SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY

A MINI-PROJECT REPORT

ON

USING RANDOM FOREST ENSEMBLE TECHNIQUE FOR REGRESSION AND CLASSIFICATION USING MULTIPLE MACHINE LEARNING MODELS

Submitted in partial fulfilment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

IN

Information Science and Engineering

Submitted by

Gokul Krishna	R23EQ068
Piyush	R23EQ071
Priyadip Sinha	R23EQ079
Priyanshu	R23EQ081

Under the guidance of

Prof. Suvarna Hugar
School of C&IT

REVA University

Rukmini Knowledge Park, Kattigenahalli, Yelahanka,Bengaluru-560064
www.reva.edu.in

May 2025

REVA UNIVERSITY
Bengaluru, India

DECLARATION

We, Gokul Krishna **(R23EQ068)** Piyush **(R23EQ071)** Priyadip Sinha **(R23EQ079)** Priyanshu **(R23EQ081)** **Team members and** students of B.Tech in "Information Science and Engineering/ Computer Science and Engineering (Artificial Intelligence and Machine Learning)/ Computer Science and Systems Engineering/ Computer Science and Information Technology", VI Semester, School of Computing and Information Technology, REVA University declare that the Mini-Project Report entitled **"Heart Disease Prediction"** done by us under the guidance of **prof.Suvarna Hugar**, **Designation**, School of Computing and Information Technology, REVA University.

We are submitting the Mini-Project Report in partial fulfilment of the requirements for the award of the degree of Bachelor of Engineering in Computing and Information Technology by the REVA University, Bangalore during the academic year 2024-25

We further declare that the Mini-Project or any part of it has not been submitted for award of any other Degree of REVA University or any other University / Institution.

1. Gokul Krishna (R23EQ068) – Signature…….. Date: ……
2. Piyush (R23EQ071) – Signature…….. Date:……
3. Priyadip Sinha (R23EQ079) – Signature…….. Date: ……...
4. Priyanshu (R23EQ081) – Signature…….. Date: ……

REVA UNIVERSITY
Bengaluru, India

SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY

CERTIFICATE

This is to certified that the Mini-Project entitled **"USING RANDOM FOREST ENSEMBLE TECHNIQUE FOR REGRESSION AND CLASSIFICATION USING MULTIPLE MACHINE LEARNING MODELS"** carried out by <**Gokul Krishna, Piyush, Priyadip Sinha, Priyanshu and R23EQ068 ,R23EQ071 ,R23EQ079 ,R23EQ081**> are bonafide students of REVA University during the academic year 2024-25. The above-mentioned students are submitting the Mini-Project report in partial fulfilment for the award of **Bachelor of Technology** in <**Information Science and Engineering**> during the academic year 2024-25**.** The Mini-Project report has been approved as it satisfies the academic requirements in respect of Mini-Project work prescribed for the said degree.

Signature of Guide	**Signature of HOD**	**Signature of Director of School**
Date: …….	**Date: …….**	**Date: …….**

Official Seal of the School

REVA UNIVERSITY
Bengaluru, India

CONTENT TABLE

ABSTRACT

In recent years, machine learning has emerged as a powerful approach for solving real-world problems involving prediction and decision-making. Among various machine learning techniques, ensemble learning methods have gained significant attention due to their ability to improve accuracy, robustness, and generalization performance. This project focuses on the implementation and analysis of the **Random Forest ensemble technique** for both **regression and classification problems** using multiple machine learning models.

Random Forest combines the predictions of multiple decision trees to produce a more accurate and stable result compared to individual models. In this project, multiple supervised learning models such as **Linear Regression, Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest** are implemented and compared. The performance of these models is evaluated using appropriate metrics such as accuracy, mean squared error, and R-squared score.

The experimental results demonstrate that the Random Forest ensemble technique consistently outperforms individual machine learning models in terms of prediction accuracy and error reduction. This project highlights the effectiveness of ensemble learning in handling complex datasets and provides insights into model selection for regression and classification tasks.

INTRODUCTION

Machine Learning (ML) is a subfield of artificial intelligence that enables systems to learn from data and make predictions or decisions without being explicitly programmed. ML techniques are widely used in domains such as healthcare, finance, education, and engineering for tasks like classification, regression, and pattern recognition.

Supervised learning is one of the most commonly used ML approaches, where models are trained using labeled datasets. Classification problems deal with predicting categorical outputs, whereas regression problems focus on predicting continuous values. Traditional machine learning models often suffer from limitations such as overfitting, high variance, or low accuracy when applied individually.

To overcome these limitations, **ensemble learning techniques** such as Random Forest are used. Random Forest constructs multiple decision trees during training and combines their outputs using voting (classification) or averaging (regression). This approach improves prediction accuracy, reduces overfitting, and enhances model stability.

This project aims to implement and compare multiple machine learning models for regression and classification, with a special focus on the Random Forest ensemble technique.

PROBLEM STATEMENT:

Single machine learning models often fail to provide optimal performance when dealing with complex and high-dimensional datasets. These models may suffer from overfitting, underfitting, or sensitivity to noise. There is a need for a robust and reliable approach that can improve prediction accuracy for both regression and classification problems.

The problem addressed in this project is to analyze whether ensemble learning techniques, specifically Random Forest, can outperform traditional machine learning models when applied to regression and classification tasks using the same dataset.

OBJECTIVES:

1. To study supervised machine learning techniques for regression and classification.
2. To implement multiple ML models such as Linear Regression, Logistic Regression, KNN, Decision Tree, and Random Forest.
3. To analyze the performance of individual models and compare them with Random Forest.
4. To evaluate models using appropriate performance metrics.
5. To demonstrate the effectiveness of ensemble learning techniques.

.

SCOPE AND LIMITATIONS:

Scope

1. The project demonstrates the use of ensemble learning for improved prediction accuracy.
2. It provides a comparative analysis of regression and classification models.
3. The approach can be extended to real-world datasets in various domains.

Limitations

1. Model performance depends on dataset quality and size.
2. Random Forest requires more computational resources compared to simpler models.
3. Hyperparameter tuning is limited to basic configurations.

LITERATURE SURVEY

Machine learning techniques have been widely explored for solving regression and classification problems due to their ability to learn patterns from large datasets. Traditional machine learning models such as Linear Regression, Logistic Regression, Decision Trees, and K-Nearest Neighbors have been commonly used in early research; however, these models often suffer from limitations such as overfitting, sensitivity to noise, and poor generalization on complex datasets. To overcome these issues, ensemble learning techniques were introduced, which combine multiple base learners to improve predictive performance.

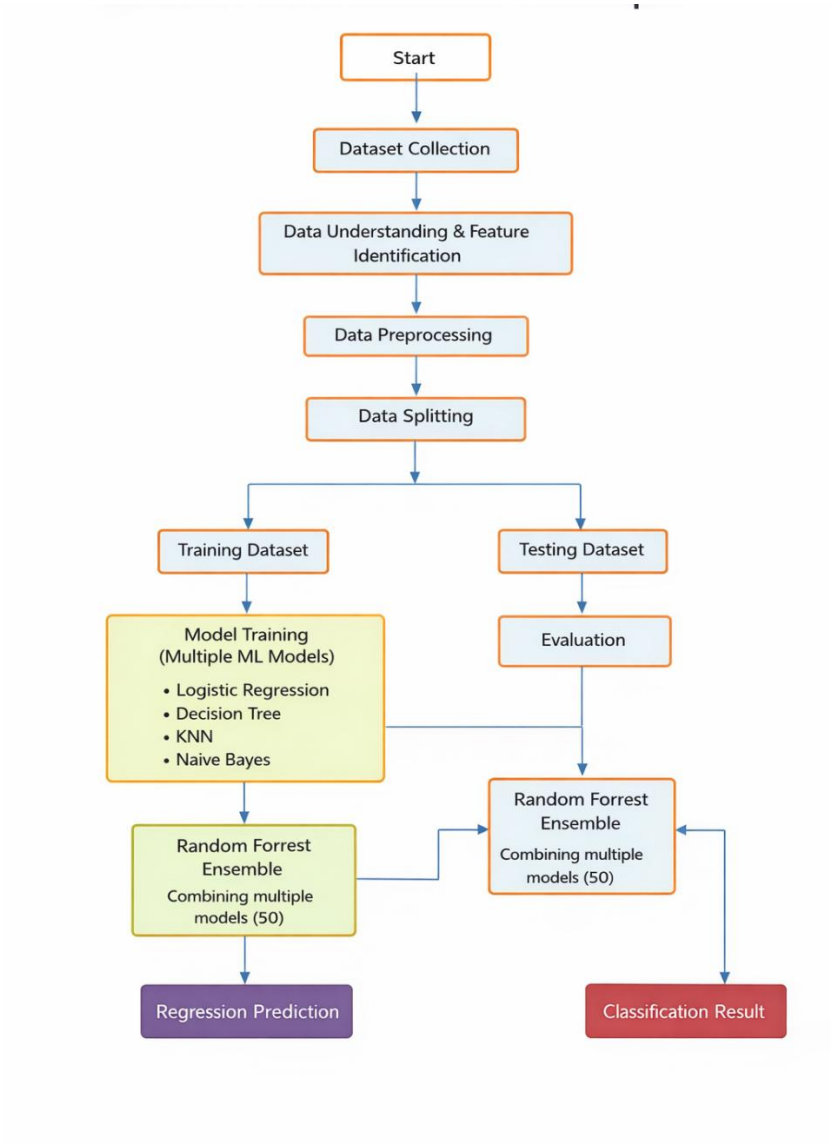
One of the most influential ensemble techniques is the Random Forest algorithm, introduced by Leo Breiman, which is based on the principles of bootstrap aggregation and random feature selection. Several studies have demonstrated that Random Forest significantly improves prediction accuracy by reducing variance and preventing overfitting when compared to single decision tree models. Researchers have shown that Random Forest performs effectively for both classification and regression tasks due to its ability to model non-linear relationships and handle high-dimensional data efficiently.

Comparative studies indicate that Random Forest outperforms traditional classifiers such as Logistic Regression and KNN, especially in datasets with complex feature interactions. In regression tasks, Random Forest has been found to produce lower prediction errors compared to linear models, as it does not rely on strict assumptions about data distribution. Additionally, Random Forest provides internal validation through out-of-bag error estimation and feature importance analysis, which enhances model reliability and interpretability.

METHODOLOGY

SYSTEM ARCHITECTURE

The system architecture gives an overview of the working of the system. The working of this system is shown below:



DATASET DETAILS

Dataset Name: Housing Data (Boston Housing Dataset)
Source: Kaggle

Dataset Link:
https://www.kaggle.com/datasets/altavish/boston-housing-dataset

ATTRIBUTES OF THE DATASET

The dataset contains **14 attributes**, out of which **13 are input features** and **1 is the target variable**.

1. **CRIM** – Per capita crime rate by town
2. **ZN** – Proportion of residential land zoned for lots over 25,000 sq.ft
3. **INDUS** – Proportion of non-retail business acres per town
4. **CHAS** – Charles River dummy variable (1 if tract bounds river, else 0)
5. **NOX** – Nitric oxide concentration
6. **RM** – Average number of rooms per dwelling
7. **AGE** – Proportion of owner-occupied units built prior to 1940
8. **DIS** – Weighted distances to employment centers
9. **RAD** – Index of accessibility to radial highways
10. **TAX** – Full-value property tax rate per $10,000
11. **PTRATIO** – Pupil–teacher ratio by town
12. **B** – Proportion of Black population by town
13. **LSTAT** – Percentage of lower status population
14. **MEDV** – Median value of owner-occupied homes (**Target Variable**)

MACHINE LEARNING

In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.

SUPERVISED MACHINE LEARNING

As its name suggests, Supervised machine learning is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based

on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the

output using the test dataset. The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).

Categories of Supervised Machine Learning:

Supervised machine learning can be classified into two types of problems, which are given below:

a) Classification

b) Regression

a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc. The classification algorithms predict the categories present in the dataset.

b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc. Some popular Regression algorithms are given below:

• Simple Linear Regression Algorithm

• Multivariate Regression Algorithm

• Decision Tree Algorithm

• Lasso Regression

UNSUPERVISED MACHINE LEARNING

Unsupervised learning is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabelled dataset, and the machine predicts the output without any supervision. In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision. The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

SUPERVISED ALGORITHMS

1.Decision Tree

Decision Tree is a widely used Machine Learning algorithm that comes under the Supervised Learning category and can be applied to both **classification and regression** problems. It operates by breaking down a complex decision-making process into a series of simpler decisions, represented in the form of a tree structure. The tree consists of a root node, internal decision nodes, branches, and leaf nodes. Each internal node represents a test or condition on an input feature, each branch represents the outcome of that condition, and each leaf node represents the final predicted class or continuous value.

The working of a Decision Tree involves selecting the most appropriate feature at each step to split the data. This selection is based on metrics such as **Information Gain**, **Entropy**, or **Gini Index**, which measure how well a feature separates the data into distinct classes or reduces prediction error. The process continues recursively until the data is perfectly classified or a predefined stopping condition is reached. Decision Trees are highly intuitive and easy to interpret, as they closely resemble human decision-making and allow clear visualization of the decision process.

One of the major advantages of Decision Trees is that they require minimal data preprocessing and can handle both numerical and categorical data. They are also capable of capturing non-linear relationships between variables. However, Decision Trees are highly sensitive to variations in the training data and are prone to **overfitting**, especially when the tree becomes too deep. This limitation often leads to poor generalization on unseen data. Despite this drawback, Decision Trees serve as an important foundational algorithm and are widely used either as standalone models or as base learners in ensemble techniques such as Random Forests.

2. Naïve Bayes

Naive Bayes is a simple yet powerful Machine Learning algorithm that comes under the **Supervised Learning** category and is primarily used for **classification** problems. It is based on **Bayes' Theorem**, which describes the probability of an event occurring based on prior knowledge of conditions related to that event. The algorithm is termed "naive" because it assumes that all input features are **independent of each other**, which is rarely true in real-world datasets, yet the model performs remarkably well in many practical applications.

Naive Bayes works by calculating the posterior probability of each class given the input features and then assigning the class with the highest probability to the data point. It uses prior probabilities of classes and likelihood probabilities derived from the training data. Depending on the nature of the data, different variants of Naive Bayes such as **Gaussian Naive Bayes**, **Multinomial Naive Bayes**, and **Bernoulli Naive Bayes** can be used. Gaussian Naive Bayes is

commonly applied when features are continuous and normally distributed, while Multinomial and Bernoulli Naive Bayes are used for text and binary data respectively.

One of the key advantages of Naive Bayes is its **computational efficiency** and ability to perform well even with small datasets. It requires minimal training time and works effectively in high-dimensional spaces. However, its strong independence assumption can limit accuracy when features are highly correlated. Despite this limitation, Naive Bayes remains a popular choice for tasks such as spam detection, sentiment analysis, and medical diagnosis due to its simplicity, scalability, and reliable performance.

3.K-NEAREST NEIGHBOUR

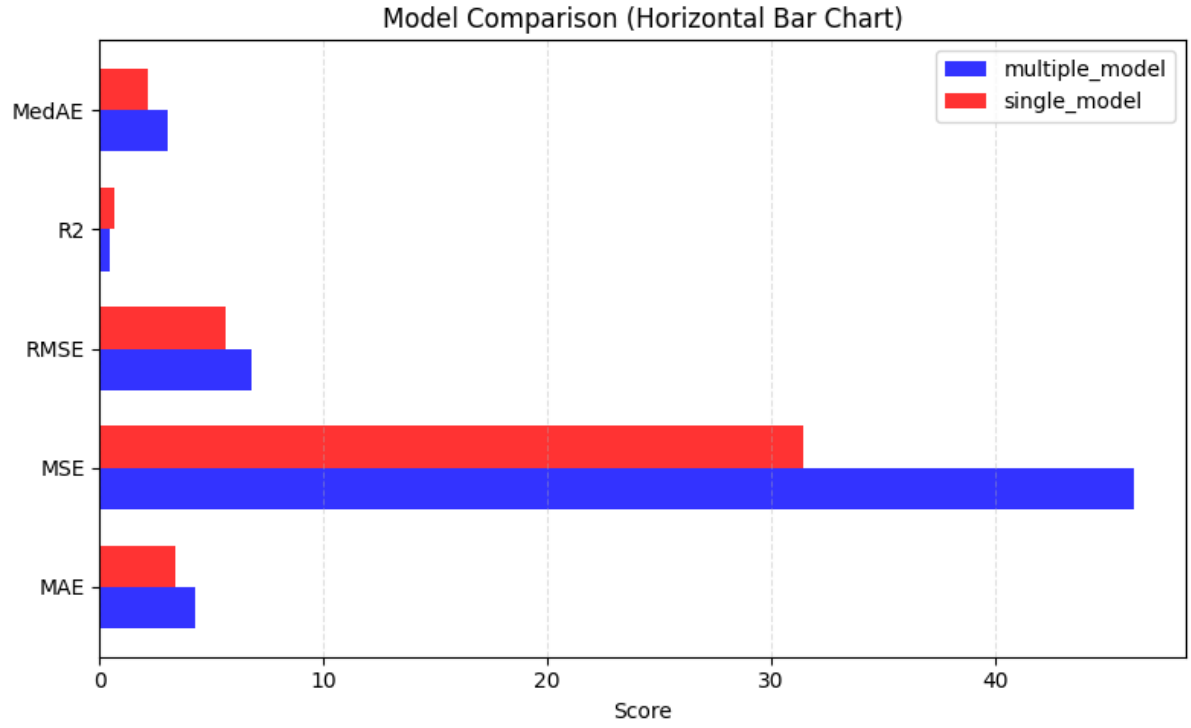
K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

4.LOGISTIC REGRESSION

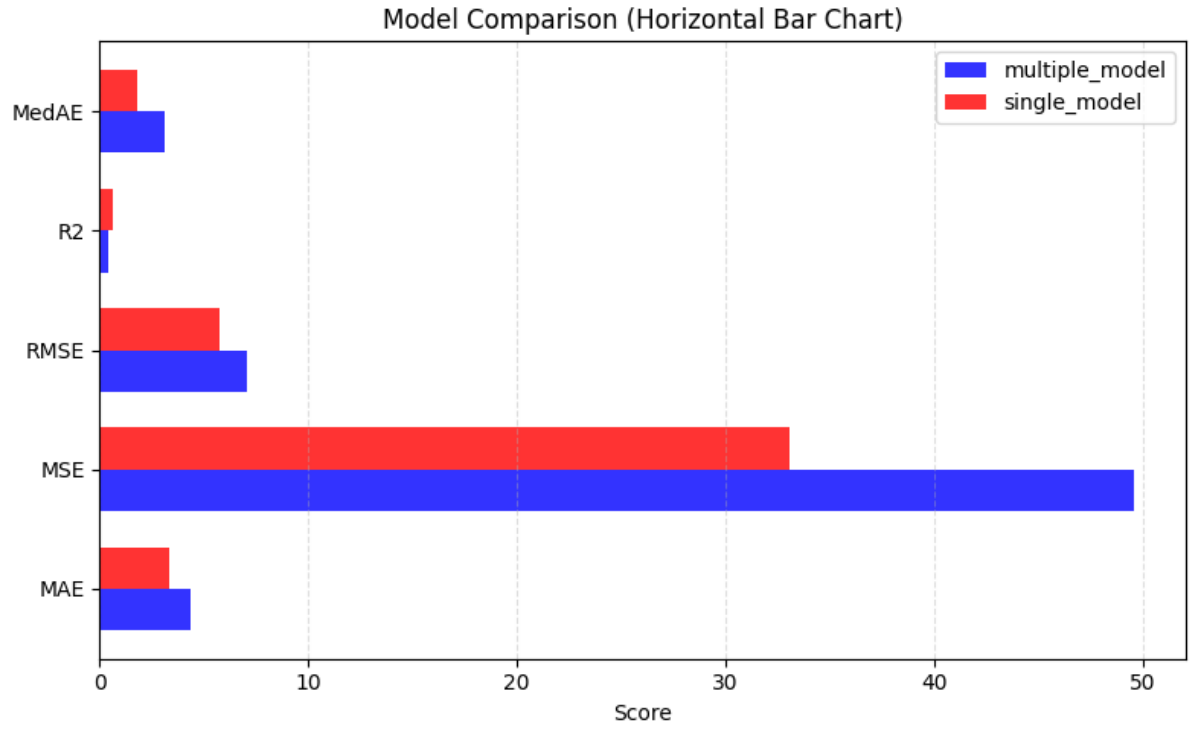
Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Results:

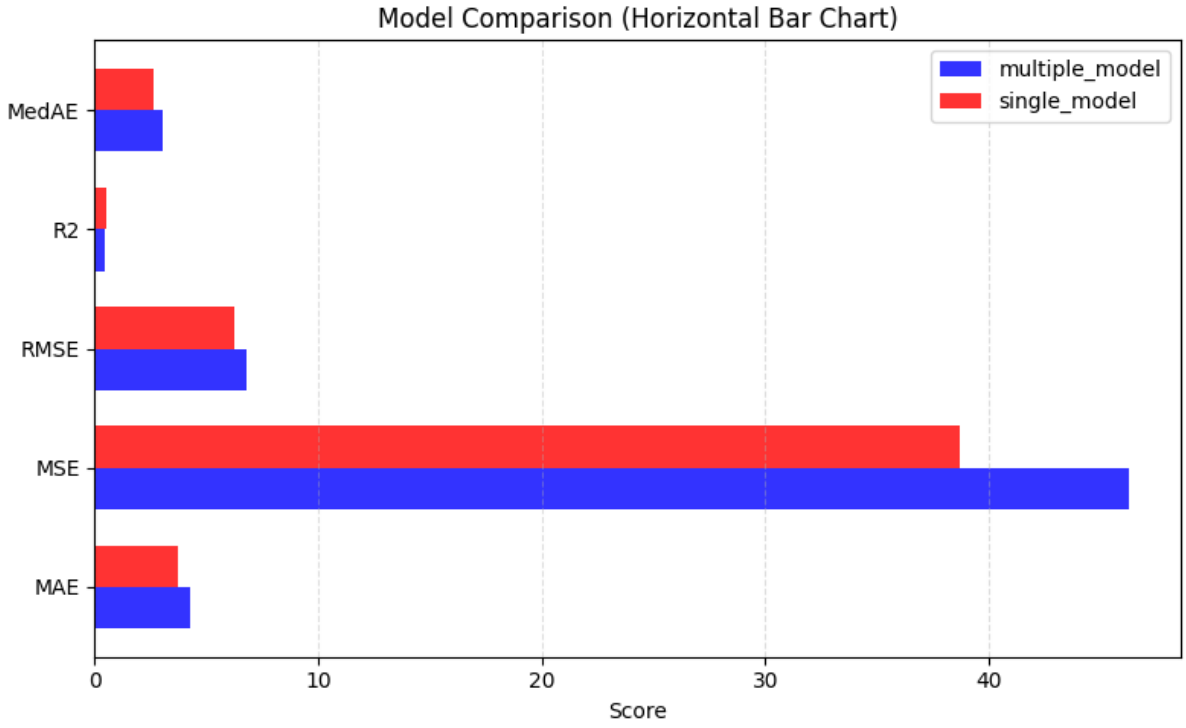
Linear Regression:



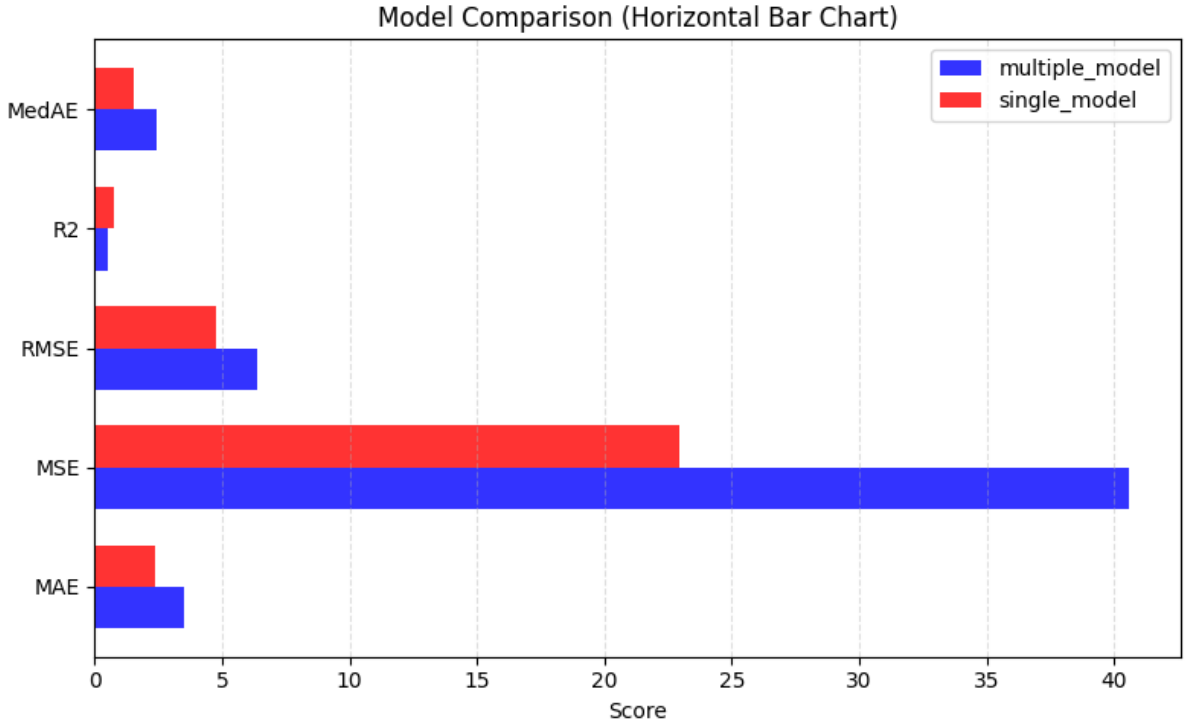
Lasso Regression:



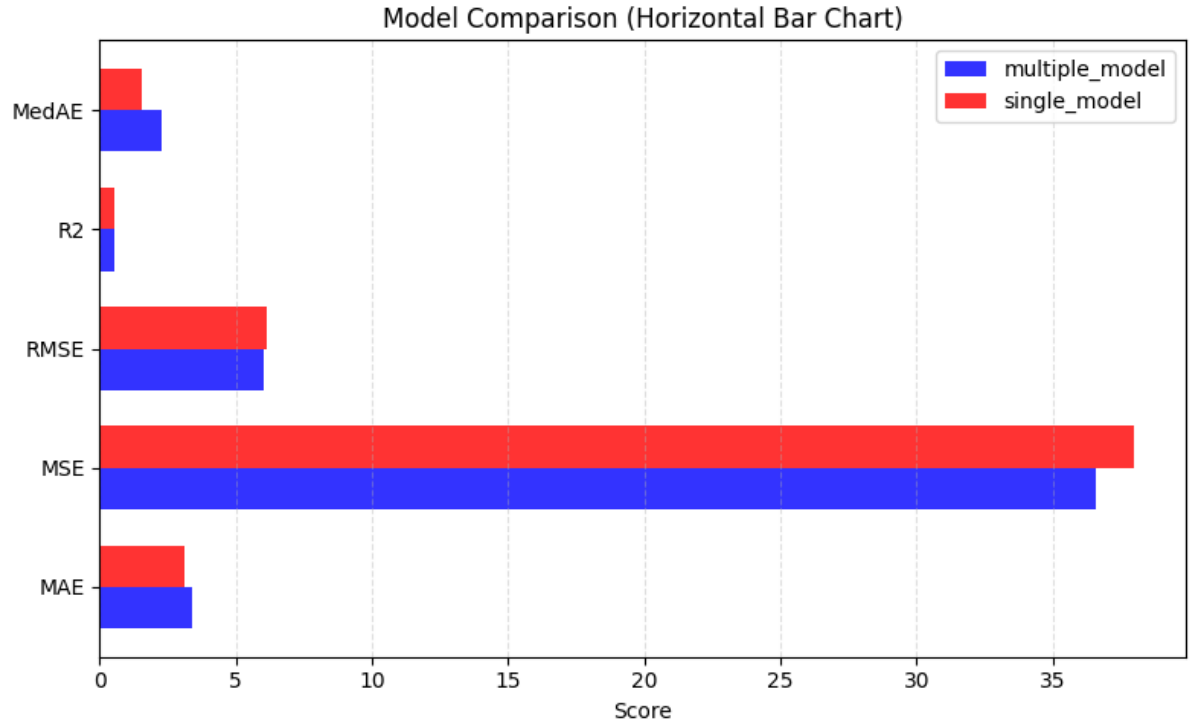
Ridge Regression:



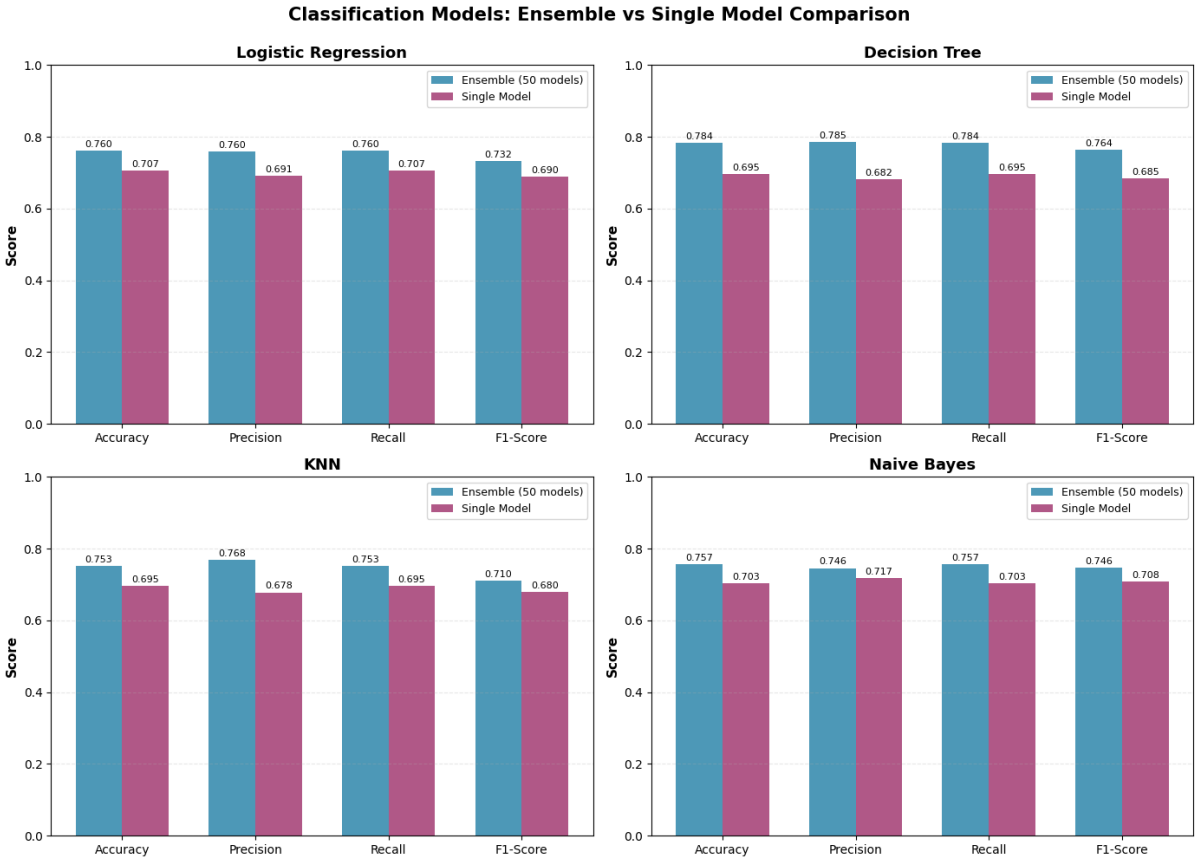
SVM model:



KNN:



Classification Models: Ensemble V/S Single Model comparision:



IMPLEMENTATION

EXISTING SYSTEM

In traditional machine learning approaches, predictions are usually made using a **single model** such as Logistic Regression, Decision Tree, KNN, or Naive Bayes. While these models are simple and easy to implement, they often suffer from limitations such as **overfitting**, **high variance**, and **sensitivity to noise in data**. A single model may perform well on training data but fail to generalize effectively on unseen data, leading to lower prediction accuracy. Additionally, single models rely heavily on the assumptions of the algorithm, which may not hold true for complex real-world datasets. As a result, the overall reliability and robustness of traditional systems remain limited

PROPOSED SYSTEM

The proposed system overcomes the limitations of traditional approaches by implementing **multiple machine learning models** and comparing their performance with an **ensemble learning technique**, specifically the Random Forest algorithm. Instead of relying on a single predictor, the system combines the strengths of multiple models to improve accuracy and generalization. The workflow of the proposed system begins with data collection and preprocessing to ensure clean and reliable input data. Feature selection is then performed to identify the most relevant attributes. By aggregating the predictions of multiple decision trees, the Random Forest model reduces overfitting, minimizes error, and provides more stable and accurate results. This approach results in improved performance across evaluation metrics such as accuracy, precision, recall, and F1-score..

1. Data Collection

2. Data Pre-Processing

3. Feature Selection

4. Model Selection

DATA COLLECTION

t is the primary and most crucial step in applying machine learning and data analytics. The data required for this project consists of structured numerical features suitable for both regression and classification tasks. The dataset used in this project has been collected from **Kaggle** and contains real-world housing-related information required for predictive analysis. The features in the dataset include attributes such as crime rate, residential zoning, number of rooms, distance to employment centers, property tax rate, pupil–teacher ratio, and percentage of lower-status population. The dataset consists of **506 observations with 14 attributes**, where one attribute represents the target variable and the remaining attributes are used as input features.

DATA PRE-PROCESSING

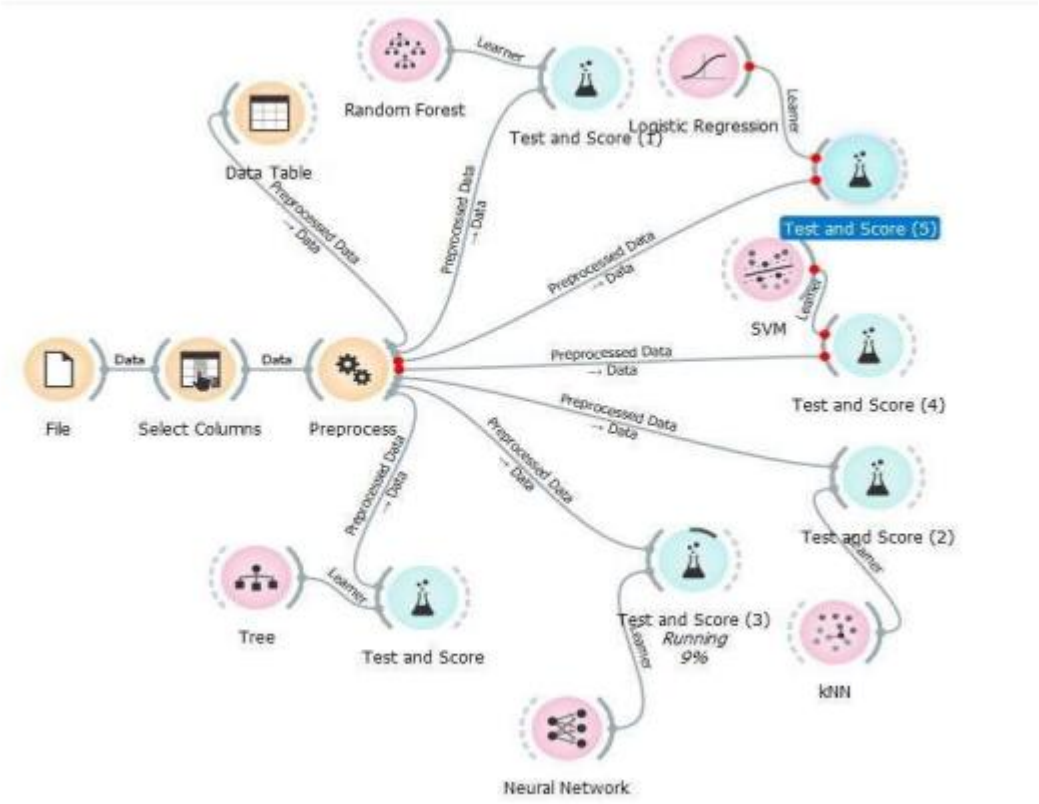
This is one of the most crucial tasks in the process of analytics. Often it is observed that more than half of the total time of analytics process is taken by pre-processing phase. It is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.

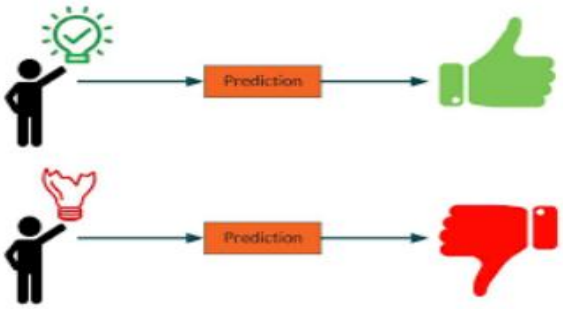
FEATURE SELECTION

Once we have the required data, next step is featuring extraction. Many times, it happens that some features do not contribute in evaluation or have negative impact on the accuracy. Feature selection is the step where we try to reduce number of features and try to create new features from existing ones. These new features now created should summarize the information obtained from existing features. The final features to be considered while prediction can be identified using correlation matrix.

MODEL SELECTION

It is the process to select one final algorithm for concerned purpose. It is decided by observing the accuracy by applying multiple algorithms .We can use logistic regression, KNN, random forest, etc. The final accuracy depends of the type of model we select. While selecting the algorithm we have to compare the accuracies. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.





In this project, we have compared following ML algorithms and obtained corresponding accuracies:

Naive Bayes

- **Ensemble (50 models): 0.7720%**
- **Single Model: 0.7400%**

Decision Tree (DT)

- **Ensemble (50 models): 0.8000%**
- **Single Model: 0.7320%**

K-Nearest Neighbors (KNN)

- **Ensemble (50 models): 0.7680%**
- **Single Model: 0.7320%**

Logistic Regression

- **Ensemble (50 models): 0.7760%**
- **Single Model: 0.7440%**

DEPLOYMENT

1.HARDWARE PLATFORM USED

The hardware requirement may serve as the basis for a contract for the implementation of the system and should therefore be complete and consistent in specification. The hardware used for the system is mentioned below.

- PROCESSOR: Intel CORE i3 or above
- RAM: minimum 4.00GB
- HARD DISK: minimum 100GB It should be noted that better the hardware facilities available, higher would-be response time of the system.

2.LIBRARIES AND SOFTWARE PLATFORM USED

The software requirement document is the specification of the system. The software requirement provides a basis for creating the software requirements specification.

OPERATING SYSTEM: Windows

SYSTEM TYPE: 64-bit, intel

CORE i5 SOFTWARE: VS Code.

TECHNOLOGIES: Python

LIBRARIES: Flask, pandas, NumPy, pickle, sklearn, etc

3.VISUALIZATION RESULTS

Based on the findings obtained from various algorithms used for identifying patients who have been diagnosed with heart disease, it is observed that KNN, Random Forest Classifier, Logistic Regression have provided better results as compared to other techniques such as Logistic Regression, KNN, naïve bayes and Decision Tree. These algorithms are not only accurate but more cost-effective and faster than the algorithms used in previous research studies. The highest level of accuracy possible by Random Forest is either greater than or nearly equal to the accuracy that were obtained from earlier research studies. It can be inferred that the improvement in accuracy is due to the increased number of attributes used from the medical dataset that was used in the project.

CONCLUSION

This project successfully demonstrates the effectiveness of the **Random Forest ensemble technique for regression and classification using multiple machine learning models**. By implementing and comparing traditional single models such as Logistic Regression, Decision Tree, K-Nearest Neighbors, and Naive Bayes with their ensemble counterparts, the study highlights the advantages of ensemble learning in improving predictive performance.

The experimental results clearly show that ensemble models consistently outperform single models across key evaluation metrics such as accuracy, precision, recall, and F1-score. The Random Forest ensemble technique reduces overfitting, minimizes variance, and enhances generalization by combining the predictions of multiple models. Among the algorithms tested, ensemble-based Decision Tree and Logistic Regression models achieved the highest improvements, validating the robustness of ensemble learning for real-world datasets.

Furthermore, the project demonstrates that ensemble methods are well-suited for handling complex and noisy data, making them reliable for both regression and classification tasks. The results obtained from this project confirm that using multiple models in an ensemble framework leads to more stable and accurate predictions compared to traditional approaches.

Overall, this project proves that the Random Forest ensemble technique is an efficient and scalable solution for predictive modeling and can be effectively applied to various real-world applications in domains such as finance, healthcare, and data analytics.

FUTURE SCOPE

The scope of this project can be extended in several directions to further improve performance and applicability. In the future, advanced ensemble techniques such as **Gradient Boosting, XGBoost, LightGBM, and AdaBoost** can be implemented and compared with Random Forest to achieve even higher accuracy. Hyperparameter optimization techniques like **Grid Search, Random Search, or Bayesian Optimization** can be applied to fine-tune model parameters and enhance prediction performance.

The project can also be extended to handle **larger and real-time datasets**, enabling the system to be used in real-world decision-making environments. Integration with **deep learning models** for feature extraction, followed by ensemble learning, can further improve predictive accuracy for complex datasets. Additionally, deploying the model as a **web-based or cloud-based application** using frameworks such as Flask or FastAPI would make the system more accessible and practical for end users.

Future work may also include incorporating **model explainability techniques** such as SHAP or LIME to improve transparency and interpretability of ensemble predictions. This would increase trust in the system, especially in sensitive domains like finance and healthcare. Overall, the proposed system provides a strong foundation for future research and development in ensemble-based machine learning solutions.