iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

HEART DISEASE PREDICTION USING ENSEMBLE APPROACH

Mrs. Rekha. P
Assistant Professor, Sri Muthukumaran Institution of Technology, Mangadu, Chennai
R. Purushothamon. G
Student, PG Student, II MCA, Sri Muthukumaran Institution of Technology, Mangadu, Chennai

---

ABSTRACT:
A heart attack, also known as a myocardial infarction, occurs when the flow of blood to a part of the heart muscle is suddenly blocked. This blockage prevents the heart from receiving enough oxygen. If blood flow is not promptly restored, the heart muscle begins to die. Machine learning algorithms are a powerful tool for predicting heart disease, allowing for more accurate results than traditional methods. Different models are compared and evaluated to identify the most accurate algorithm. Early detection is key to managing heart disease, and machine learning can help to identify at-risk individuals. Prevention of heart disease is invaluable for saving lives and reducing medical costs. Machine learning algorithms can provide the most accurate predictions possible, offering the best chance of preventing serious health issues.

Keywords:
Heart, Muscle, Machine learning, blockage, predictions, accuracy.

INTRODUCTION:

Machine learning algorithms are a powerful tool for predicting heart disease, allowing for more accurate results than traditional methods. Different models are compared and evaluated to identify the most accurate algorithm. Early detection is key to managing heart disease, and machine learning can help to identify at-risk individuals. Prevention of heart disease is invaluable for saving lives and reducing medical costs. Machine learning algorithms can provide the most accurate predictions possible, offering the best chance of preventing serious health issues.

OBJECTIVES:

The primary objective of the "Heart Disease Prediction Using Ensemble Approach" is to develop a highly accurate, robust, and generalizable predictive model that can effectively identify individuals at risk of heart disease. By leveraging the strengths of multiple machine learning models through ensemble techniques, the goal is to create a system that improves upon the limitations of individual models, offering more reliable predictions that can be utilized in clinical settings.

EXISTING SYSTEM:

Normal traditional methods depending on doctor’s analysis and their experience. The second one comes less accurate primitive prediction model like logistic regression, decision tree classifier.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

LITERATURE SURVEY:

s.no	Year	Title	Author	Algorithm	Disadvantages
1	2017	An Ensemble Random Forest Algorithm for Insurance Big Data Analysis	WEIWEI LIN1,2, ZIMING WU1 , LONGXIN LIN3 , ANGZHAN WEN1 , AND JIN LI	Random Forest	**A forest is less interpretable than a single decision tree**
2	2019	An Adaptive Estimation of Distribution Algorithm for Multipolicy Insurance Investment Planning	Wen Shi, Student Member, IEEE, Wei-Neng Chen	EDA	Anticipating the Unforeseen
3	2020	Research on the UBI Car Insurance Rate Determination Model Based on the CNN-HVSVM Algorithm	CHUN YAN1 , XINDONG WANG 1 , XINHONG LIU2 , WEI LIU 3 , (Member, IEEE), AND JIAHUI LIU1	CNN	**A small change in the dataset can make the tree structure unstable which can cause variance**
4	2022	Blockchain and AI-Empowered Healthcare Insurance Fraud Detection: An Analysis, Architecture, and Future Prospects	KHYATI KAPADIYA1 , USHA PATEL , RAJESH GUPTA , MOHAMMAD DAHMAN ALSHEHRI	SVM	Choosing a “good” kernel function is not easy
5	2021	Clustering Application for Data-Driven Prediction of Health Insurance Premiums for People of Different Ages	Tallal Omar, Mohamed Zohdy	K-means	Require high memory – need to store all of the training data

PROPOSED SYSTEM & ADVANTAGES:

Does not need any human analysis. It should give more accurate prediction. Different machine learning models should be used. Random forest classifier show the highest accuracy. Ensemble learning method using multiple decision trees. Trees generated by randomly selecting features and samples. Final prediction made by averaging or majority voting of trees. Uses bagging and bootstrapping to reduce variance. Provides feature importance measurement.

METHODOLOGY: -

Suppose you want to predict the gender of a commercial customer. Collect data about height, weight, occupation, salary, shopping cart, etc. from your customer database. You know the gender of each customer, but only male or female. The purpose of the classifier is to assign probabilities of whether you are male or female (i.e. a label) based on information (i.e. features collected from you). Once the model learns to recognize males or females, it can use new

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

data to make predictions. For example, suppose you just received new information from an unknown customer and want to know if the customer is male or female. If the classifier predicts Male = 70%, it means that the algorithm has 70% confidence that this customer is male and she is 30% female. A label can consist of two or more classes. The example above has only two classes, but there are dozens of classes (glass, table, shoes, etc.) if the classifier needs to predict an object. Each object represents a class).

SYSTEM ARCHITECTURE:

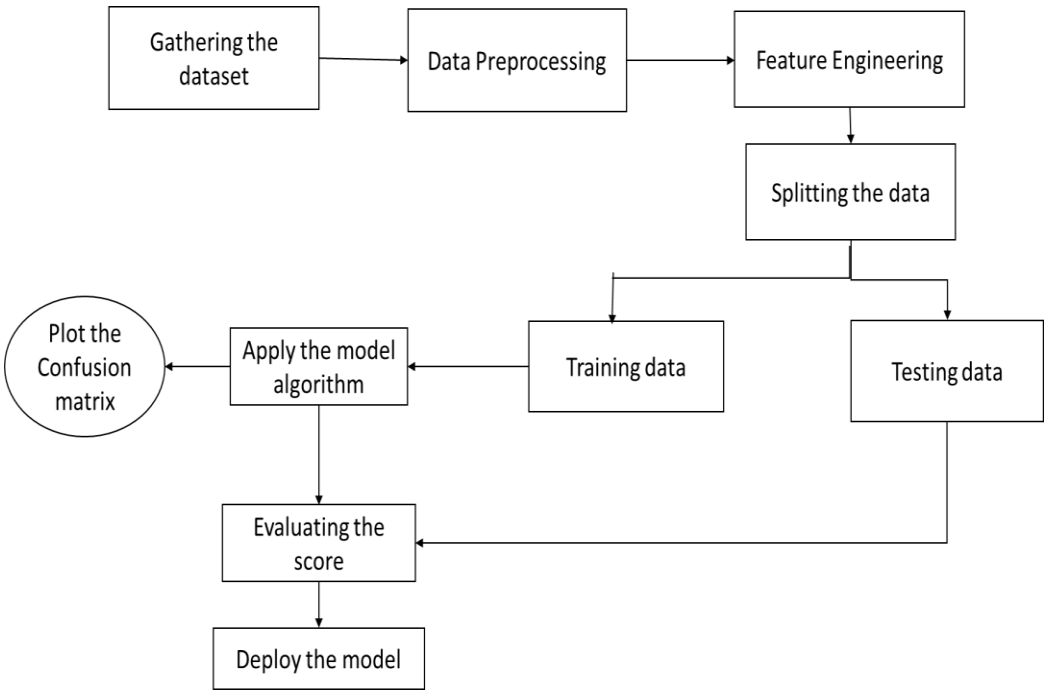

Figure 1. System Architecture

ALGORITHMS USED: -
LINEAR REGRESSION:
A machine learning algorithm built on supervised learning is linear regression. Activate a regression task. Run a regression task. Regression models target predictors based on independent variables. Different regression models differ based on the type of relationship between dependent and independent variables considered and the number of independent variables used. The dependent variable in regression has many names. This is sometimes called the outcome variable, criterion variable, endogenous variable, or regression sand. Independent variables are sometimes called exogenous variables, predictor variables, or regressors.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

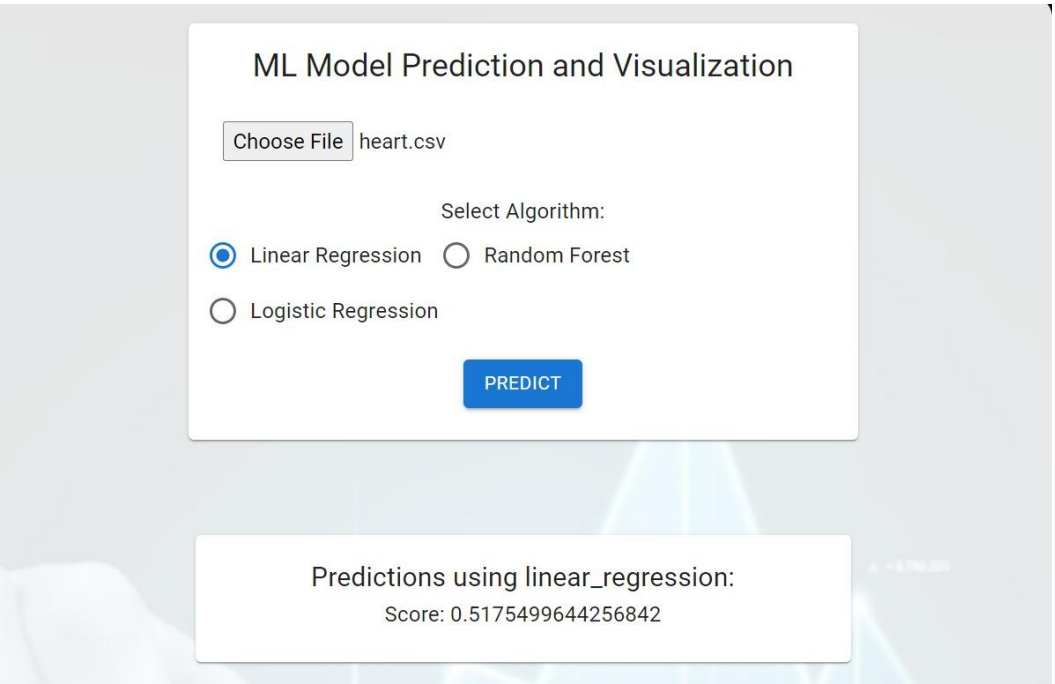

Figure 2. Linear Regression

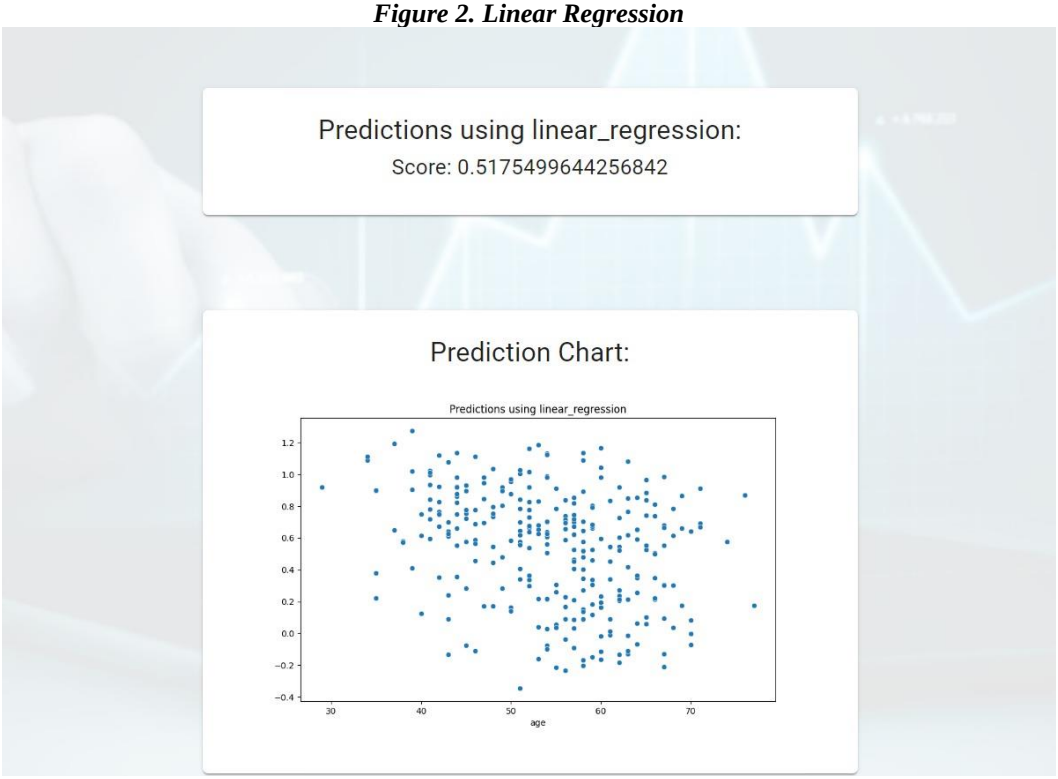

Figure 3. Linear Regression

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

LOGISTIC REGRESSION:

Logistic regression is one of the most popular machine learning algorithms that falls under supervised learning techniques. Logistic regression predicts the output of a categorical dependent variable. Using a specific collection of independent factors, it is used to predict a categorical dependent variable. As a result, the outcomes must be discrete or categorical. Can be true or false, 0 or 1, yes or no, and so forth. But instead of giving exact values as 0 and 1, it gives probability values between 0 and 1. Logistic regression is very similar to linear regression except for how it is used. Linear regression is used to solve regression problems and logistic regression is used to solve classification problems.

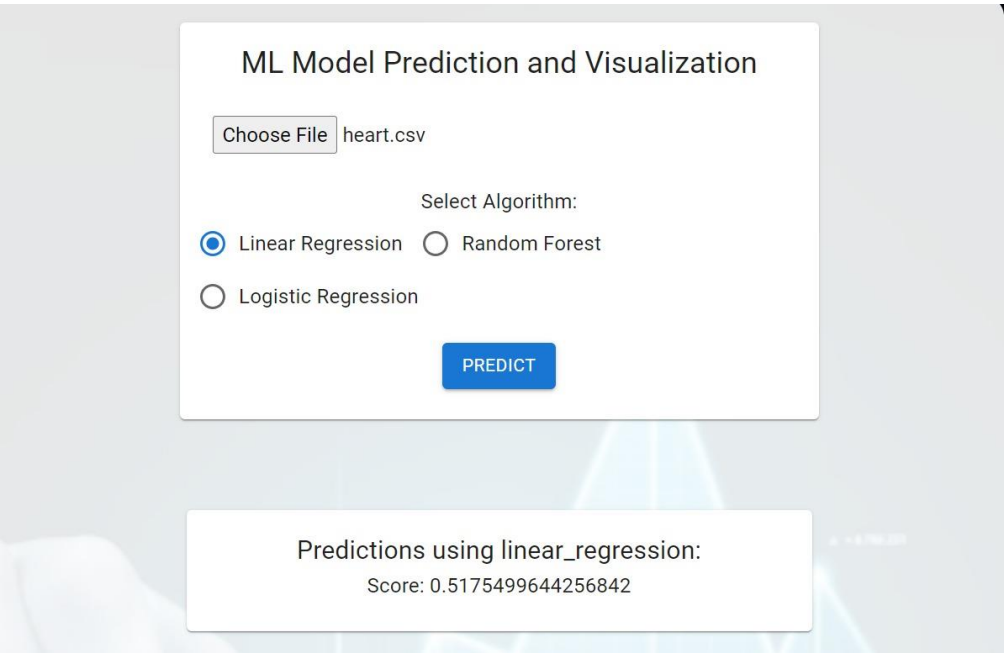

Figure 4. Logistic Regression

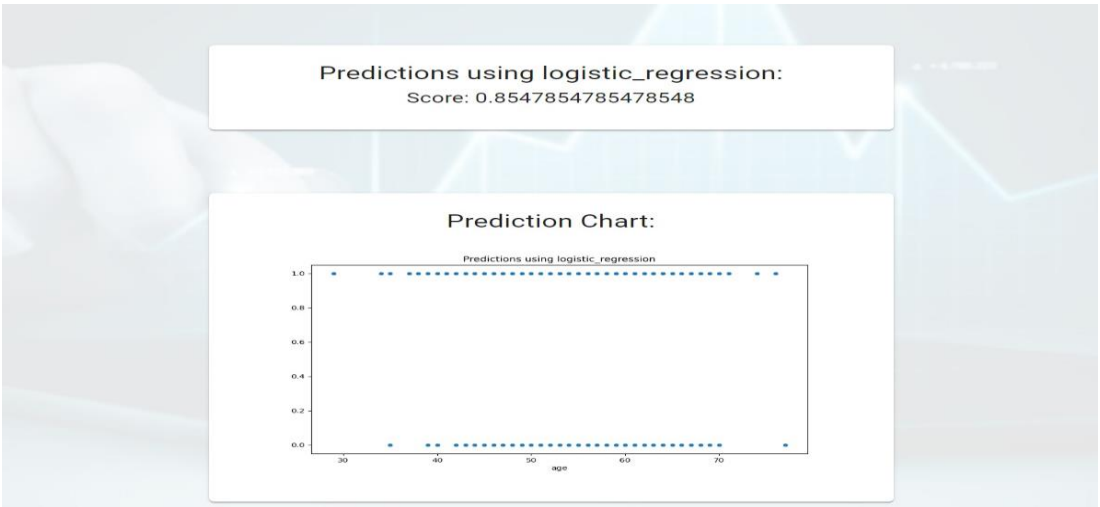

Figure 5. Logistic Regression

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

RANDOM FOREST:

Random Forest is a machine learning algorithm that can be used for classification and regression tasks. It belongs to the ensemble learning family of algorithms, which means that it combines multiple individual models to improve the accuracy and robustness of the predictions.

In Random Forest, a large number of decision trees are trained on different random subsets of the data. Each tree is grown independently and can split the data based on a random subset of features. The final prediction is then made by aggregating the predictions of all individual trees, typically by taking the majority vote in classification tasks or the average in regression tasks.

Random Forest is a popular algorithm in machine learning due to its good performance, scalability, and interpretability. It can handle large datasets with high dimensionality, handle missing values and outliers, and provide insights into the importance of different features in the prediction.

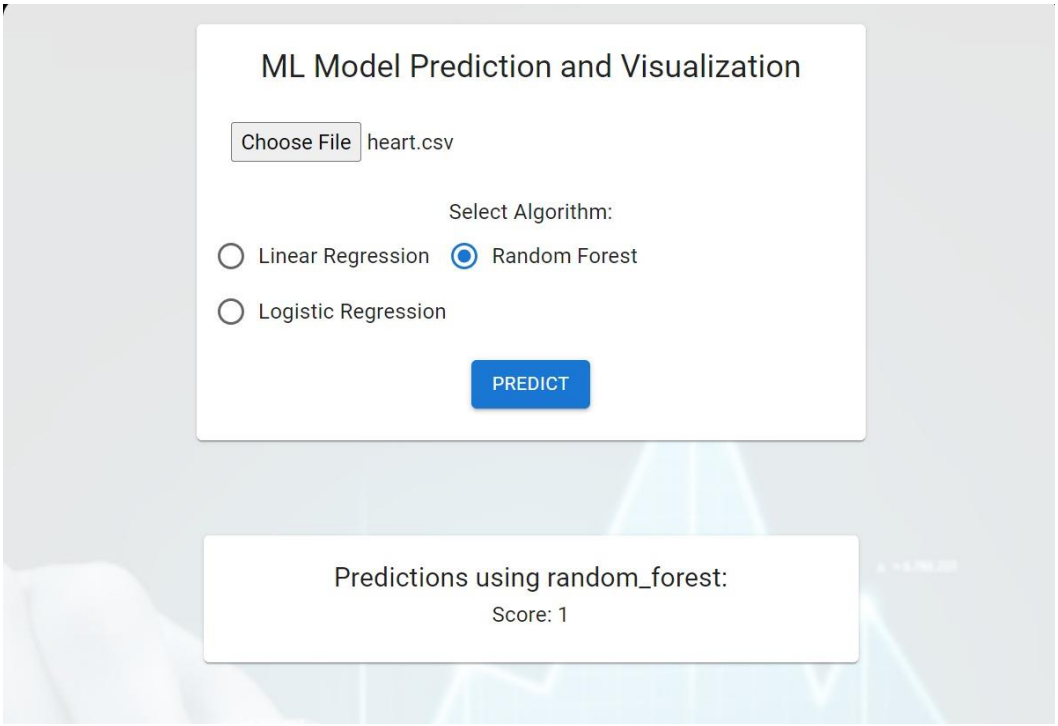

Figure 6. **Random Forest**

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

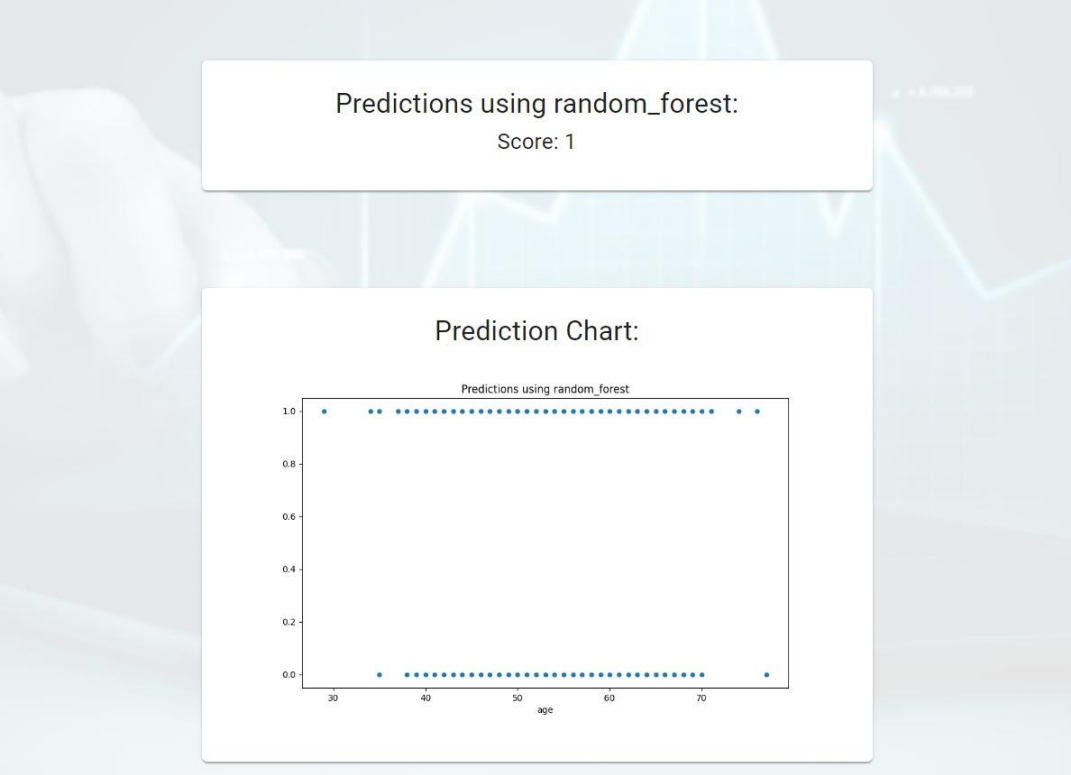

Figure 7. **Random Forest**

•**Decision Trees**: Random Forest algorithm is based on Decision Trees. Decision Trees are a type of supervised learning algorithm that can be used for both classification and regression tasks. The basic idea behind decision trees is to split the data into smaller subsets based on different criteria, such as the value of a certain feature, until the subsets become pure or homogeneous. Each split creates a new node in the tree, and the final prediction is made by traversing the tree from the root to a leaf node.

•**Ensemble learning**: Random Forest algorithm belongs to the family of ensemble learning algorithms. Ensemble learning combines multiple individual models to improve the accuracy and robustness of the predictions. In Random Forest, multiple decision trees are trained on different random subsets of the data, and their predictions are aggregated to make the final prediction.

•**Feature Selection**: In Random Forest, each decision tree is grown on a random subset of features. This process is called feature selection or feature subsampling. The purpose of feature selection is to reduce the correlation between the individual trees and increase the diversity of the ensemble. In addition, feature selection can also help to reduce overfitting and improve the generalization performance of the model.

•**Bootstrap Aggregating**: The training data for each decision tree is obtained through a process called bootstrap aggregating or bagging. Bagging involves sampling the data with replacement to create multiple training sets that are used to train the individual trees. The purpose of bagging is to reduce the variance of the individual trees and improve their stability.

•**Out-of-Bag Error**: In Random Forest, some samples are not used in the training of certain trees due to the random selection of the data subsets. These samples are called out-of-bag (OOB) samples. OOB samples can be used to estimate the generalization error of the model without the need for cross-validation or a separate test set.

•**Importance of Features**: Random Forest can also provide insights into the importance of different features in the prediction. The importance of each feature is estimated based on the decrease in impurity or information gain caused by its use in the splits of the decision trees.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

Random Forest is a powerful algorithm that can be used for a wide range of applications, such as classification, regression, feature selection, and outlier detection. Its performance is often comparable or superior to other popular algorithms such as support vector machines (SVMs) and neural networks. However, Random Forest can be computationally expensive and may not be suitable for very large datasets with many features.

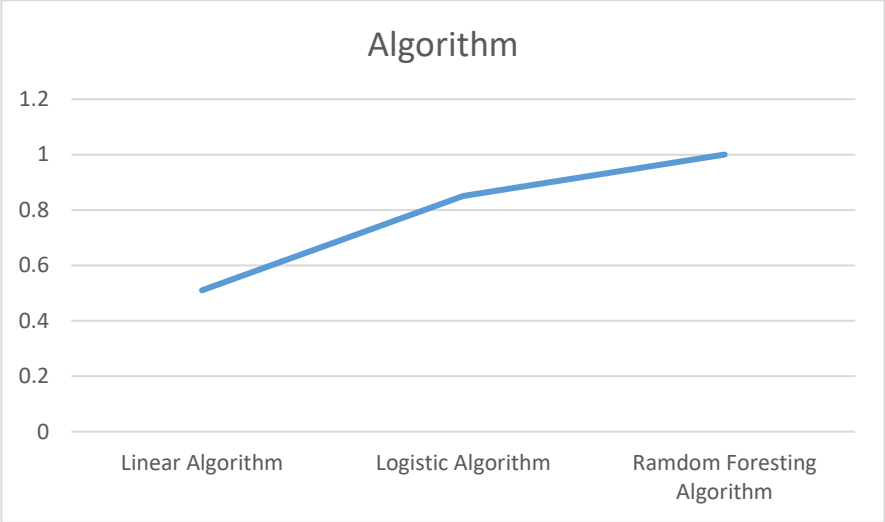

Figure 8. Algorithm

CONCLUSION: -

Various techniques were adopted to preprocess the data to suite the requirement of analysis. Feature selections were made to optimize the performance of machine learning algorithms. Ensemble prediction gave better accuracy when combined using Random forest algorithm as combiner. Better feature selection techniques can be applied to further improve the accuracy.

FUTURE ENHANCEMENT:

Future enhancements for a heart disease prediction system using an ensemble approach could include advanced ensemble techniques like stacking, integration with deep learning models, and real-time data processing. Improved feature engineering, data augmentation, and personalized predictions could enhance accuracy. Explainability tools like SHAP or LIME would increase transparency, while integration with Electronic Health Records (EHRs) would streamline clinical use. Scalability through distributed computing and rigorous clinical validation, alongside a user-friendly interface, would ensure broader adoption and reliability, ultimately leading to better patient outcomes and more informed healthcare decisions.

REFERENCE:

[1] Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01
[2]]Prerana T H M, Shivaprakash N C et al “Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS”,Vol 3, PP: 90-99 ©IJSE, 2015
[3] Salam Ismaeel, Ali Miri et al “Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis”,IEEE Canada International Humanitarian TechnologyConference,DOI:10.1109/IHTC.2015.7238043, 03 September 2015
[4]F BrainBoudi,’Risk Factors for Coronary Artery Disease’ ,2016. [Online]Available: https://emedicine.medscape.com/article/164163 -overview.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

[5]National Health Council,'Heart Health Screenings',2017.[Online]Available:http://www.heart.org/HEARTORG/Conditions/Heart-HealthScreenings_UCM_428687_Article.jsp#.WnsOAeeYPIV
[6]ScikitLearn,'MLPCLassifier',Available:http://scikitlearn.org/stable/modules/generated/sklearn.neural_network.MLPClassifi er.html
[7] Prediction System for heart disease using Naïve Bayes *Shadab Adam Pattekari and Asma Parveen Department of Computer Science and Engineering Khaja Banda Nawaz College of Engineering
[8] Comak E, Arslan A (2012) A biomedical decision support system using LS-SVM classifier with an efficient and new parameter regularization procedure for diagnosis of heart valve diseases. J Med Syst 36:549–556
[9] Ahmed Fawzi Otoom , Emad E. Abdallah, Yousef Kilani ,Ahmed Kefaye and Mohammad Ashour (2015) Effective Diagnosis and Monitoring of Heart Disease ISSN: 1738-9984 IJSEIA