015). Prediction and diagnosis of heart disease by data mining techniques. Journals of Multidisciplinary Engineering Science and Technology, 2, 164-168.
6. Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29(10), 685-693.
7. Sharmila, R., & Chellammal, S. (2018). A conceptual method to enhance the prediction of heart diseases using data techniques. International Journal of Computer Science and Engineering.
8. Kumari, M., & Godara, S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction. International Journal of Computer Science and Technology, 2(2), 10-15.
9. Parvathi, I., & Rautaray, S. (2014). Survey on data mining techniques for the diagnosis of diseases in the medical domain. International Journal of Computer Science and Information Technologies, 5(1), 838-846.
10. .Purusothaman, G., & Krishnakumari, P. (2015). A survey of data mining techniques on risk prediction: Heart disease. Indian Journal of Science and Technology, 12, June.
11. Subha, R., Anandakumar, K., & Bharathi, A. (2016). Study on cardiovascular disease classification using machine learning approaches. International Journal of Applied Engineering Research, 11(6), 4377-4380.
12. Negi, P., & Bisht, M. (2022). Analysis and prediction of heart attack using machine learning models. In 7th International Conference on Computing, Communication and Security (ICCCS). https://doi.org/10.1109/ICCCS55188.2022.10079409
13. NBSHARE Notebooks. (n.d.). Confusion matrix diagram. NBSHARE. Available at https://www.nbshare.io/notebook/626706996/Learn-And-Code-Confusion-Matrix-WithPyton