iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** III **Month of publication:** March 2022

DOI: https://doi.org/10.22214/ijraset.2022.40768

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 10 Issue III Mar 2022- Available at www.ijraset.com

Prediction of Heart Disease Using Machine Learning Algorithms

Shriniket Dixit[1], Pilla Vaishno Mohan[2], Shrishail Ravi Terni[3]
[1, 2, 3]*Student, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India.*

Abstract: ***In living organisms, the heart plays an important function. Diagnosis and prediction of heart diseases necessitates greater precision, perfection, and accuracy because even a minor error will result in fatigue or death. There are multiple death cases related to the heart, and the number is growing rapidly day by day. The scope of this study is restricted to discovering associations in CHD data using three super- vised learning techniques: Logistic Regression, K-Nearest Neighbour, and Random Forest, in order to improve the prediction rate. As a result, this paper conducts a comparative analysis of the results of various machine learning algorithms. The trial results verify that Logistic Regression algorithm has achieved the highest accuracy of 89% com- pared to other ML algorithms implemented.***

Keywords: Machine Learning, Logistic Regression, K-Nearest Neighbour, Random Forest, Python, Heart Disease, Prediction model, Healthcare

I. INTRODUCTION

Heart disease has risen to become one of the leading causes of death all over the world. Ac- cording to the World Health Organization, cardiac illnesses claim the lives of 17.7 million people each year, accounting for 31% of all fatalities worldwide. Heart disease has become the top cause of death in India as well. As a result, it is essential to be able to forecast heart-related disorders in a reliable and precise manner. Data on various health-related con- cerns is compiled by medical institutions all over the world. These data can be used to gain significant information utilizing a variety of machine learning techniques. However, the amount of data collected is enormous, and it is frequently noisy.

II. PROBLEM-STATEMENT

We analyzing the various machine learning algorithms and finding the best to predict the presence or absence of heart disease. The tar- get we will be exploring is binary classification which is 0 to show the absence of heart disease and 1 to show the presence of heart disease.

III. PROPOSED METHOD

We are going to use various machine learning algorithms to predict the target. We will be using a number of different features about a person to predict whether they have heart dis- ease or not. The dependent variable is whether or not a patient has heart disease, while the independent variables are the patient's many medical characteristics. The various machine learning algorithms used for our model will be Logistic Regression, K-Nearest Neighbours, and Random Forest. We will compare the scores of all these models by splitting our data into training and testing in an approximate 80:20 ratio. We will also tune the hyper parameters for all these models to yield the best results. And finally conclude the best prediction model for our heart disease dataset.

Flow Diagram

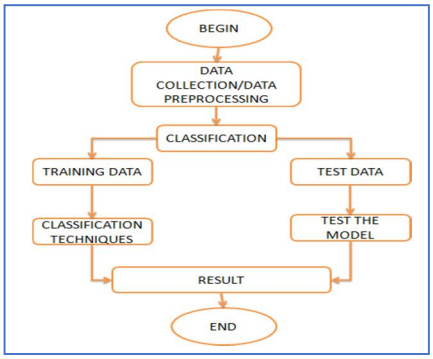


Fig 1-Flow diagram

IV. LITERATURE SURVEY

1) [Archana singh, Rakesh Kumar (2020) heart disease prediction using machine learning algorithms in this particular publication various machine learning algorithms such as linear regression, decision tree, support vector ma- chine and k- nearest neighbour is used. When they performed the analysis of algorithms on the basis of the dataset whose attributes are shown in a research paper and on the basis of the confusion matrix, they found KNN is the best one. [7.]SVM has an accuracy of 87%, decision tree of 79%, and k-nearest neighbour has an accuracy of 74%. For the future scope more machine learning approach will be used for the best analysis of heart dis- eases and for earlier prediction of diseases so that the rate of a number of deaths can be reduced if people are informed of the illness.

2) Jaymin Patel, Prof.Tejalupadhyay, Dr. Samir Patel (2016) used machine learning and data mining techniques to predict cardiac dis- ease. In this research paper, they have analyzed the experimental results, it is concluded thatj48 tree technique turned out to be the best classifier for heart disease prediction because it contains more accuracy and the least total time to build. Weka is an open-source software tool that consists of an accumulation of ma- chine learning algorithms for data mining undertakings. It contains apparatuses for information pre-processing, regression, visualization, classification, association rules and clustering. [8.] The best algorithm is j48 based on UCI data haste with the highest accuracy i.e. 56.76% and the total time to build the model is 0.04 seconds while LMT algorithm has the lowest accuracy i.e. 55.77% and the total time to build a model in 0.39seconds. There is an only marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.

3) Rajesh n, T manesha, Shaik hafeez, Hari krishna (2018) prediction of heart disease using machine learning algorithms. The Naive Bayes treats every variable as independent which helps it to predict even if variables don't have proper relation. We used decision trees and a combination of algorithms for the pre- diction of heart disease based on the above attributes. When the dataset is small Naive Bayes algorithm gives accurate results and when the dataset is large decision trees give the accurate results. Naïve Bayes will not give accurate results every time we need to consider the results of different algorithms and by all its results if a prediction is made it will be accurate.

4) V.v. Ramalingam*, Ayantan Dandapath, M Karthik raja(2018).They published a paper named; A survey on the use of machine learning techniques to forecast heart disease'. Algorithms and techniques used are -. Naive Bayes, support vector machine, random forest, ensemble model, decision tree, and k – nearest neighbour. Models based on Naïve Bayes classifier were computationally very fast and have also performed well. In the vast majority of cases, SVM performed admirably. Because they employ many algorithms to overcome the problem of overfitting, random forest and ensemble models have fared exceptionally well. A lot of research may be done on the best algorithm ensemble to employ for a specific sort of data.

5) Marjia sultana; afrin haider; Mohammad shorif uddin (22-24 Sept. 2016). They published a paper [6]. In this paper, two data sets (collected and UCI standard) are used separately for each data mining technique. This paper performed an experiment using diverse data mining techniques to find out a more accurate technique for heart disease prediction. [6.]Their findings show that for heart disease prediction the performances of Bayes net and SVM classifiers are the optimum among the investigated five classifiers: Bayes net, sma, kstar, mlp and j48. The prediction of heart disease requires a huge size of data that is too complex and massive to process and analyze by conventional techniques. Various experts employ a variety of data mining approaches to solve various problem.

V. METHODOLOGY IMPLEMENTATION

A. *Preprocessing*

We have collected data from various reliable sources from the internet. After analysing various factors, we have reached to a conclusion that 13 independent variables will determine 1 target variable. To do this we will have to split the target variable from the rest. If we can reach 96% accuracy at predicting whether or not a patient has heart disease during the proof of concept, we'll pursue this project.

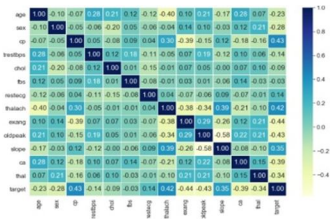


Fig 2-Correlation Matrix

VI. TRAINING AND TEST SPLIT

The train and split procedure is used to divide the data the dataset into two halves.

1) Train split

2) Test split

The model designed will first train on the train split where it tries to learn the patterns in the data. Then based on the patterns it has learnt it will tested on the test split. In this entire process choosing the test split size is also very important. A rule thumb is to use 80% of your data to train on and the other 20% to test on.

VII. MACHINE LEARNING MODELS

Machine learning models are majorly classified as supervised and unsupervised. If the model is supervised, it is divided into two categories: regression and classification. We will focus on the following machine learning models:

1) *Logistic Regression:* It is a basic classification algorithm which predicts the probability of a target variable.

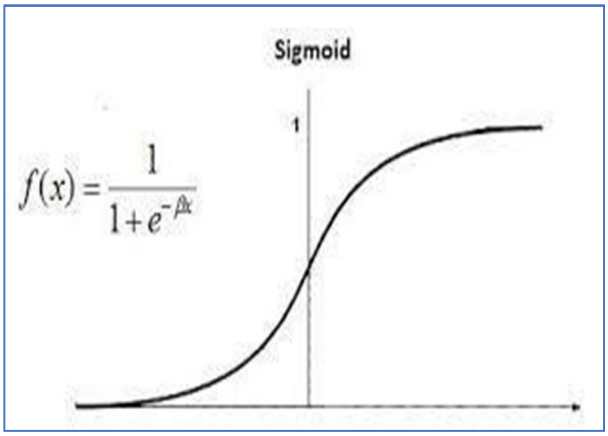


Fig 3-Logistic Regression

2) *K-nearest Neighbours:* It's a ma- chine learning algorithm that's supervised. The idea behind nearest neighbour methods is to find a predetermined number of training samples that are closest in distance to the new point and use them to predict the mark. It makes no assumptions about the data and is typically used for classification tasks where little to no prior knowledge of the data distribution is available. Finding the k closest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the identified data points to it is the aim of this algorithm.

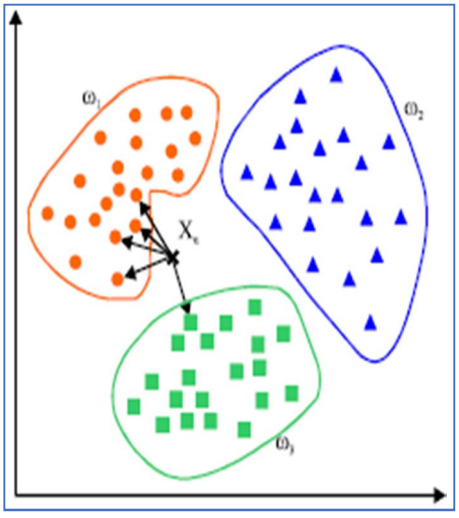


Fig 4-KNN

3) *Random Forest:* Random forest is a supervised machine learning algorithm that can be used to solve problems in both classification and regression. It builds decision trees out of data samples, then gets predictions from each of them before voting on the best solution.

$$RFfi_i = \frac{\sum_{j \in all\ trees} normfi_{ij}}{T}$$

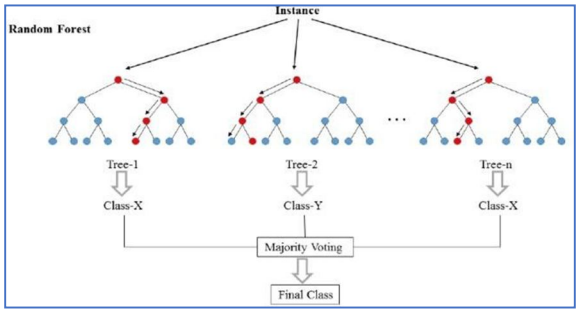


Fig 5-Descision Tree

VIII. RESULTS OBTAINED BY MACHINE LEARNING MODELS

1) 'Logistic regression': 0.8852459016393442,
2) 'knn': 0.6885245901639344,
3) 'random forest': 0.8360655737704918

IX. HYPER-PARAMETER TUNING AND CROSS VALIDATION

A hyperparameter is a parameter whose value is set before the model is allowed to train on the train split. Tuning the hyper parameters helps to increase the efficiency of a model. Not all the hyperparameters are to be considered any context. Choosing the right hyperparameters is also an im- portant task.

The best accuracy obtained for KNN is when the number of nearest neighbours is 11 with an accuracy score of 0.7540983606557377

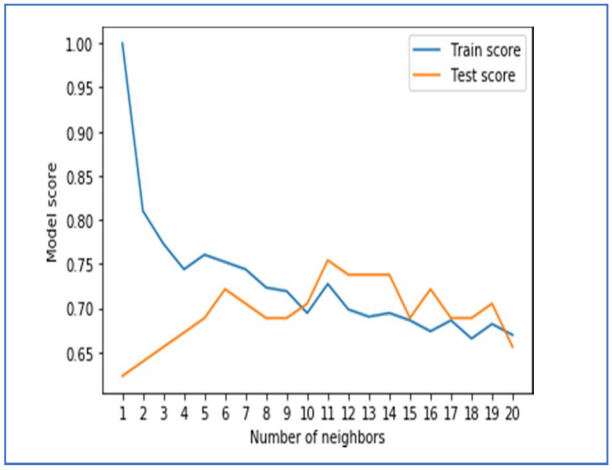


Fig 6-Tuning of Hyperparameter for KNN

The best parameter found for logistic regression is {'solver': 'liblinear', 'c': 0.23357214690901212} with a accuracy score of 0.8852459016393442

The best parameter found for random forest is {'n_estimators': 210, 'min_samples_split': 4, 'min_samples_leaf': 19, 'max_depth': 3} with a accuracy score of 0.8688524590163934

X. COMPARE WITH YOUR EXISTING MODEL

Sno.	Algorithm	Accuracy Found By Us	Accuracy Of Base Research Paper
1.	Logistic Regression	89%	--
2.	Random Forest	87%	--
3.	Decision Tree	--	79%
4.	KNN	75%	74%
5.	SVM	--	87%

In our base research (Paper 1) we found that the machine learning algorithms used were KNN, SVM, Decision Tree and the highest accuracy achieved was 87%. Also there was a lack of tuning of hyperparameters. In our re- search paper we worked on ensemble learning algorithms like Random Forest , Logestic Regression, KNN. And after tuning the hyperparameters we found that the highest accuracy is achieved through Logistic Regression with a accuracy rate of 89%

XI. RESULTS

After tuning the hyper parameters for KNN, Logistic Regression, Random forest and selecting the best ones we found the following results for accuracy:
KNN: 0.6885245901639344
Logistic Regression: 0.8852459016393442
Random Forest: 0.8360655737704918

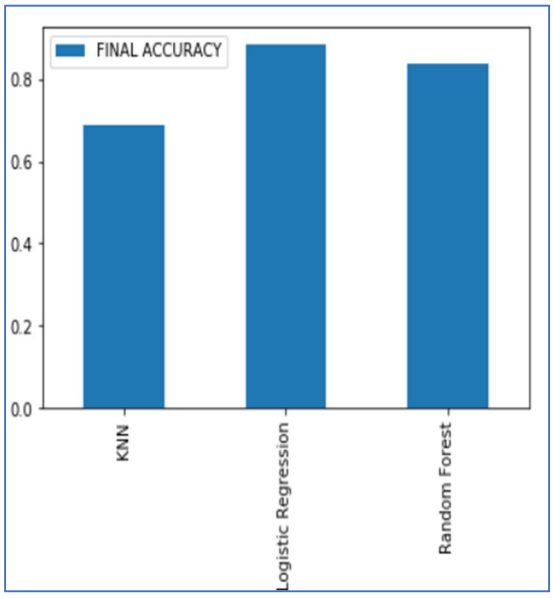


Fig 7-Accuracy Comparison

Among these we can see that random forest with a certain set of hyperparameters Logistic Regression performs the best.
Now we will find the other metrics for the logistic regression model.

A. *ROC Curve*

The metric compares the true positive rate with the false positive rate.
The True Positive Rate (TPR) is defined as follows:

$$TPR = \frac{TP}{TP+FN}$$

The False Positive Rate (FPR) is defined :

$$FPR = \frac{FP}{FP+TN}$$

It also provides us with AUC scores which denotes the area underneath the ROC curve

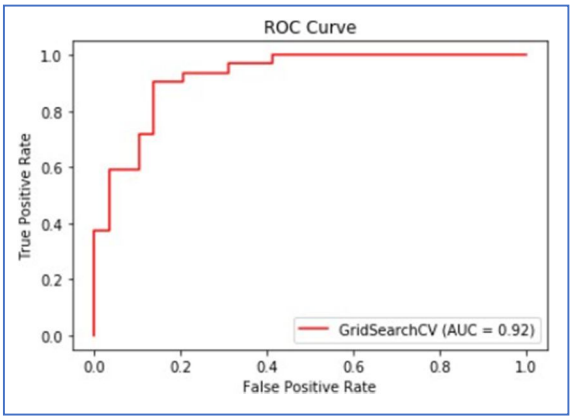


Fig 8-ROC Curve

B. *Confusion Matrix*

A confusion matrix is a table that is used to describe the output of a classification model/classifier by comparing the true values of the training and test datasets. It is divided into four parts, each of which is defined as follows:

1) True positives (TP): These are cases in which we expected yes (they have the disease) and they do.
2) Real negatives (TN): We predicted they wouldn't have the disorder, and they don't.
3) False positives (FP): We expected that they will have the disease, but they don't. (This is often referred to as a "Type I error.")
4) False negatives (FN): We expected that they will not have the disorder, but they do. (This is often referred to as a "Type II error.")

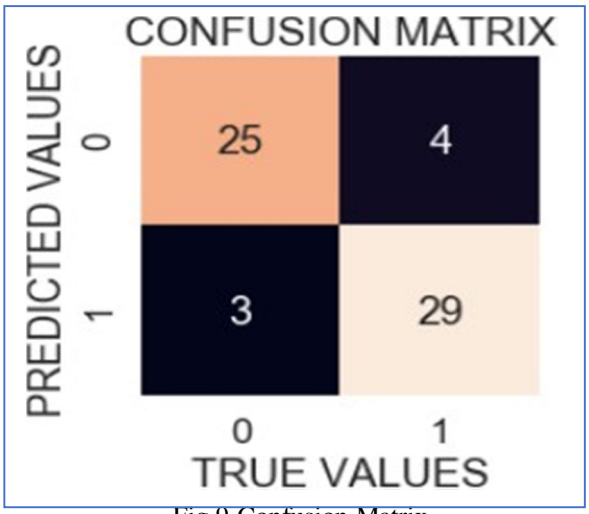


Fig 9-Confusion Matrix

C. Classification Report

The Classification report is used to find the quality of predictions from a classification algorithm. It helps us to find how many predictions are correct and how many are wrong. More specifically, it gives us an understanding of True negatives and False Negatives, True Positives and False Positives, and uses them to predict the metrics of a classification

The main metrics found by the Classification report are accuracy, precision, recall, and f1- score.

The model's accuracy is expressed in decimal form. Precision refers to a classifier's ability to avoid labelling a negative occurrence as positive. Recall - This metric indicates the per- centage of true positives that were successfully classified. The F1 score is a weighted harmonic mean of precision and recalls, with 1.0 being the highest and 0.0 being the poorest. F1 Score = 2*(Recall * Precision) / (Recall + Precision) Support - The number of samples used to calculate each metric. Support - The number of samples used to calculate each metric.

	precision	recall	f1-score	support
0	0.89	0.86	0.88	29
1	0.88	0.91	0.89	32
accuracy			0.89	61
macro avg	0.89	0.88	0.88	61
weighted avg	0.89	0.89	0.89	61

Fig 10-Classification Report

D. Cross Validation Score

The statistical method of cross-validation is majorly used for measuring the skill of machine learning models. The k-fold cross-validation is used to test how a machine learning model per- forms with different sets of data.

As our data set consists of 303 entries using 5-folds of cross-validation along with the Logistic Regression model and with the best hyperparameters yielded the following results:

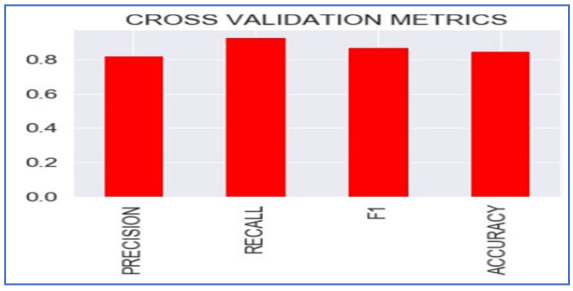


Fig 11-Cross Validation Metrics

E. Feature Importance

It refers the techniques that assign a score to the input attributes/features with respect to the fact that which feature has the highest contribution in predicting the results for a given ma- chine learning model. For finding it we will use the coef_ attribute .The coef_ attribute is the coefficient of the features in the decision function. We can note that negative coef_ attribute denotes the presence of negative correlation.

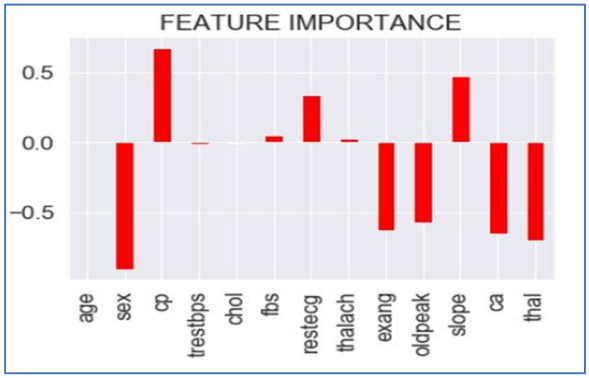


Fig 12-Feature Importance

XII. CONCLUSION

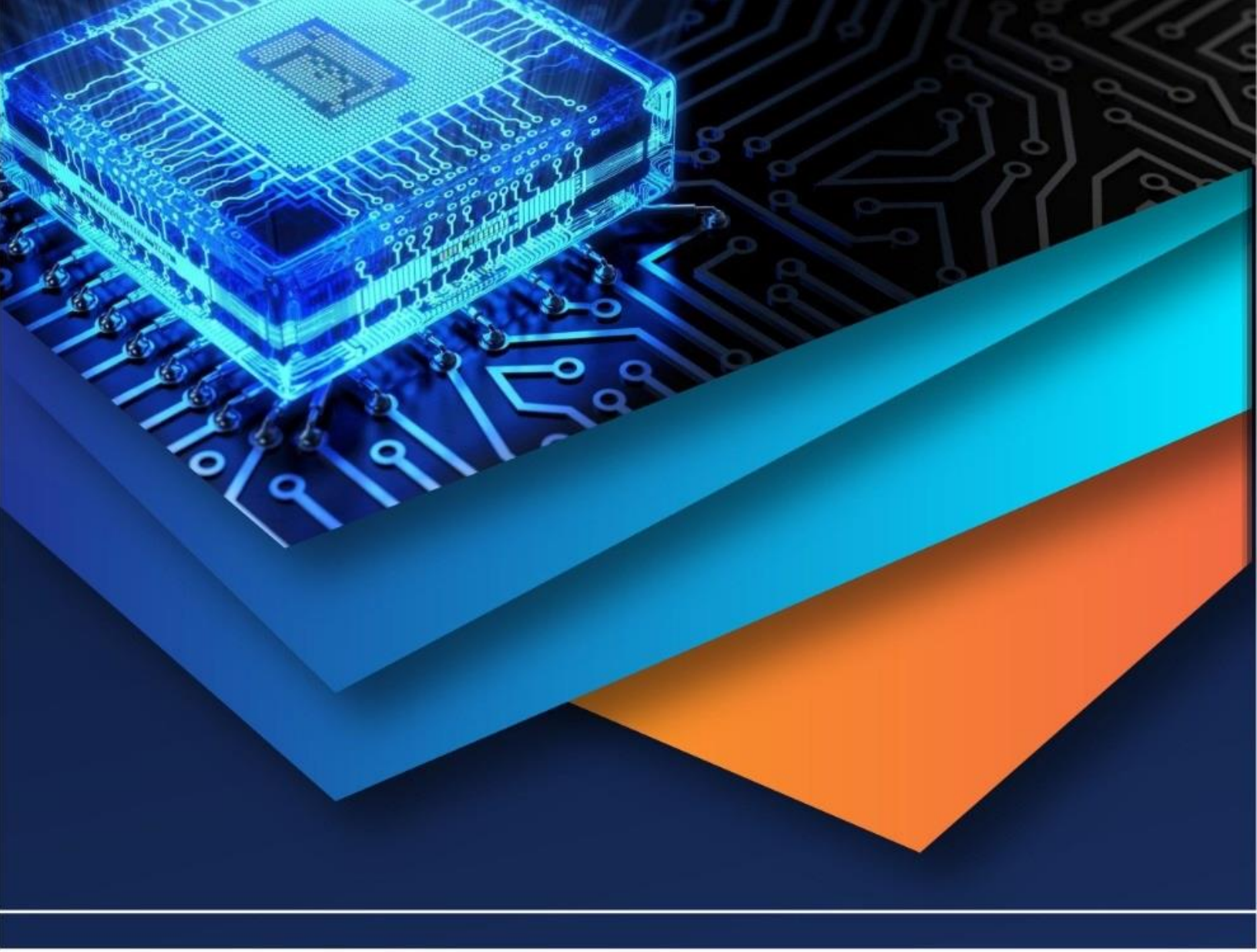
With the rising number of deaths due to heart disease, it is becoming increasingly important to build a system that can effectively and accurately forecast heart disease. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of KNN, Logistic Regression and Random Forest for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Logistic regression algorithm is the most efficient algorithm with accuracy score of 89% for prediction of heart disease. Accuracy of the algorithms in ma- chine learning depends upon the dataset that used for training and testing purpose.

XIII. FUTURE SCOPE

In the future, the work could be improved by creating a web application premised on the logistic regression algorithm and by using a larger dataset than the one used in this study, which would help to provide better outcomes and aid health professionals in predicting heart disease efficiently and effectively.

REFERENCES

[1] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.
[2] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.
[3] Rajesh, N., T, M., Hafeez, S., & Krishna, H. (2018). Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering & Technology, 7(2.32), 363-366. doi:http://dx.doi.org/10.14419/ijet.v7i2.32.15714
[4] 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
[5] Kaur, A., & Arora, J. (2018). HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY. International Journal of Advanced Research in Computer Science, 9(2).
[6] "Sultana, M., Haider, A., & Uddin, M. (2016). Analysis of data mining techniques for heart disease prediction. 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 1-5.
[7] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia technology, 10, 85-94.
[8] Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. Advances in Computational Sciences and Technology, 10(7), 2137-2159.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY