IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction And Risk Analysis Using Machine Learning Techniques

[1]Omkar Singh, [2]Amit Kumar Pandey, [3]Sushilkumar Sarode, Sandeep Odichuya4

[1] 1HOD (Department of DATA SCIENCE),2,3,4PG Students

1,3,4Department of DATA SCIENCE, 2Department of IT,

Thakur College of Science and Commerce

Thakur Village, Kandivali (East), Mumbai-400101, Maharashtra, India

Abstract: Heart disease, encompassing various cardiovascular conditions, is a leading cause of death and disability globally, including heart, attacks, strokes, and other heart-related issues, with preventable risk factors such as unhealthy lifestyle choices playing a significant role. Heart disease (also known as cardiovascular disease / CVD) refers to any disease or disorder of the heart and blood vessels. Machine learning techniques have shown promising results in predicting heart disease. Our approach algorithms like Logistic regression, K-Nearest Neighbor, Random Forest, and Decision Tree were predicted using attributes such as age, chest pain, blood pressure, and cholesterol levels. By wielding breakthroughs in machine learning (ML), the key aim of this study is to design prognostic models for identifying cardiovascular disease. Predicting those at the highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature death. Applying these classification models can help to predict the chances of cardiovascular disease.

Keywords: Cardiovascular Disease, Heart Disease, K-Nearest Neighbor (KNN), Decision Tree, Logistic Regression, Random Forest.

I. INTRODUCTION

Cardiovascular malady (CVD) may be an infection that includes the heart or blood vessels. CVDs incorporate a lesson of maladies that incorporates:

Coronary course infections (e.g. angina, heart assault), heart disappointment, hypertensive heart malady, rheumatic heart illness, cardiomyopathy, arrhythmia, innate heart malady, valvular heart illness, carditis, aortic Aneurysms, fringe supply route illness, thromboembolic illness, and venous thrombosis.[3][4]. The paper looks at the utilization of machine learning models in anticipating heart malady by utilizing K-NEAREST NEIGHBOR (KNN), LOGISTIC REGRESSION, Choice TREE, and Arbitrary Woodland is assessed. This prescient Demonstration shows potential in supporting healthcare specialists in reducing heart malady-connected fatalities. The point is that the distinctive strategies examined classify heart maladies more successfully to accomplish better exactness and execution evaluation. Ponders have appeared that the Arbitrary Timberland Calculation accomplishes remarkable execution, coming to precision rates of up to 95% in Distinguishing potential heart malady cases. To decrease the number of individuals who die. From heart infection, we ought to do it rapidly and have a way to identify it. Machine learning (ML) has appeared to be viable in helping make choices and forecasts from the huge amount of data produced by the healthcare industry. K-nearest neighbor Calculations of machine learning are utilized to anticipate and classify persistent heart infections. Restorative conclusion is respected as an imperative however complicated assignment that has to be executed precisely and effectively.

Methodology

This research employed a methodology to
assess the effectiveness of different classification Techniques, including Logistic Regression, Random Forest, Decision Tree, KNN. These are the supervised learning algorithms used mainly in classification tasks.

Logistic Regression

Logistic Regression is a supervised Classification algorithm used for predictive analysis based on probability theory. It employs a logistic function to estimate probabilities, establishing a relationship between the dependent variable and one or more independent variables. By utilizing the sigmoid function as a cost function, logistic regression ensures that predictions remain Within a range of 0 to 1. The effectiveness of this algorithm heavily depends on how well the data is presented. Therefore, the key feature. The dataset is selected using recursive and backward elimination techniques to enhance the model's accuracy and performance.

Random Forest

Random Forest is a powerful learning algorithm used for both classification and regression tasks. It constructs multiple decision trees from different subsets of data and aggregates their predictions to improve accuracy and reduce variance. Unlike a single
decision tree, Random Forest delivers more reliable results by minimizing overfitting, Although it operates slower due to multiple tree computations, it effectively handles categorical variables but does not process null values. The accuracy of the model increases with the number of trees used. In the proposed approach, 75% of the data is utilized for training and 25% for testing, ensuring optimal model performance. The Random Forest Classifier is applied to train the model, and predictions are obtained from individual trees. Each outcome entropy is calculated separately, refining the final predictions for better reliability.

K – Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) algorithm identifies relationships between dataset values
and predictions by utilizing a non-parametric approach, meaning it does not rely on predefined parameters or assumptions about data distribution. Unlike traditional machine learning models that learn weights during training, KNN is considered a lazy classifier as it stores training data and performs computations during classification rather than training. The algorithm determines the category of a new data point by analyzing its proximity to existing classes. One of KNN's key strengths is its simplicity and effectiveness, particularly in handling noisy datasets. It can manage large
datasets efficiently while allowing flexible decision boundaries. The dataset is typically divided into training and testing sets using the train-test-split function, where the training set is used for model learning, and the testing set evaluates performance on real-world or unseen data. To enhance efficiency, techniques such as KD-Trees or Ball Trees can be employed for faster nearest-neighbor searches.

Decision Tree

A Decision Tree algorithm is a widely used machine learning technique for predicting outcomes based on a series of conditions. It works by breaking down complex decisions into simple steps, following a tree-like structure where each decision leads to further subdivisions until a final outcome is reached. This approach helps in making predictions by systematically analyzing available data and identifying the most significant patterns. The model follows a step-by-step process, learning from past examples to classify new instances accurately. By structuring decisions in an organized manner, the algorithm enhances prediction reliability and is especially useful for
an application that requires clear and interpretable
Decision-making.

The application of machine learning (ML) techniques in heart disease prediction has gained significant traction in recent years, with researchers exploring various algorithms and data types to enhance predictive accuracy and clinical utility. This literature review examines recent advancements in the field, focusing on the performance of different ML algorithms and their potential impact on cardiovascular health management. To increase the predictive accuracy of a machine learning algorithm, this study compares four algorithms, including KNN (K-Nearest Neighbor), Decision Tree, Random Forest and Logistic Regression. 14 attributes, including age, sex, cp, trestbps, chol, FBS, restecg, thalach, exang, oldpeak, slope, ca, thal, target are used.

The accuracy, recall, and precision of each algorithm are calculated to determine the most accurate model. ML is a sub-method used for difficult models and algorithms that grant themselves to prediction this is known as forecasting analytics. These analytical models allow researchers, data scientists, data engineers, and data analysts to "produce good repeatable choices and outcomes" and expose "hidden patterns" by studying ancient connections and courses in the data.

Data Sources:

The primary dataset used in this analysis is the Heart Disease Dataset from the UCI Machine Learning Repository [1]. This widely-used dataset contains 303 instances with 14 attributes.

Attribute	DESCRIPTION	DATA TYPE
AGE	Age of the patient in years	Numerical
SEX	Sex of the patient(0=female,1=male)	Categorical
CP	Chest pain type(0=typical angina, 1=atypical angina, 2=non-anginal pain,3=asymptomatic)	Categorical
TRESTBPS	Resting blood pressure (mm Hg)	Numerical
CHOL	Serum cholesterol (mg/dl)	Numerical
FBS	Fasting blood sugar > 120 mg/dl (1=true, 0=false)	Categorical
RESTECG	Resting electro cardio graphical results (0=normal, 1=ST-T wave abnormality, 2=probable or definite left ventricular hypertrophy)	Categorical
THALACH	Maximum heart rate achieved during exercise	Numerical
EXANG	Exercise-induced angina (1 = yes,0 = no)	Categorical
OLDPEAK	ST depression induced by exercise relative to rest	Numerical
SLOPE	The slope of the peak exercise ST segment (0=upsloping, 1=flat, 2=down sloping)	Categorical
CA	Number of major vessels coloured by fluoroscopy (0-3)	Numerical
THAL	A blood disorder called thalassemia (0=normal, 1=fixed defect, 2=reversible defect)	Categorical
TARGET	Presence of heart disease (0=np, 1=yes)	Categorical

1. AGE: The age of the patient in years.
2. SEX: The sex of the patient (1 = male; 0 = female).
3. CP: Chest pain type, which can take four values: typical angina, atypical angina, non-anginal pain, or asymptomatic.
4. TRESTBPs: The resting blood pressure (in mm Hg) of the patient.
5. CHOL: The serum cholesterol (in mg/dl) of the patient.
6. FBs: Fasting blood sugar (in mg/dl) greater

than 120 mg/dl or not (1 = true; 0 = false).
7. RESTECG: Resting electrocardiographic results, which can take three values: normal, having ST-T wave abnormality, or showing probable or definite left ventricular hypertrophy.
8. THALACH: Maximum heart rate achieved during exercise. exang: Exercise-induced angina (1 = yes; 0 = no). OLDPEAK: ST depression induced by exercise relative to rest.
9. Exercise Induced Angina (exang) : Whether the patient experienced chest pain during work.
10. ST Depression (oldpeak): A measure of how much the ST segment in an ECG drops (related to heart stress). Example: If oldpeak = 1.4, it means there was 1.4 mm of ST depression.
11. Number of Major Vessels Colored by Fluoroscopy (ca): Number of major blood vessels visible using X-ray imaging.
12. SLOPE: The slope of the peak exercise ST segment, which can take three values: upsloping, flat, or down sloping. CA: The number of major vessels (0-3) colored by fluoroscopy.
13. THAL: A blood disorder called thalassemia, which can take three values: normal, fixed defect, or reversible defect.
14. TARGET: The presence of heart disease (1 = yes; 0 = no).

Data Preprocessing and Feature Selection:

1. Missing values were done by replacing the null values with mode to maintain data integrity. Feature selection was performed to identify the most relevant attributes, which will eliminate redundant or less significant variables to reduce overfitting to the models.
2. Standardization was applied to normalize the data, ensuring all features had a uniform scale, which is essential for models sensitive to varying magnitudes.
3. Applying these techniques will ensure the performance of the model maintain accuracy and avoid overfitting.

Feature selection techniques, including correlation analysis, mutual information, and recursive feature elimination, were often employed to identify the most relevant attributes for prediction.
Model Training and Evaluation Approach The studies analyzed generally followed a structured methodology for training and evaluating machine learning models:
1. Data Splitting
The dataset was typically partitioned into 75% for training and 25% for testing to assess model performance on unseen data.
2. Cross-Validation
To enhance reliability, K-fold cross-validation (commonly with k=5 or k=10) was applied, ensuring that performance metrics were not biased toward a specific data split.
3. Hyperparameter Optimization
Techniques such as grid search and random search were implemented to fine-tune model parameters, maximizing predictive accuracy and minimizing Overfitting.

Result and Discussion

1. Evaluation of metrics
The confusion matrix, also known as an error matrix, provides a detailed breakdown of the model's classification performance by displaying correct and incorrect predictions for each class. It allows you to visualize the performance of an algorithm. To assess the effectiveness of the trained model, various evaluation metrics were utilized, with the confusion matrix being a primary tool. A confusion matrix also referred to as an error matrix, is a tabular representation used to evaluate the performance of a classification model by comparing its predicted outputs with the actual values from the test dataset. Accuracy, precision, and recall help evaluate the quality of classification models in machine learning.

Accuracy

Accuracy shows how often a classification ML model is correct overall. Accuracy is a metric that measures how often a machine learning model correctly predicts the outcome. You can calculate accuracy by dividing the number of correct predictions by the total number of predictions.
Accuracy = (TP+TN)/(TP+TN+FP+FN) Where,
True Positive (TP) – Observation is positive

and predicted to be positive.
False Positive (FP) –Observation is negative
but predicted to be negative.
True Negative (TN)-Observation is positive
and predicted to be negative.
False Negative (FN)-Observation is negative
but predicted to be negative.

Recall

Recall is a metric that measures how often a
The machine learning model correctly identifies
Positive instances (true positive) from all the actual positive samples in the dataset. You can calculate recall by dividing by number of true positives by the number of positive instances. The latter includes true positive (successfully identified cases) and false negative results (missed cases).
Recall = TP/(TP+FN)

Precision

Precision is a metric that measures how often
a machine learning model correctly predicts the positive class. You can calculate precision by dividing the number of correct positive predictions (true positives) by the total number
of instances, the model predicted a positive
(both true and false positives).
Precision = TP/(TP+FP)

Result

Negative The result of the 14 feature classification models indicates the predictive performance of each model in determining the existence of certain outcomes, such as heart failure or heart disease. After comparing several feature selection approaches, the KNN K-Nearest Neighbour, Random Forest algorithm outperforms the other algorithms in terms of accuracy. The research evaluates the dataset using three different algorithms, comparing their results, by this methodology the research achieves a prediction accuracy of 88.52% in determining whether a patient is suffering from a heart disease (0 denotes absence, and 1 denotes presence). The Random Forest and (KNN) K-Nearest Neighbor algorithms yield the highest accuracy of approximately 88.52% compared to another algorithm.

Accuracy (%)

LOGISTIC REGRESSION	85.25
K-NEAREST NEIGHBOR	88.52
DECISION TREE	83.61
RANDOM FOREST	83.61

Precision

LOGISTIC REGRESSION	0.85
K-NEAREST NEIGHBOR	0.84
DECISION TREE	0.73
RANDOM FOREST	0.89

Recall

LOGISTIC REGRESSION	0.82
K-NEAREST NEIGHBOR	0.93
DECISION TREE	0.86
RANDOM FOREST	0.86

	F1 Score
LOGISTIC REGRESSION	0.84
K-NEAREST NEIGHBOR	0.88
DECISION TREE	0.79
RANDOM FOREST	0.87

Conclusion

This study has evaluated multiple classification models for heart disease prediction and comparing all the models at last to see which algorithm has performed the best. The results indicate that Random Forest (Before Tuning) achieves the highest accuracy, making it the most effective model for this dataset. K-Nearest Neighbor (KNN) also performs well, followed closely by Logistic Regression, demonstrating the reliability of these models for classification tasks in medical data analysis.

While Decision Tree (After Tuning) shows a notable improvement compared to its initial state, it still underperforms relative to Random Forest. Interestingly, Random Forest (After Tuning) does not surpass its pre-tuned counterpart, suggesting that the default parameters may already be well-optimized for this dataset.

Future exploration can be enhanced by using deep learning methods and better feature selection for more larger and real-world datasets.

References

1. Yadav, Shivani. "Heart Disease Prediction Using Machine Learning." INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (2024): n. pag.

2. Rufes, Patricia et al. "Heart Disease Prediction Using Machine Learning." International Research Journal on Advanced Engineering Hub (IRJAEH) (2024): n. pag.

3. Bhajibhakare, Prof. M. M.. "Heart Disease Prediction using Machine Learning." International Journal for Research in Applied Science and Engineering Technology (2019): n. pag.

4. Ramalingam, V V & Dandapath, Ayantan & Raja, M. (2018). Heart disease prediction using machine learning techniques: A survey. International Journal of Engineering & Technology. 7. 684. 10.14419/ijet.v7i2.8.10557.

5. Omkar Singh, Amit Pandey, Rishidev Mishra, Pooja Pandey, "HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.12, Issue 3, pp.g356-g359, March 2024,

6. Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease. 2015 Sep;7(1):129-37.

7. Puneet Misra, "MACHINE LEARNING ALGORITHMS FOR OPTIMISING HEART DISEASE PREDICTION THROUGH HYPERPARAMETER TUNING", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.5, Issue 1, pp.310-315, January 2017

8. Ravindra yadav, Upendra singh, "SURVEY ON HEART DISEASE PREDICTION BY USING MACHINE LEARNING TECHNIQUE ", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.7, Issue 1, pp.44-51, March 2019