International Journal of Core Engineering & Management
Volume-8, Issue-01, 2025 ISSN No: 2348-9510

A COMPARATIVE STUDY OF MACHINE LEARNING MODELS FOR HEART DISEASE PREDICTION

Shresta Malviya
Independent Researcher, Frisco, Texas, US
malviyashresta@gmail.com

Abstract

Heart disease, one of the leading causes of death in the world, is a problem that is rising among the population. The early diagnosis can help mitigate outcomes such as heart attacks. By using efficient Machine Learning models, the early detection and diagnosis can save lives. This study deals with the comparison of classification models, Logistic Regression and Random Forest, in order to prevent fatal risks of heart disease. Logistic Regression classifies based on the probability of the target being met. Random Forest classifies based on decision trees and bootstrapping. Both have their unique advantages and drawbacks. Logistic Regression is easy to set up and efficient but assumes a linear relationship. Random Forest is flexible and resistant to overfitting but too many trees can slow down the real-time predictions, reducing efficiency. 6 models were created in this study, 3 Logistic Regression models and 3 Random Forest models. The difference among the models was the features given into the model. The study concludes that Logistic Regression performs better than Random Forest for heart disease prediction.
Index Terms – heart disease, comparison, Logistic Regression, Random Forest, machine learning

I. INTRODUCTION

According to the World Health Organization, heart disease is one of the leading causes of death in the world [1]. "Heart Disease" is a term referring to any disease in which the heart is not completely efficient. The latter stages of heart disease can lead to heart attacks and heart failures [2[. The prediction of heart disease in a timely manner can help mitigate this disastrous outcome. The application of machine learning (ML) algorithms in the prediction and diagnosis of heart disease among patients allows timely diagnosis and analysis [3]

This study compares two classification models, logistic regression and random forest, to mitigate the serious risks associated with heart disease. Using the Cleveland Heart Disease dataset sourced from Kaggle, In addition, the research explores the significance of specific features and data within the models. In total, six models were developed for this study—three using Logistic Regression and three using Random Forest—differentiated by the features included in each model.

International Journal of Core Engineering & Management
Volume-8, Issue-01, 2025 ISSN No: 2348-9510

II. DATA SET

1. Data Set Understanding

Feature	Definition	Value
age	Age	29-77 (years)
sex	Gender	0: Female 1: Male
cp	Chest Pain Level	0: Typical Angina 1: Atypical Angina 2: Non-Anginal Pain 3: Asymptomatic
trestbps	Resting Blood Pressure(BP)	94-200 (mm of Mercury)
chol	Serum Cholesterol	126-564(mg/dL)
fbs	Fasting Blood Sugar greater than 120 mg/dL	0: False 1: True
restecg	Resting Electrocardiographic Results	0: Normal 1: ST-T wave abnormality 2: Visible, or probability of, left ventricular hypertrophy
thalach	Maximum Heart Rate	71 to 202 (BPM)
exang	Exercise Induced Angina	0: False 1: True
oldpeak	Stress Test Depression induced by Exercise relative to rest	0 to 6.2 (mm)
slope	Slope of the Peak Exercise ST Segment	0: Up / Positive 1: Flat / Horizontal 2: Down / Negative
ca	Major Vessels	0 to 3 (major vessels)
thal	Thallium Heart Rate	0: Normal 1: Fixed Defect 2: Reversible Defect
target	Diagnosis of Heart Disease	0: Healthy heart 1: Diseased heart

Table 1: Values

The data set is the Cleveland data set obtained from the Kaggle ML repository. It has 14 features

International Journal of Core Engineering & Management
Volume-8, Issue-01, 2025 ISSN No: 2348-9510

and 303 rows. The 14th feature is the target, 0 stands for healthy heart while a 1 stands for a diseased heart. The features and their definitions are below.

2. Data Distributions and Cleaning

To ensure the highest performance among the models, skews in the data set and rows with outliers, rows with missing data and rows with repeated data must be fixed. If the data is skewed too much towards one of the values (example : 90% male and 10% female) then the data set is not fit for a proper prediction and diagnosis. If too many rows are repeated then this data set is unworthy for the study. If continuous values are missing then the average of that feature will be assumed; however, if a discrete value is missing then the whole row will be deleted. Outliers in the data can skew our predictions, if there are too many extreme outliers the data set is not fit for the study; however, outliers only matter for the continuous values.

```
df.isna().sum()

0
```

Figure 1: Missing Values

```
duplicated=df[df.duplicated(keep=False)]
duplicated.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
163	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1
164	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1

Figure 2: Duplicated Values

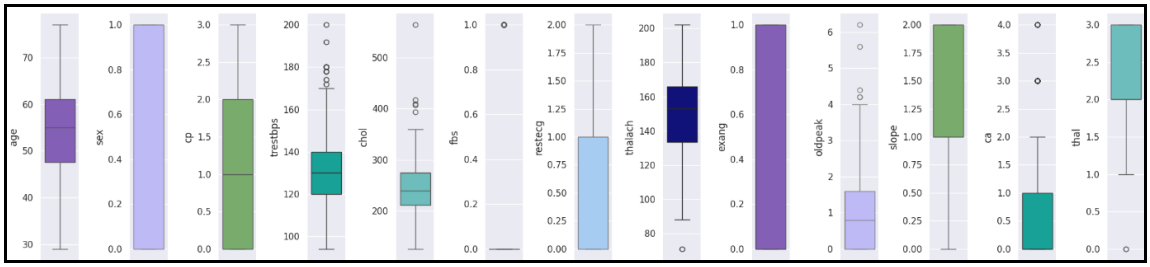

Figure 3: Outliers

Based on figures 1, 2 and 3 the data set is fit to use for the study: there are no missing values, there is only 1 duplicate out of the 303 rows, there aren't extreme outliers that can affect our model.

International Journal of Core Engineering & Management
Volume-8, Issue-01, 2025 ISSN No: 2348-9510

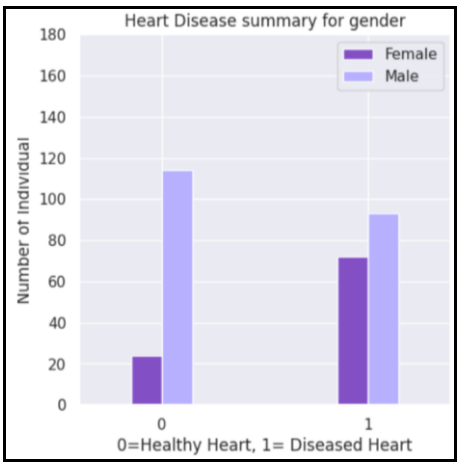

Figure 4: Gender Distribution

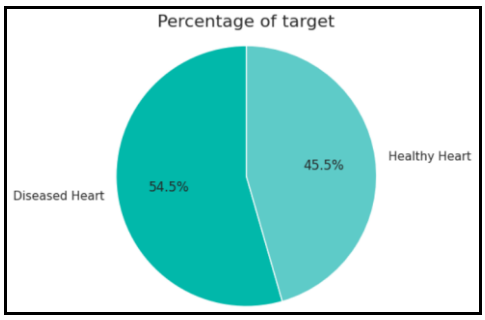

Figure 5: Target Distribution

The distributions in figures 4 and 5 show that the percentage of diseased heart to healthy heart is almost 1:1 preventing extreme skews; the female to male ratio isn't equal, it is possible for skews to occur. Though the data set isn't ideal, it is worthy to be used for the study.

III. MODELS AND METHODOLOGY

Logistic Regression(LR) and Random Forest(RF) are the two models compared in this project. LR is known to quickly provide accurate predictions based on mathematical possibility [4]. RF used decision trees and bootstrapping to provide accurate decisions [5]. Both are classifiers, they will determine whether a patient has a healthy heart or a diseased heart. This project has 6 models, 3 are LR models and 3 are with RF.

International Journal of Core Engineering & Management
Volume-8, Issue-01, 2025 ISSN No: 2348-9510

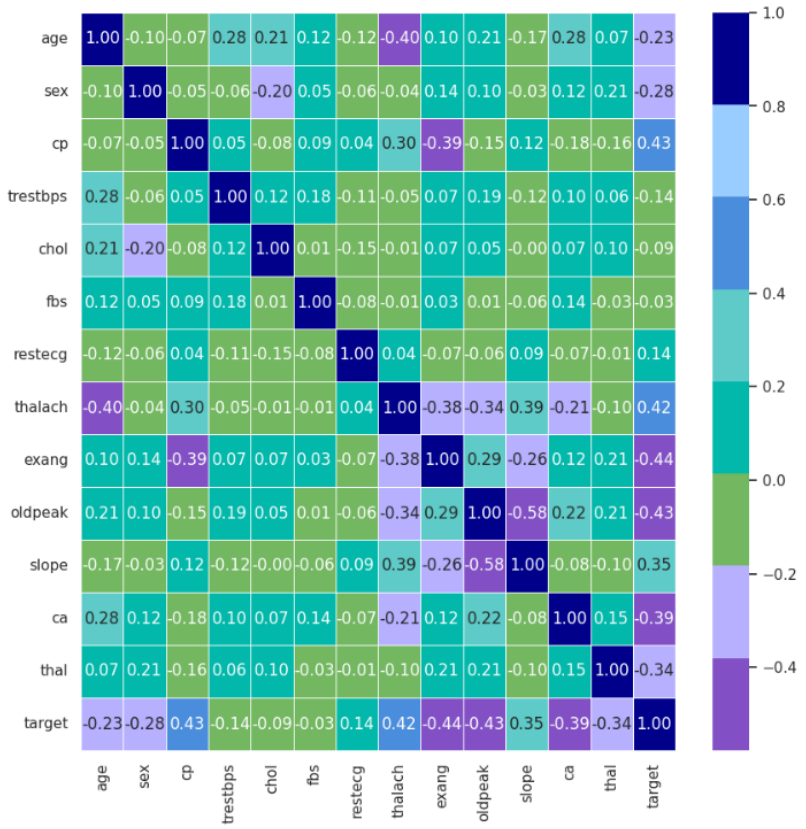

Figure 6: Correlation Map

Figure 6 shows that the 2 most positively correlated features are cp and thalach, while the 2 most negatively correlated features are exang and oldpeak. LR-All-Feature (LRAF) and RF-All Feature (RFAR) are given all the 13 features from the data set. LR-Correlated-Features (LRCF) and RF-Correlated-Features (RFCF) are only given the 4 most correlated features (thalach, cp, exang and oldpeak) from the data set: 2 of the most positively correlated and 2 of the most negatively correlated. LR-2-Correlated-Features (LR2CF) and RF-2-Correlated-Features (RF2CF) are only given the 2 most positively correlated features: cp and thalach. The models are split 80% for training and 20% for testing. The rows for testing and training are decided randomly.

For the comparative study, the dataset was taken from the Kaggle repository as comma separated value format. Then, data was processed using a Google Colab Notebook. Python libraries such as pandas, matplotlib, seaborn, sklearn and numpy were used for processing the algorithms and analyzing the data in Exploratory Analysis. After splitting the testing and training data to 80% and 20% respectively, the LRAF and RFAF models were created. Using the correlation heat map (figure 6) the 4 most correlated features were found. Using correlated features the LRCF, RFCF, LR2CF, and RF2CF models were created.

IV. MODEL EVALUATION

1. Key Metrics

International Journal of Core Engineering & Management
Volume-8, Issue-01, 2025 ISSN No: 2348-9510

In order to evaluate which model is most accurate for the diagnosis of heart disease the key metrics will be Accuracy, Recall, AUC, Precision and F1 Score on the test data. Of these the focus will be Recall in the presence of a diseased heart (Recall(1)) and Accuracy. This is due to the fact that a higher Recall(1) will prevent diseased patients from being diagnosed as healthy. Accuracy is important as it evaluates how well the model predicts heart disease overall. A Recall(1) of 1 means that our model did not predict any diseased patients as healthy. A lower Recall(1) defeats the purpose of our model as it does not mitigate the effects of a delayed diagnosis.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (1)$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{All\ Samples} \quad (2)$$

2. Model Comparison

The worst models were RF2CF and LR2CF as they had the lowest Accuracy and lowest Recall(1) values. It is seen that LRAF, LRCF and RFAF had the highest Accuracy while LRCF and RFCF had the highest Recall(1) values.

COMPARISON OF ML MODELS							
METRICS	LR-AF	LR-CF	LR-2CF	RF-AF	RF-CF	RF-2CF	Best Score
Accuracy Train Data	0.85	0.79	0.77	**1**	0.98	0.88	High (Best=1)
Accuracy Test Data	**0.82**	**0.82**	0.67	0.79	0.77	0.67	High (Best=0.82)
Precision_0	0.79	**0.84**	0.62	0.76	0.77	0.65	High (Best=0.84)
Recall_0	**0.82**	0.75	0.71	0.79	0.71	0.61	High(Best=0.82)
F1_Score_0	**0.81**	0.79	0.67	0.77	0.74	0.63	High(Best=0.81)
Precision_1	**0.84**	0.81	0.72	0.81	0.77	0.69	High(Best=0.84)
Recall_1	0.82	**0.88**	0.64	0.79	0.82	0.73	High(Best=0.88)
F1_Score_1	0.83	**0.84**	0.68	0.80	0.79	0.71	High(Best=0.84)
AUC	**0.89**	0.86	0.76	0.88	0.83	0.70	High(Best=0.89)

Figure 7: Comparison of Models

Figure 7 shows the exact key metric values. LRCF is the model with the highest Accuracy on test data and a higher Recall(1) value; while LRAF may have high performance in key metrics other than Recall(1), LRCF is the most efficient model among the 6 models. Since both LRAF and LRCF were 2 of the most efficient models, Logistic Regression models do comparably better than Random Forest models.

3. Limitations and Challenges

LR and RF have both advantages and disadvantages. Keep in mind, LR assumes linearity while RF uses immense amounts of memory due to its computational complexity. LR's limitations only reduce accuracy by a marginal amount while RF's has virtually no effect. Other than model disadvantages, the data set also plays a crucial role. While data set bias is inevitable, during exploratory data analysis, efforts to eradicate bias were made. Thus, the model comparisons are accurate.

International Journal of Core Engineering & Management
Volume-8, Issue-01, 2025 ISSN No: 2348-9510

V. CONCLUSION

The Logistic Regression classifier operates on the probability of meeting the target, while the Random Forest classifier is based on decision trees and bootstrapping methods. Each model has its distinct advantages and limitations: Logistic Regression is straightforward and efficient but relies on the assumption of a linear relationship, whereas Random Forest is flexible and less prone to overfitting, though using too many trees can hinder real-time prediction efficiency.

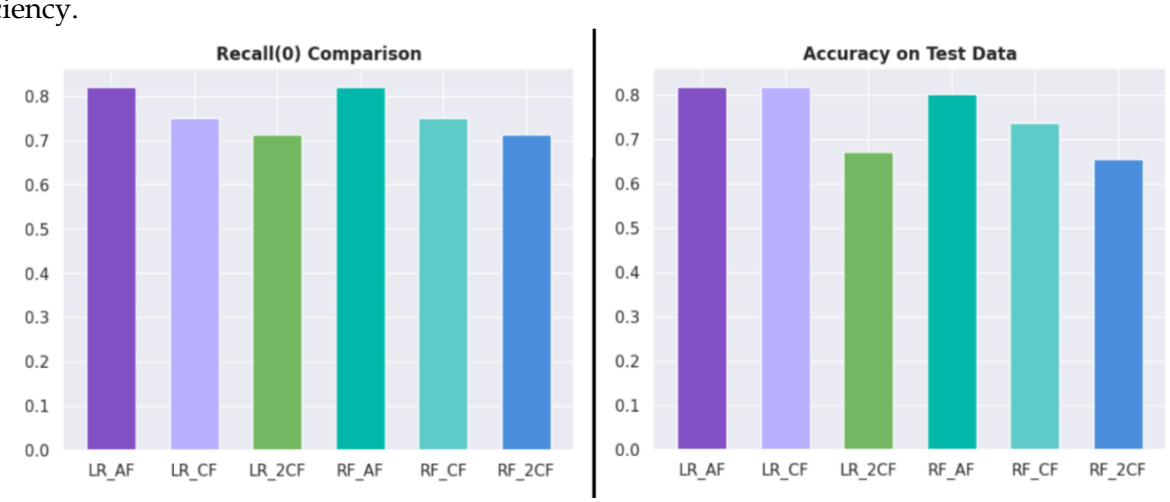

Figure 8: Metrics Comparison

Figure 8 shows that Logistic Regression with the 2 most positively correlated features and Random Forest with the 2 most positively correlated features (LR2CF and RF2CF respectively) performed the worst: they had the lowest Recall(1) values and lowest Accuracy on test data. Since the only difference among these 2 models compared to the others was the amount of data given, it is clear that giving a model too little data, while being correlated, doesn't help increase the performance of the model and can instead decrease the performance. Therefore, it is advisory to ensure the classification model is given more than 2 positively correlated features.

The performance of the Random Forest models that weren't given only 2 features (RFAF and RFCF) was lower than that of the Logistic Regression models. Upon looking at the Accuracy for training data it is visible that both the Random Forest models over fit according to the training data. Random Forest All Feature had 100% Accuracy and Random Forest with Correlated Features had an Accuracy of 98% for training data, while the Accuracy for test data dropped by almost 30%. Based on these significant differences, Random Forest is prone to be sensitive and overfit when given too much data.

Through the graphs and charts it is evident that, when given the right amount of data, Logistic Regression can perform comparably better than Random Forest. Both Logistic Regression models with all features and 4 correlated features outperformed the other 4 models. Logistic Regression with all features was more efficient than Logistic Regression with 4 correlated features with regards to all the key metrics. However, the focus of this study is on Recall(1) and Accuracy. Logistic Regression with 4 correlated features (LRCF) performed better than Logistic Regression for all features (LRAF) for Recall(1) by almost 10%. Therefore, Logistic Regression

with 4 correlated features was the best model among the 6 for the diagnosis of heart disease. This also tells us that the correlation of data to the prediction is important, giving a model with correlated data increases the performance of the model.

For the future, comparisons between K-Nearest Neighbor, Neural Networks and other models will help determine which model is the best choice for the diagnosis of heart disease.

1. Disclosure of Funding
This study has not been funded nor will I gain any financial benefit from the publication of this study. The study has been done due to the pure interest in Machine Learning and health benefits through the application of it in the medical field.

REFERENCES

1. Adizul ahmad and Huseyin. Prediction of heart disease based on machine learning using jellyfish optimization algorithm. NIH, 13(14)(PMCID10378171), 2023.
2. Sarah McDermott, Chun Shing Kwok, Hayley Burke. Missed opportunities in the diagnosis of heart failure: Evaluation of pathways to determine sources of delay to specialist evaluation. NIH,19(4)(PMCID:9169019):247–253, 2022.
3. Karen L. Howard. Artificial intelligence in healthcare: Benefits and challenges of machine learning technologies for medical diagnostics. US Government Accountability Office, 22(104629), 2022.
4. Cortina-Borja M Issitt R W and Bryant W. Classification performance of neural networks versus
5. logistic regression models: Evidence from healthcare practice. Cureus, 14(2)(22443), 2022.
6. Madhumita Pal and Smita Parija. Prediction of heart diseases using random forest. Journal of
7. Physics, 1817(012009), 2021.