iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62085

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Analysis of Machine Learning Algorithms for Heart Disease Prediction

Prof. Rupatai Lichode[1], Shafiya Siddique[2], Vaishnavi Sriramula[3], Bhoomika Vishwakarma[4], Jyoti Yadav[5]
Computer Science and Engineering, Dr. Babasaheb Ambedkar Technological University, Lonere

Abstract: ***In recent times, heart disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This project makes use of heart disease clinical dataset available in UCI machine learning repository. The proposed Work predicts the chances of heart disease and classifies patient's risk level by implementing different data Mining techniques such as Decision Tree, Logistic Regression, support vector machine, k-Nearest Neighbor, Random Forest and Gradient Boosting. Thus, this paper Presents a comparative study by analyzing the performance of different machine learning algorithms. With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to Predict heart diseases effectively and accurately. The Aim of the study was to find the most efficient ML Algorithm for detection of heart diseases. This study compares the accuracy score of all above mentioned machine learning algorithms for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 85.24% for prediction of heart disease when Compared to all other Classification algorithms used in this analysis which will help to provide better Results and help health professionals in predicting the heart disease effectively and efficiently.***

Keywords: Machine Learning, Random Forest, Clinical Dataset, Prediction, Classification Algorithms, Data Science.

I. INTRODUCTION

Heart Diseases have shown a tremendous hit in this modern age. As doctors deal with precious human life, it is very important for them to be perfect in their results. Thus, an application was developed which can predict the vulnerability of heart disease, given basic symptoms like age, gender, pulse rate, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, exercise induced angina, ST depression ST segment the slope at peak exercise, number of major vessels colored by fluoroscopy and maximum heart rate achieved. This can be used by doctors to recheck and confirm on their patient's condition. In the existing surveys they have considered only 10 features for prediction, but in this proposed research work 14 necessary features are taken into consideration. Also, this paper presents a comparative analysis of machine learning techniques like Random Forest (RF), Logistic Regression, Support Vector Machine (SVM) in the classification of cardiovascular disease. By the comparative analysis, machine learning algorithm Random Forest has proven to be the most accurate and reliable algorithm and hence used in the proposed system. Coronary illness has the biggest level of passing on the planet. In 2012, around 17.5 million individuals kicked the bucket from coronary illness, implying that it comprises of the 31% of every single worldwide passing. Besides, coronary illness loss of life rises each year, It is relied upon to develop more than 23.6 million by 2030. The exploration from the January 2017 demonstrated that the main source of death worldwide is cardiovascular infections. The cardiovascular malady is considered as a world's biggest killer and is currently taking the top position in the record of ten reasons for passing in the previous 15 years and in 2015 was numeration for fifteen million passing. Various human lives could be spared by diagnosing on schedule. Along these lines, diagnosing the syndrome is significant and an exceptionally muddled undertaking. Mechanizing this procedure would conquer the issues with the diagnosis. The utilization of AI in ailment arrangement is normal and researchers are especially fascinated in the advancement of such frameworks for simpler following and analysis of cardiovascular diseases. Since ML permits PC projects to ponder from information, building up a model to perceive ordinary examples and having the option to settle on choices dependent on assembled data, it doesn't have hitches with the deficiency of utilized medicinal database. The proposed model is to amass significant information relating all components identified with coronary illness and parameters impacting it, train the information according to the proposed calculation of AI and foresee how solid is there a probability for a patient to get a coronary illness. The relationship with the diabetes related credits is considered to set up the impact.

The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. One of the main causes of illness and death among the world's population is heart disease. The prediction of cardiovascular illness is one of the most important subjects in the field of data analysis. The prevalence of cardiovascular disease has been rapidly increasing worldwide since a few years ago. Numerous research have been conducted in an effort to pinpoint the most crucial heart disease risk factors and precisely calculate the overall risk. Because it results in death without any overt symptoms, heart disease is sometimes known as the "silent killer." The ability to make decisions about lifestyle changes for high-risk individuals significantly depends on the early detection of cardiac disease, which reduces consequences. The vast amount of data produced by the healthcare industry has made machine learning an effective tool for prediction and decision-making. By evaluating patient data that uses a machine-learning algorithm to categorise whether a patient has heart disease or not, this study hopes to predict future cases of heart disease. Machine learning methods can be extremely helpful in this situation. There is a common set of basic risk factors that determine whether or not someone will ultimately be at risk for heart disease, despite the fact that heart disease can manifest itself in various ways. We may say that this technique can be very well adapted to accomplish the prediction of heart disease by gathering the data from many sources, classifying them under appropriate algorithms, and then analysing using the different dataset.

II. METHODOLOGY

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression, Random Forest Classifiers, etc which can be helpful for practitioners or medical analysts to accurately diagnose Heart Disease. This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model. The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology (Figure 1.) includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the preprocessing stage where we explore the data. Data preprocessing deals with the missing values, cleaning of data and normalization depending on algorithms used. After pre-processing of data, classifier is used to classify the pre-processed data. The classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier, etc. Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective Heart Disease Prediction System (EHDPS) has been developed using different classifiers. This model uses 14 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction

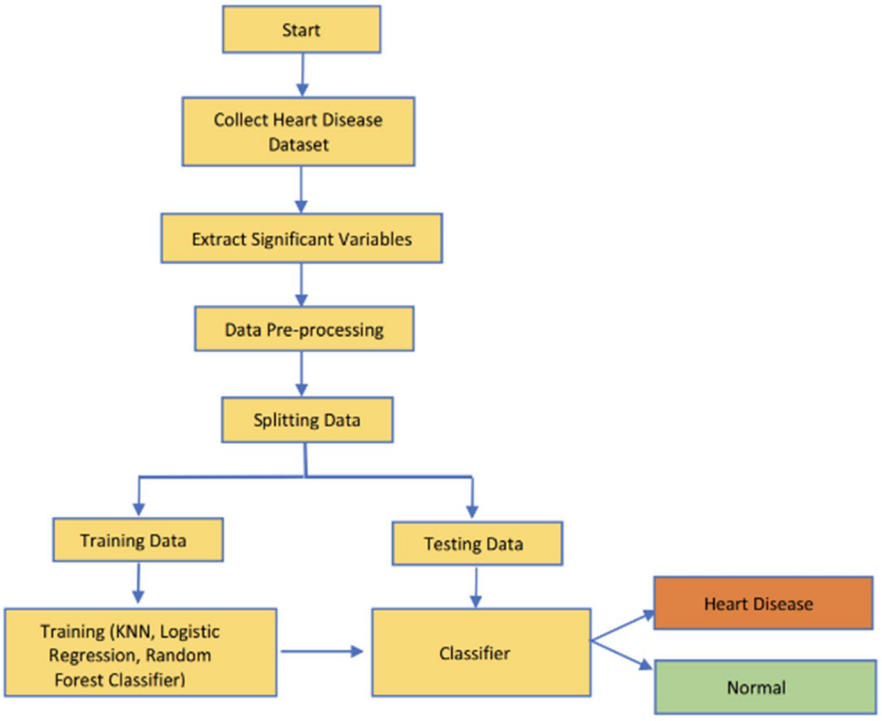


Figure 1. Proposed Model

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

The methodology of the code implementation is:

1) Importing the Libraries

1. Importing the Libraries

```
In [1]: import pandas as pd
```

Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data. Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant.

2) Importing the Dataset

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

3) Taking Care of Missing Values & Taking Care of Duplicate Values

```
In [3]: data.isnull().sum()

Out[3]: age         0
        sex         0
        cp          0
        trestbps    0
        chol        0
        fbs         0
        restecg     0
        thalach     0
        exang       0
        oldpeak     0
        slope       0
        ca          0
        thal        0
        target      0
        dtype: int64
```

Part of a data analysis or machine learning project is checking for missing values in the DataFrame using the .isnull() method, and then summing up the count of missing values for each column using the .sum() method.

data.isnull().sum() essentially gives you the count of missing values for each column in the DataFrame data. This is a common operation in data preprocessing to identify and handle missing data appropriately before analysis or modeling.

Data cleaning is the method of preparing a dataset for machine learning algorithms. It includes evaluating the quality of information, taking care of missing values, taking care of outliers, transforming data, merging and deduplicating data, and handling categorical variables. This basic process is required to ensure if the information is ready for machine learning algorithms, as it helps to diminish the hazard of blunders and enhances the accuracy of the models.

Data merging and deduplication in machine learning is the method of combining two or more datasets into one and expelling any duplicate data points. Usually done to guarantee that the information utilized to construct the machine learning models is accurate and complete. Data merging includes combining datasets to preserve the integrity of the information, whereas deduplication includes recognizing and evacuating any duplicate data points from the dataset.

4) Data Preprocessing

The columns in a DataFrame data are categorized into two lists based on the number of unique values they contain:

cate_val: This list contains column names where the number of unique values is less than or equal to 10. These columns are likely categorical or discrete variables.

cont_val: This list contains column names where the number of unique values is greater than 10. These columns are likely continuous variables.

Data Processing is a task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative. Using Machine Learning algorithms, mathematical modelling and statistical knowledge, this entire process can be automated.

Data preprocessing is a crucial step in building machine learning models for heart disease prediction. Here's a general outline of the steps involved:

Data Collection: We Gathered a dataset containing relevant features (such as age, gender, cholesterol levels, blood pressure, etc.) and the target variable (whether the individual has heart disease or not).

Data Cleaning: Checked for missing values, outliers, and inconsistencies in the dataset. Handled missing values by imputation (replacing them with mean, median, or mode of the column) or removing the corresponding rows/columns if they are too significant.

Feature Selection/Extraction: Identify the most relevant features for predicting heart disease. This can involve techniques like correlation analysis, feature importance ranking, or domain knowledge.

Data Transformation: The data is transformed.

Normalization/Standardization: Scale the numerical features to a similar range to prevent some features from dominating others. Normalization scales the data between 0 and 1, while standardization scales it to have a mean of 0 and a standard deviation of 1.

One-Hot Encoding: Convert categorical variables into a binary format to make them usable for machine learning algorithms.

Handling Imbalanced Data: If the dataset is imbalanced (i.e., one class is significantly more frequent than the other), employ techniques such as oversampling, undersampling, or synthetic data generation to balance the classes.

Feature Engineering: Create new features from existing ones if necessary. For example, you might calculate the body mass index (BMI) from height and weight.

Splitting the Data: Divide the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance.

Feature Scaling: Scale the features if necessary to ensure that they have similar ranges, which can help improve the performance of certain machine learning algorithms (e.g., SVM, k-NN).

Dimensionality Reduction (optional): If dealing with high-dimensional data, consider techniques like Principal Component Analysis (PCA) or feature selection algorithms to reduce the number of features while retaining most of the information.

Cross-Validation: Use techniques like k-fold cross-validation to assess the model's performance more accurately and reduce overfitting.

Data Augmentation (optional): If the dataset is small, consider augmenting it with synthetic data to improve model generalization.

Final Data Checks: Before feeding the data into the model, perform a final check to ensure that everything is in the correct format and no errors are present.

5) Splitting The Dataset Into The Training Set And Test Set

The dataset is typically divided into two subsets:

Training Data: This subset of the dataset is used to train the machine learning model. It consists of input data (features) and corresponding output labels (target variable). The model learns patterns and relationships from this data.

Test Data: This subset of the dataset is used to evaluate the performance of the trained model. It also consists of input data and corresponding output labels, but it is kept separate from the training data. The model makes predictions on this data, and the actual output labels are compared against the predicted labels to assess how well the model generalizes to new, unseen data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

The purpose of splitting the dataset into training and test sets is to assess the model's performance on data it hasn't seen during training. This helps to detect overfitting, where the model learns to memorize the training data rather than generalize from it.

```
In [22]: from sklearn.model_selection import train_test_split
```

In this method, From sklearn.model_selection import train_test_split: This line imports the train_test_split function from the scikit-learn library, which is used to split the dataset into training and testing sets. The train_test_split function from scikit-learn is commonly used to split the dataset into training and test sets.

6) Logistic Regression

Logistic regression is an important technique in the field of artificial intelligence and machine learning (AI/ML). ML models are software programs that you can train to perform complex data processing tasks without human intervention. ML models built using logistic regression help organizations gain actionable insights from their business data. They can use these insights for predictive analysis to reduce operational costs, increase efficiency, and scale faster.

Logistic Regression is a statistical and machine-learning technique classifying records of a dataset based on the values of the input fields. It predicts a dependent variable based on one or more sets of independent variables to predict outcomes. It can be used both for binary classification and multi-class classification.

7) SVM

SVM can be imagined as a surface that maximizes the boundaries between various types of points of data that is represented in multi-dimensional space also known as hyperplane. It can be used for both binary classification and multi-class classification.
using scikit-learn to train a Support Vector Machine (SVM) classifier on some data. SVM stands for Support Vector Machine are a set of supervised learning methods used for classification, regression and outliers detection. After training the model, you're making predictions on the test set and calculating the accuracy score.

8) KNeighbors Classifier

KNeighborsClassifier is used to instantiate the kNN classifier. You can specify the number of neighbors through the n_neighbors parameter.
We then fit the classifier to the training data using the fit method.
After that, we make predictions on the test data using the predict method.
Finally, we calculate the accuracy score by comparing the predicted labels with the true labels of the test data.

9) Decision Tree Classifier

A Decision Tree is a supervised machine learning algorithm used for classification and regression tasks. It creates a model that predicts the value of a target variable based on several input features. The algorithm splits the dataset into subsets based on the values of one feature at a time, aiming to maximize the homogeneity of the target variable within each subset. This process is repeated recursively until the subsets either contain data points belonging to the same class or reach a maximum specified depth.
Decision Tree Classifier: A Decision Tree Classifier is specifically used for classification tasks. It builds a decision tree model that predicts the class label of a data point based on its input features.

10) Random Forest Classifier

Random Forest is an ensemble learning method used for classification and regression tasks. It builds multiple decision trees during training and combines their predictions to improve accuracy and reduce overfitting.
Random Forest Classifier: Random Forest Classifier is an ensemble learning method specifically used for classification tasks. It consists of a collection of decision trees, where each tree is trained on a random subset of the training data and a random subset of the features. Random Forests are robust against overfitting because they combine multiple decision trees, each trained on different subsets of the data.
They are less sensitive to noisy data and outliers compared to individual decision trees. Random Forests provide feature importance scores, which can be useful for feature selection and understanding the data.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

11) Gradient Boosting Classifier

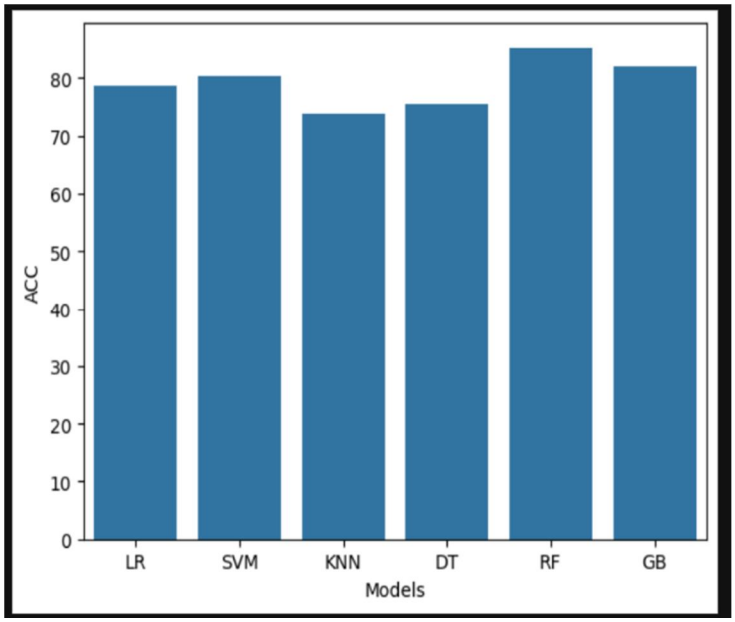
The Gradient Boosting Classifier is another ensemble learning method, like Random Forest, commonly used for classification tasks. However, instead of building multiple trees independently and combining their predictions, gradient boosting builds trees sequentially, with each new tree trying to correct errors made by the previous ones.

Gradient Boosting can capture complex patterns in the data by sequentially refining the model. It typically achieves high predictive accuracy, often outperforming other algorithms on structured/tabular data. It handles both numerical and categorical features without the need for one-hot encoding.

12) Finding Best Algorithm

[67]:

	Models	ACC
0	LR	78.688525
1	SVM	80.327869
2	KNN	73.770492
3	DT	75.409836
4	RF	85.245902
5	GB	81.967213



Random Forest is indeed one of the commonly used algorithms for heart disease prediction, and it often performs quite well in this domain. It has proved highest accuracy algorithm in our project showing following benefits:

Accuracy: Random Forest tends to have high accuracy in both classification and regression tasks. Since heart disease prediction is typically a binary classification problem (presence or absence of heart disease), Random Forest can effectively learn the complex relationships between various risk factors and the likelihood of heart disease.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

Robustness to Overfitting: Random Forest is less prone to overfitting compared to some other algorithms, such as decision trees. This is because it aggregates the predictions of multiple decision trees, which helps reduce variance and improve generalization to unseen data.

Handling of Nonlinear Relationships: Heart disease prediction often involves complex and nonlinear relationships between risk factors (e.g., age, cholesterol levels, blood pressure) and the presence of heart disease. Random Forest is capable of capturing these nonlinear relationships, making it suitable for such tasks.

Feature Importance: Random Forest provides a feature importance measure, which can help identify the most important risk factors contributing to heart disease prediction. This can be valuable for understanding the underlying factors driving the predictions and for medical interpretation.

Ease of Use and Interpretability: Random Forest is relatively easy to use and requires minimal hyperparameter tuning. Additionally, the ensemble nature of Random Forest allows for some level of interpretability by examining individual decision trees within the forest.

That said, while Random Forest is often a good choice, the "best" algorithm can depend on various factors including the specific characteristics of the dataset, the size of the dataset, computational resources, and the desired balance between interpretability and predictive performance. Therefore, it's always a good practice to experiment with multiple algorithms and evaluate their performance on the specific task at hand before making a final decision.

13) Prediction on New Data

In [72]:
```
import pandas as pd
```

In [73]:
```
new_data = pd.DataFrame({
    'age':52,
    'sex':1,
    'cp':0,
    'trestbps':125,
    'chol':212,
    'fbs':0,
    'restecg':1,
    'thalach':168,
    'exang':0,
    'oldpeak':1.0,
     'slope':2,
    'ca':2,
    'thal':3,
},index=[0])
```

In [74]:
```
new_data
```

Out[74]:

	age	sex	cp	trestbps	chol	fbs	restecg
0	52	1	0	125	212	0	1

In [75]:
```
p = rf.predict(new_data)
if p[0]==0:
    print("No Disease")
else:
    print("Disease")
```

```
No Disease
```

Prediction on new data in machine learning refers to the process of using a trained machine learning model to make predictions or estimations on data points that were not part of the training set. Once a machine learning model is trained on a dataset, it learns patterns and relationships within that data to make predictions on new, unseen data.

The process typically involves feeding the new data points into the trained model, which then applies the learned patterns to generate predictions or classifications for the new data. This prediction can take various forms depending on the type of machine learning task, such as regression (predicting a continuous value), classification (predicting a category or class label), or clustering (grouping data points into clusters).

The performance of the model on new data can be evaluated using metrics such as accuracy, precision, recall, or mean squared error, depending on the specific task and the nature of the data. It's important to note that the quality of predictions on new data depends heavily on the quality of the training data, the chosen model architecture, and the tuning of model parameters.

14) Save Model Usign Joblib

```
In [76]: import joblib
```

joblib is a library in Python that provides utilities for saving and loading Python objects (including NumPy arrays and other data structures) to and from disk. It's particularly useful for saving trained machine learning models, as well as large NumPy arrays efficiently.

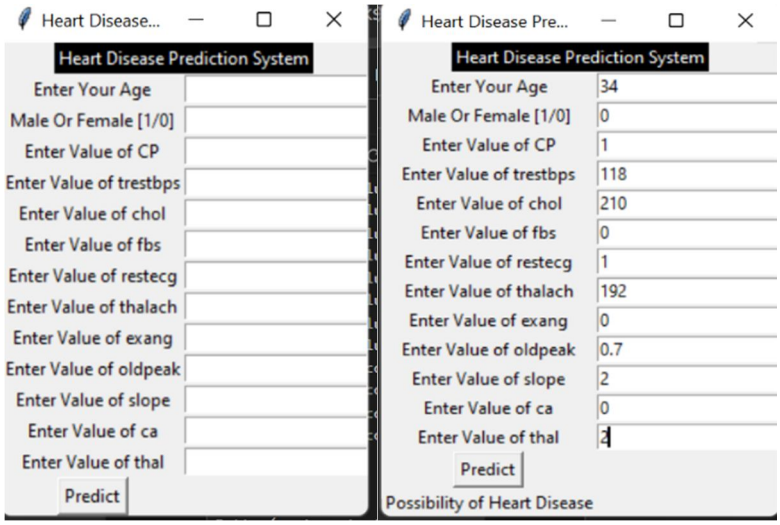
We trained a Random Forest Classifier on the this dataset. After training, we save the trained model to disk using joblib.dump. We then load the saved model using joblib.load. Finally, we make predictions using the loaded model and evaluate its performance. joblib is particularly efficient for large NumPy arrays and can serialize objects much faster than the standard pickle library. It's commonly used in the machine learning community for saving and loading trained models.

15) Creating GUI

from tkinter import *

import joblib

Tkinter is a standard GUI (Graphical User Interface) library in Python, commonly used for creating desktop applications with graphical interfaces.



III. RESULTS

In our project, we explored various machine learning algorithms for the task of heart disease prediction. We focused on six key algorithms: logistic regression, decision tree, K-nearest neighbors (KNN), support vector machine (SVM), gradient boosting, and random forest. Our aim was to identify the algorithm or combination of algorithms that provide the highest accuracy in predicting the likelihood of heart disease. We got result as random forest.

In our project, we observed that ensemble methods like random forest and gradient boosting generally outperformed individual algorithms in terms of accuracy. However, logistic regression and decision trees also provided valuable insights and competitive performance. Therefore, a combination of these algorithms or ensemble methods could be a prudent approach for heart disease prediction, balancing accuracy, and computational efficiency.

IV. CONCLUSION

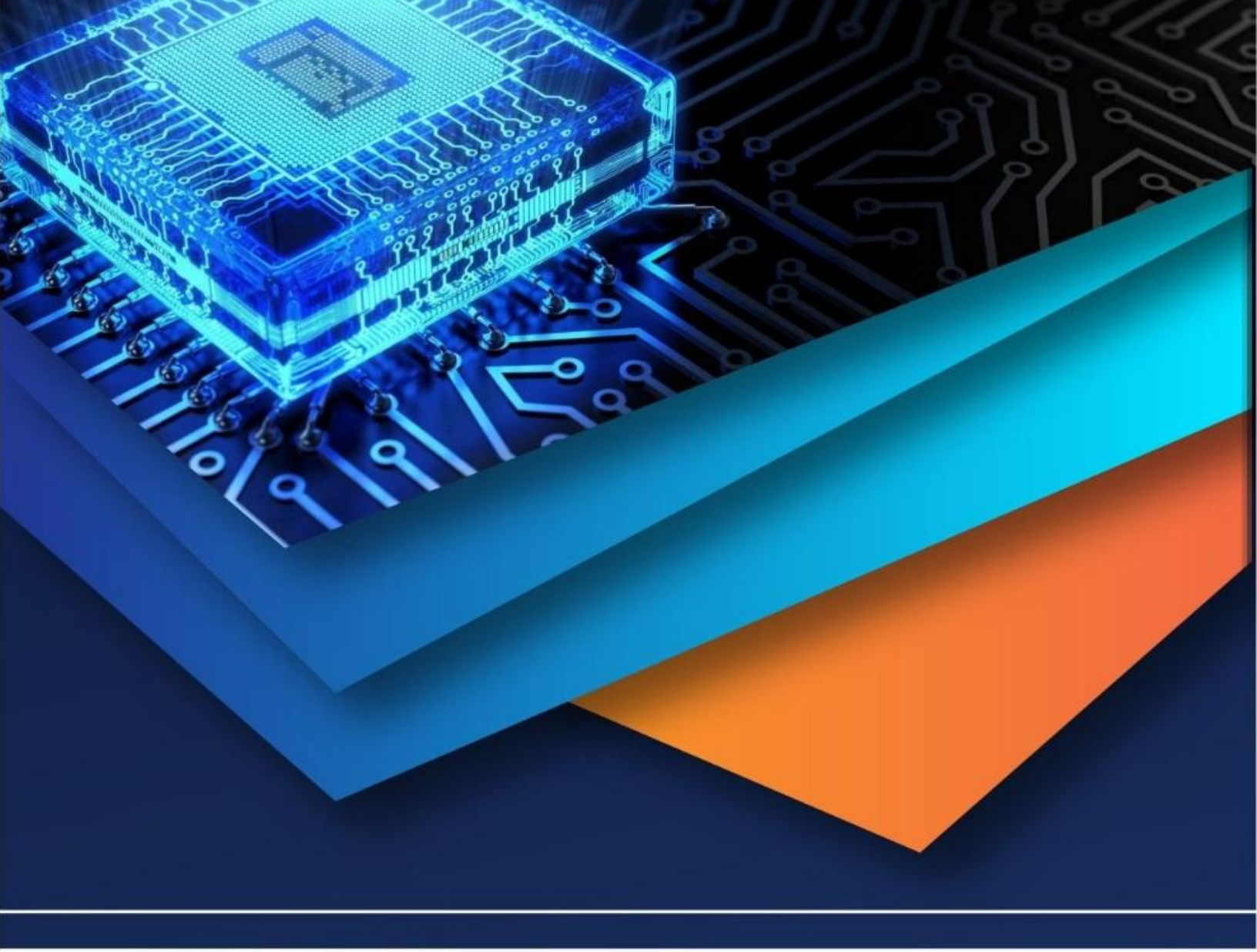
The primary reason for conducting this study is to propose a model for predicting the development of heart disease. Additionally, the goal of this research is to determine the optimum classification method for detecting the likelihood of cardiac disease. Six classification algorithms, namely Logistic Regression, Decision Tree, and Random Forest, etc are employed at various levels of evaluations in a comparative study and analysis to support this work. Although these machine learning methods are widely utilised, predicting cardiac disease is a crucial task requiring the highest level of accuracy. Consequently, a variety of levels and assessment strategy types are used to evaluate these algorithms. This will enable scientists and medical professionals to create a better world.

Making forecasts and diagnosing ailments has never been simple for medical professionals when it comes to heart conditions. Due to this, people can take the necessary action to treat heart disease before it gets worse if it is discovered in its early stages anywhere in the world. The three main causes of heart disease-drinking alcohol, smoking cigarettes, and not exercising-have become serious issues in recent years. The health care industry has produced a substantial amount of data over time, which has made machine learning capable of providing effective outcomes in prediction and decision-making.

Our goal in this study is to identify the best factors that can improve the prediction accuracy of heart disease and finding the most effective variables to raise the accuracy of heart disease prediction. Evaluation criteria, namely accuracy, specificity, sensitivity, and area under the ROC curve, are employed to verify the efficacy of the proposed approach on a public dataset comprising patients of both genders. The primary benefits of applying machine learning for heart disease prediction are that it reduces the complexity of the doctors time, is patient- and cost-friendly, and manages the largest (enormous) amount of data through feature selection and the random forest algorithm. Early diagnosis of cardiovascular disease can help with lifestyle modifications for high-risk patients, which can lower complications and be a significant medical milestone.

REFERENCES

[1] A comparative study on heart disease prediction using machine learning techniques" by J. V. Eswari, M. Hemalatha, and S. Indumathi, published in the International Journal of Computer Applications in 2017.
[2] "Heart Disease Prediction System Using Machine Learning" by Nikhil Kumar Singh, V. K. Jain, and Manoj Diwakar, published in the International Journal of Scientific Research in 2017.
[3] "Prediction of heart disease using machine learning algorithms" by Bharath Bhushan Natarajan, Navjyoti Singh, and Karthik Balasubramanian, published in the International Journal of Engineering Technology Science and Research in 2016.
[4] "Comparative Study of Machine Learning Algorithms for Predictive Analysis of Heart Disease" by R. P. Santosh Kumar, M. Ravi Teja, and G. Lavanya, published in the International Journal of Advanced Research in Computer and Communication Engineering in 2017.
[5] "Comparison of Data Mining Techniques for Predictive Modeling of Heart Disease" by Priti Chandra, Tanupriya Choudhury, and Prashant Singh Rana, published in the International Journal of Computer Applications in 2012.
[6] "Heart Disease Prediction Using Machine Learning Algorithms" by Ankita Shukla and Bharti W. Gawali, presented at the 2019 International Conference on Communication and Electronics Systems (ICCES).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY