Cardiovascular Diseases Prediction Analysis Using Machine Learning Algorithms

Safana Parveen P[1], Dr C Thiyagarajan[2]
[1]Dept of Computer Application
[2]Associate Professor, Dept of Computer Application
[1, 2] PSG College of Arts & Science, Coimbatore, India

Abstract- *Cardiovascular diseases(CVDs) remain a leading global cause of morbidity and mortality. Early detection and accurate risk assessment are crucial for effective prevention and intervention. This study presents a comprehensive machine learning framework for predicting heart diseases based on a wide range of clinical and demographic features. Information science assumes an urgent part in handling tremendous measures of information in the field of medical services. As coronary illness expectation is a complex undertaking, there is a need to robotize the expectation cycle to stay away from gambles related to it and alarm the patient well in advance. Several machine learning algorithms, including logistic regression, random forests, support vector machines, XGBoost, KNN, decision tree, and naive bayesare evaluated and compared using cross-validation to determine the most effective model to predict the accurate value.*

Keywords- Heart Attack, Cardiovascular disease, logistic regression, random forests, support vector machines, XGBoost, KNN, decision tree, and naive bayes.

I. INTRODUCTION

Cardiovascular infections (CVDs) keep on being a worldwide wellbeing challenge, liable for a critical extent of mortality and horribleness around the world. Timely and accurate prediction of heart disease risk is crucial for effective prevention and early intervention strategies. Machine learning, with its capacity to analyze complex datasets and uncover intricate patterns, has emerged as a promising tool for improving the prediction and diagnosis of heart diseases. This delves into the realm of machine learning-based approaches for predicting heart disease. It provides a comprehensive overview of the state-of-the-art methodologies, accuracy, prediction value and techniques employed in this critical domain. The convergence of healthcare and artificial intelligence has given rise to a plethora of studies aimed at harnessing the power of machine learning to enhance cardiovascular risk assessment. Throughout this review and explore the various machine learning algorithms commonly applied to heart disease prediction, including logistic regression, randomforests, support vector machines (SVM), XGBoost, k-Nearest Neighbors (KNN), decision tree, and naive bayes. By using Logistic regression, Random Forest we achieve a maximum of 94.51% accuracy value [1] in this review compared with others this gives the highest prediction percentage.

II. MACHINE LEARNING ALGORITHM CLASSIFICATIONS ANDMODELS

Machine learning models are computational algorithms and statistical techniques that enable computers to learn and make predictions or decisions without being explicitly programmed to do so. These models play a central role in various fields, including artificial intelligence, data science, and predictive analytics. Various types of classification algorithms are available for data analysis which are mentioned below:

2.1 Logistic Regression:

Logistic Regression is a well-known AI calculation utilized for parallel grouping issues. It belongs to the family of supervised learning algorithms, where the goal is to predict a binary output variable based on one or more input variables, also known as features.

The algorithm works by modelling the probability of the output variable belonging to the positive class given the input features. This is achieved by fitting a logistic function to the input data using a set of training examples with known output labels.

2.2 Random Forest:

Random Forest is a popular machine learning algorithm used for classification and regression tasks. A troupe technique joins numerous choice trees to make expectations. In irregular backwoods, numerous choice trees are made utilizing various subsets of the preparation information and highlights. This random sampling of data and features helps to reduce overfitting and improve the accuracy of the model.

2.3 XGBoost:

Regression and classification problems are common applications of the popular machine learning algorithm known as XGBoost (Extreme Gradient Boosting). It is an outfit learning strategy that consolidates numerous frail models to make a more grounded model. XGBoost works by training decision trees iteratively to fix mistakes made by previous models. The calculation allots higher loads to the misclassified relevant pieces of information and lower loads to the accurately arranged data of interest, which is known as slope supporting.

In addition to gradient boosting, XGBoost includes several key features that improve its performance, such as regularization, handling missing values, and parallel processing. The algorithm also uses a customized loss function that balances the trade-off between model accuracy and computational efficiency.

2.4 Naive Bayes:

Credulous Bayes is utilized for arrangement errands which depends on Bayes' hypothesis, which depicts the likelihood of an occasion happening given earlier information or proof.

Because Naive Bayes assumes that the input features are independent of one another, the probability of another feature being present does not change if one feature is present or not. The algorithm can function effectively even with high-dimensional data thanks to this assumption, which makes the calculations simpler.

Using Bayes' theorem, the probability of each class is calculated using the input features in the Naive Bayes algorithm:

P(C) is the prior probability of class C, P(xi|C) is the probability of feature xi given class C, P(xi|C) is the probability of feature xi given class C, and P(x1, x2,..., xn) is the probability of the input features. P(C|x1, x2,..., xn) = P(C) * P(x1|C) * P(x2|C) *...

2.5Support Vector Machine (SVM):

An artificial intelligence calculation known as Support Vector Machine (SVM) is used to examine order and relapse. It is a supervised learning algorithm that determines the optimal decision boundary between the various data classes.

A kernel function is used to map the data points in SVM to a higher-dimensional space. The best hyperplane for separating the various data classes can be found with the assistance of this transformation. As the decision boundary, the hyperplane with the greatest margin between classes is chosen.

2.6k-Nearest Neighbors:

The k-Closest Neighbors (k-NN) calculation is a straightforward and natural Machine Learning calculation utilized for both grouping and relapse undertakings. It is a sort of occasion based or languid learning calculation, meaning it doesn't fabricate a model during preparation however rather retains the whole preparation dataset. While making forecasts on new data of interest, it takes a gander at the "k" closest pieces of information from the preparation set and uses them to decide the result.

2.7 Decision Trees:

Decision Trees are a popular machine learning model used for both classification and regression tasks. They are a non-linear, supervised learning algorithm that can handle both categorical and numerical data. Decision Trees are known for their simplicity and interpretability, making them useful in various applications, including healthcare, finance, and marketing.

III. ANALYSIS OF CLASSIFICATION METHODS BASED ON THE PERFORMANCE

Analysis of Classification Methods Based on the Performance:

S.No	Author Name	Published Year	Classification Methods	Performance Obtained
1	Niloy Biswas, Md Mamun Ali, Md AbdurRahaman, Minhajul Islam, Md. Rajib Mia, Sami Azam, Kawsar Ahmed, Francis M. Bui, Fahad Ahmed Al-Zahrani and Mohammad Ali Moni	2023	Logistic regression, Decision tree, Random Forest, Support Vector Machine, Naive Bayes, KNN [1]	using Logistic regression, Random Forest achieve accuracy level of 94.51% [1]

2	Manjula P, Aravind U R, Darshan M V, Halaswamy M H, Hemanth E	2022	Decision tree, Logistic Regression, Naive Bayes, KNN, Random Forest, XGBoost, SVM [2]	Random Forest algorithm gives high accuracy score of 90.16% [2]
3	Rachna Jain, PreetiNagrath, Harshit Jindal, Sarthak Agrawal, Rishabh Khera	2021	Logistic regression, Random Forest and KNN [3]	KNN algorithm gives high accuracy of 88.52% [3]
4	Malavika G, Rajathi N, Vanitha V, Parameswari P	2020	Logistic Regression, KNN, Support Vector Machine, Decision Tree, Naive Bayes, Random Forest [4]	Random Forest gives high accuracy of 91.80% [4]
5	Neha Nandal, Lipika Goel, ROHIT TANWAR	2022	Support Vector Machines, Logistic Regression, Naïve Bayes and XGBoost [5]	XGBoost provides best accuracy of 92% among others. [5]
6	S. Mohan, C. Thirumalai and G. Srivastava	2019	KNN, Decision Trees, SVM, Random Forest, Genetic algorithm, and Naive Bayes [6]	Using hybrid random forest with a linear model (HRFLM) achieve accuracy level of 88.7% [6]
7	M. S. Amin, Y. K. Chiam, and K. D. Varathan	2019	kNN, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine, Neural Network.[7]	a hybrid technique with Naïve Bayes and Logistic Regression achieves an accuracy of 87.4% in heart disease prediction.[7]
8	C. Beulah ChristalinLatha, S. Carolin Jeeva	2019	Bayes Net, Naive Bayes, Random Forest, C4.5, Multilayer perceptron, PART [8]	using Bayes Net, Naive Bayes, Random forest, Multilayer perceptron achieve accuracy level of 85.48%[8]
9	K. Srinivas, B. K. Rani, and A. Govrdhan	2010	ODANB and Naïve Bayes methods are used in various datasets Heart-statlog [9]	Naïve Bayes observes better results with accuracy of 83.70% [9]
10	ApurbRajdhan ,Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli	2020	Decision Tree, Logistic Regression, Random Forest, and Naive Bayes algorithms to predict heart disease using UCI machine learning repository dataset [10]	Random Forest algorithm is the most efficient algorithm with accuracy score of 90.16% [10]

IV. CONCLUSION

In this comprehensive review, we have explored the application of machine learning algorithms in the prediction of heart disease—a critical task with far-reaching implications for healthcare and public health. Heart disease remains a leading cause of morbidity and mortality worldwide, emphasizing the pressing need for accurate risk assessment and early intervention. The utilization of machine learning models has shown great promise in addressing this challenge. Through an extensive examination of the literature, we have observed a 94.51% accuracy level using logistic regression and random forest in Niloy Biswas et al [10] paper.

REFERENCES

[1] Niloy Biswas, Md Mamun Ali, Md AbdurRahaman, Minhajul Islam, Md. Rajib Mia, Sami Azam, Kawsar Ahmed, Francis M. Bui, Fahad Ahmed Al-Zahrani and Mohammad Ali Moni, "Machine Learning-Based Model to Predict Heart Disease in Early Stage Employing Different Feature Selection Techniques, " Hindawi, BioMed Research International Volume 2023, Article ID 6864343, 15 pages https://doi.org/10.1155/2023/6864343

[2] Manjula P, Aravind U R, Darshan M V, Halaswamy M H, Hemanth E, 2022, Heart Attack Prediction Using Machine Learning Algorithms, International Journal Of Engineering Research & Technology (IJERT) ICEI – 2022 (Volume 10 – Issue 11)https://www.ijert.org/heart-attack-prediction-using-machine-learning-algorithms

[3] Harshit Jindal et al, " heart disease prediction using machine learning algorithms," 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072 https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072

[4] Malavika G, Rajathi N, Vanitha V, Parameswari P, "Heart Disease Prediction Using Machine Learning Algorithms,"Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 11 (2020) Pp-24-27 https://bbrc.in/wp-content/uploads/2021/01/Galley-Proof-006.pdf

[5] Nandal N, Goel L and TANWAR R. Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis [version 1; peer review: 1 approved]. F1000Research 2022, https://doi.org/10.12688/f1000research.123776.1

[6] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707. https://ieeexplore.ieee.org/document/8740989

[7] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, vol. 36, pp. 82–93, 2019 [7] R. Roman, J. Zhou, and J. Lopez, "On the features and challengesof security and privacy in distributed internet of things," ComputerNetworks, vol. 57, no. 10, pp. 2266–2279, 2013.

[8] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, article 100203, 2019.

[9] K. Srinivas, B. K. Rani, and A. Govrdhan,"Applications of data mining techniques in healthcare and prediction of heart attacks," International Journal on Computer Science and Engineering, vol. 2, no. 2, pp. 250–255, 2010.

[10] ApurbRajdhan ,Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, "Heart Disease Prediction using Machine Learning", International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 04 (April 2020) https://www.ijert.org/heart-disease-prediction-using-machine-learning