Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction

Mohona Haque[1], Md. Amanullah Shah [2]
Professor Dr. Md. Musfique Anwar[3]
[1] Department of Electrical and Computer Engineering (ECE) , North South University (NSU)
Email: mohona.haque@northsouth.edu
[2]Department of Electrical and Computer Engineering (ECE) , North South University (NSU)
Email: amanullah.shah@northsouth.edu
[3] Department of Electrical and Computer Engineering (ECE) , North South University (NSU)
Email:musfique.anwar@northsouth.edu

***Abstract*—Heart disease is on the rise, and our modern lives may be why. To catch it early and save more lives, researchers are turning to machine learning (ML) to predict who's most at risk. This study throws down the gauntlet, comparing different ML algorithms to see which one is the champion at identifying potential heart disease cases. The ultimate goal? An accurate and reliable tool for doctors to use, reducing deaths from this preventable disease. The contenders in this competition are Five ML algorithms: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and k-Nearest Neighbor.**

***Index Terms*—Heart Disease Prediction, Machine Learning, Comparative Analysis, Model Evaluation, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and k-Nearest Neighbor.**

I. Introduction

The human heart is one of the most important organs in the human body. It's the device that circulates oxygen-rich blood to different parts of the body. The heart works 24 into 7 to ensure that all other organs get the right amount of oxygen-rich blood, and any interference in its functions would have an adverse effect on other organs' appropriate functioning, which can be catastrophic. Heart disease or cardiovascular disease, is a dangerous medical disorder caused by the heart's failure to perform its circulation functions properly.If a patient ignores the disease's early symptoms, which appear to be warning signs,the patient will have no time to recover and will eventually die on the spot. A heart attack is the medical term for this. It occurs because the purpose of the arteries is to give oxygen rich blood to the heart, but plaque forms as a result of fatty and other substances, which disrupts the functioning of a normal artery and converts it into a narrowed coronary artery. As a result, blood flow might be slowed or entirely stopped. Controllable risk factors and uncontrolled risk factors are the two types of risk variables that cause coronary artery disease. Diabetes, smoking, obesity or overweight, cholesterol, hypertension, less physical activities are all controllable risk factors. Age, sex, previous medical conditions and history are all uncontrollable risk factors. In the last decade, heart disease is the top cause for the death of people worldwide According to a WHO report, about 17.9 million people die each year as a result of cardiovascular disorders, with coronary heart disease and brain stroke accounting for 80% of these deaths [1]. A variety of laboratory tests and imaging examinations can be used to identify cardiovascular disease. However, the patient's medical and family history, risk factors, and physical examination are the most important aspects of diagnosis. We can synchronize the results and predict the existence of disease from findings and processes using statistical data. Doctors can make better decisions with the help of automation and deep learning.

II. Literature Review

S. Musfiq Ali et.al conducted research on the Cleveland database, with 10-fold cross validation and achieved a highest accuracy of 91.2% for GNB [2]. In 2021, A. Kondababu et. al.[3] used the Cleveland dataset to study comparative analysis and found the HRFLM technique, a combination of Random Forest(RM) and Linear Method(LM) to have the highest accuracy.Rohit Bharti et. al.[4] used a dataset with 13 features, preprocessed data using Isolation Forest, and found that KNeighbors classifier performed the best. Sfruti Sarah et. al.[5] compared various models for Heart Disease Prediction and found that LR had the maximum accuracy of 85.25%.Lubana Riyaz et. al.[6] performed a survey of various ML algorithms and their performances for heart disease prediction and found that highest average prediction accuracy was achieved by ANN with 86.91% and the lowest with C4.5 decision tree with 74.0%.Xiao-Yan Gao et. al.[7] found that bagging ensemble learning algorithm with Decision Tree and Principle component analysis feature extraction method performed the best. Abdullah et. al developed a Data Mining model to increase the accuracy for heart Disease Prediction, the model was based on RF classifier [8]. Sonam Nikhar et al.[9] used Cleveland Dataset with 303 instances and 19 attributes with GNB, DT technique.They also discovered that

Decision Tree has a higher accuracy than the Nave Bayes Classifier. Ravindra Yadav et al.[10] deployed ML approach for the Cardio Vascular Disease Prediction Survey, which included the DT, GNB , Neural Networks , Deep Learning and SVM. The decision tree's conclusion is generated using ID3, CART Cpercent.0,CYT , and J48. Devansh Shah et.al,[11] used Cleveland database with 303 instances and 14 attributes for heart disease prediction. K-NN algorithm had the highest accuracy. Archana Singh et al. [12] employed the Cleveland dataset resulting in the following results: Linear Regression: 78 percent, DT: 79 percent, SVM: 83 percent, and K-NN: 87 percent accuracy.

III. Dataset

The dataset that was used in this study is available at kaggle[13] . The dataset contains 13 Attributes, excluding the predicted attribute. The "target" field is the predicted attribute making a total of 14 columns. The dataset consists of 303 patients.

- Number of Entries: The dataset consists of 303 entries, ranging from index 0 to 302.
- Columns: There are 14 columns in the dataset corresponding to various attributes of the patients and results of tests.
- Data Type: Most of the columns (13 out of 14) are of the int64 data type.Only the oldpeak column is of the float64 data type.
- Missing Values: There don't appear to be any missing values in the dataset as each column has 303 non-null entries.

Based on the data types and the feature explanations we had earlier, we can see that 9 columns (sex, cp, fbs, restecg, exang, slope, ca, thal, and target) are indeed numerical in terms of data type, but categorical in terms of their semantics. These features should be converted to string (object) data type for proper analysis and interpretation. In this study, we delve into a dataset encapsulating various health metrics from heart patients, including age, blood pressure, heart rate, and more. Our goal is to develop a predictive model capable of accurately identifying individuals with heart disease. Given the grave implications of missing a positive diagnosis, our primary emphasis is on ensuring that the model identifies all potential patients, making recall for the positive class a crucial metric.The description of attributes of our dataset is given in the Fig 1 :

S. No	Attribute	Description	Values
1	age	Patient's age in years	Value is continuous in range [29-77]
2	sex	Patient's sex	1 : male 0 : female
3	cp	Type of chest pain	0 : asymptomatic 1 : atypical angina 2 : non-angina pain 3 : typical angina
4	trestbps	Resting blood pressure of patient (mm Hg, noted on the time at which patient was admitted to the hospital)	Value is continuous in range [94-200]
5	chol	Patient's serum cholesterol measurement in mg/dl	Value is continuous in range [126-564]
6	fbs	Fasting blood sugar of patient	1 if fbs > 120 mg/dl Else 0
7	restecg	Resting electrocardiographic results	0 : normal 1 : having abnormal _ST_T wave 2 : left ventricular hypertrophy
8	thalach	Maximum heart rate achieved by patient	Value is continuous in range [71-202] bpm
9	exang	Exercise included angina	0 denotes no 1 denotes yes
10	oldpeak	Depression in ST brought by exercise that is relative to rest	Value is continuous in range [0-6.2]
11	slope	ST segment's peak exercise slope.	0 : down sloping 1 : flat 2 : up sloping
12	ca	Count of major vessels that have been colored with fluoroscopy	0-4 value
13	thal	thalassemia value, a type of blood disorder	0 : Null 1: represents a defect that is fixed. In this condition in some parts of the heart there is no blood flow 2 : blood flow is normal 3: defect is reversible.
14	target	Is the heart disease present	0 : No 1 : Yes

Fig. 1. Description of the Attributes

IV. Exploratory Data Analysis (EDA)

For our Exploratory Data Analysis (EDA), we take it in two main steps:

1) Univariate Analysis: Here, we focused on one feature at a time to understand its distribution and range.

2) Bivariate Analysis: In this step, we explored the relationship between each feature and the target variable. This helps us figure out the importance and influence of each feature on the target outcome.

With these two steps, we aimed to gain insights into the individual characteristics of the data and also how each feature relates to our main goal: predicting the target variable.

A. Univariate Analysis

We undertake univariate analysis on the dataset's features, based on their datatype:

- For continuous data: We employ histograms to gain insight into the distribution of each feature. This allows us to understand the central tendency, spread, and shape of the dataset's distribution. Upon reviewing the histograms

of the continuous features and cross-referencing them with the provided feature descriptions, everything appears consistent and within expected ranges. There doesn't seem to be any noticeable noise or implausible values among the continuous variables.

- For categorical data: Bar plots are utilized to visualize the frequency of each category. This provides a clear representation of the prominence of each category within the respective feature. By employing these visualization techniques, we're better positioned to understand the individual characteristics of each feature in the dataset.

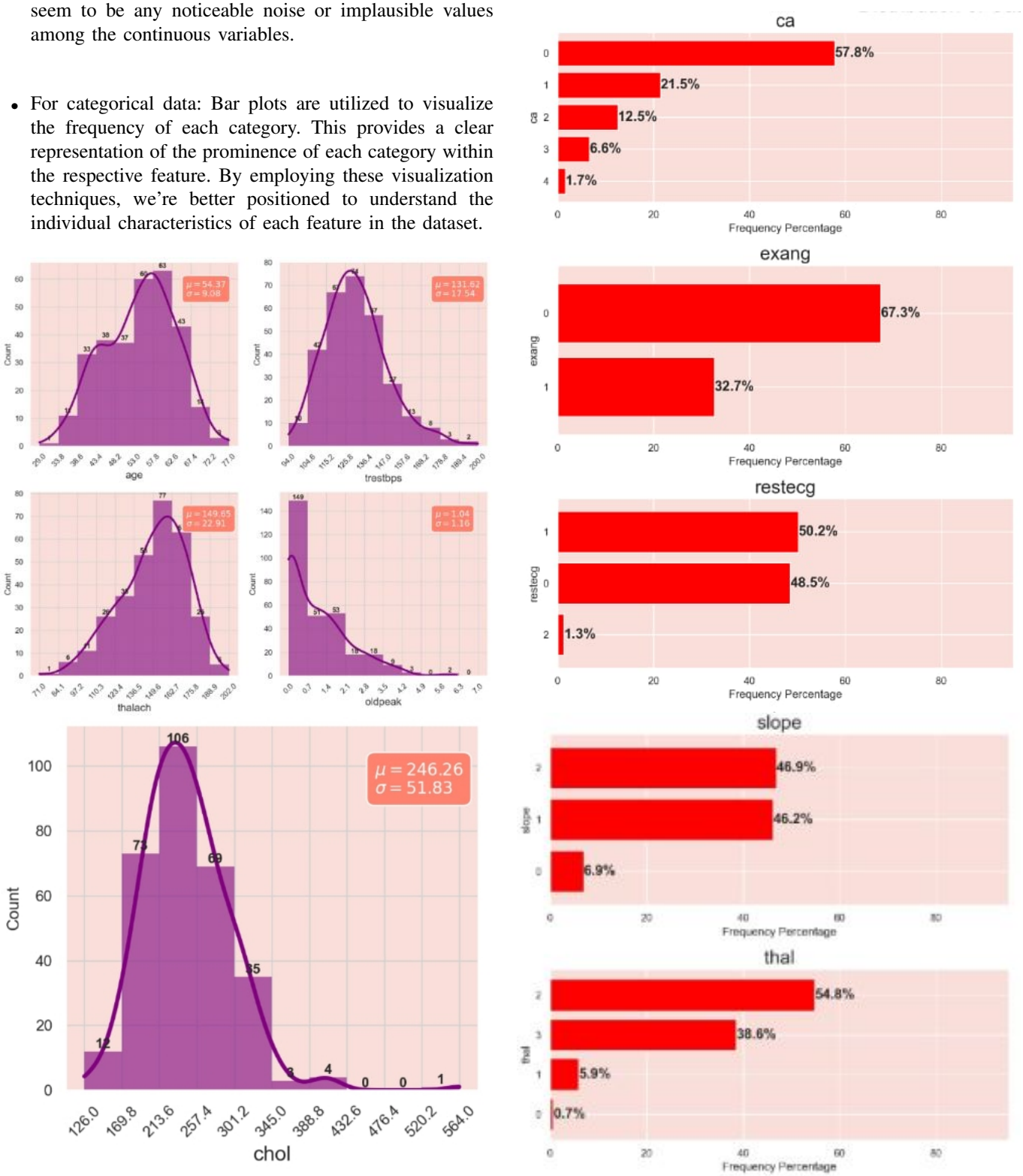

Fig. 2. Distribution Of Continous Varriable

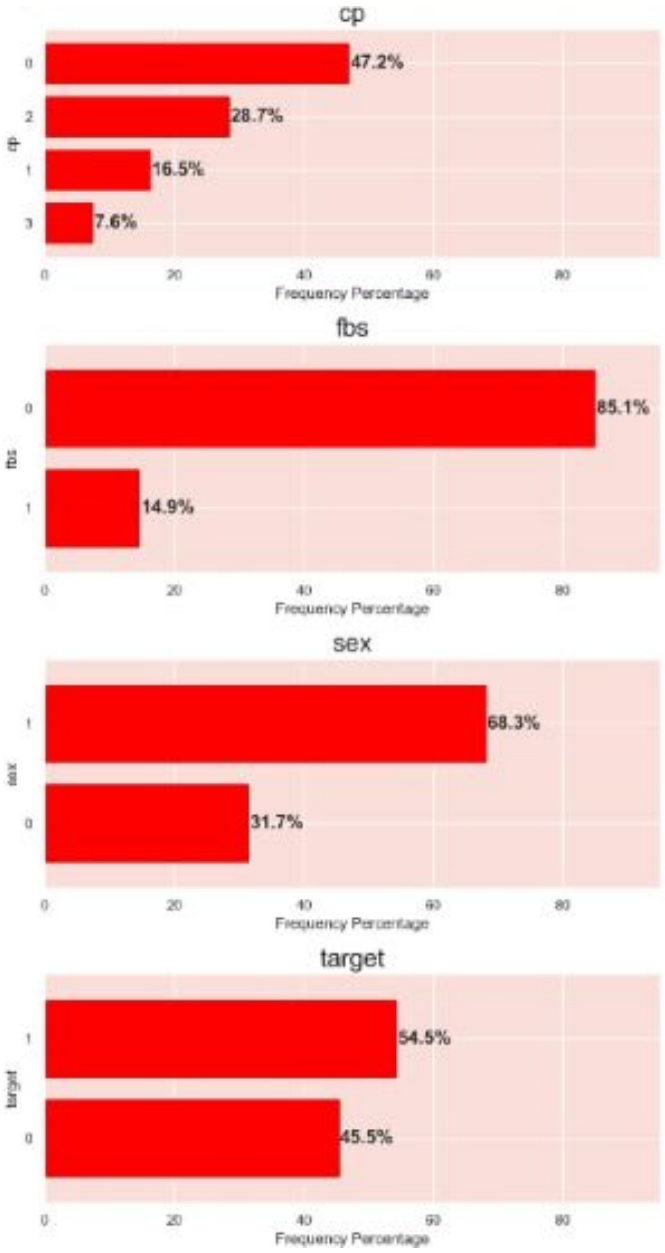

Fig. 3. Distribution Of Categorical Varriable

B. Bivariate Analysis

For our bivariate analysis on the dataset's features with respect to the target variable:

- For continuous data: We are going to use bar plots to showcase the average value of each feature for the different target classes, and KDE plots to understand the distribution of each feature across the target classes. This aids in discerning how each feature varies between the two target outcomes.
- For categorical data: We are going to employ 100% stacked bar plots to depict the proportion of each category across the target classes. This offers a comprehensive view of how different categories within a feature relate to the target.

Through these visualization techniques, we are going to gain a deeper understanding of the relationship between individual features and the target, revealing potential predictors for heart disease.

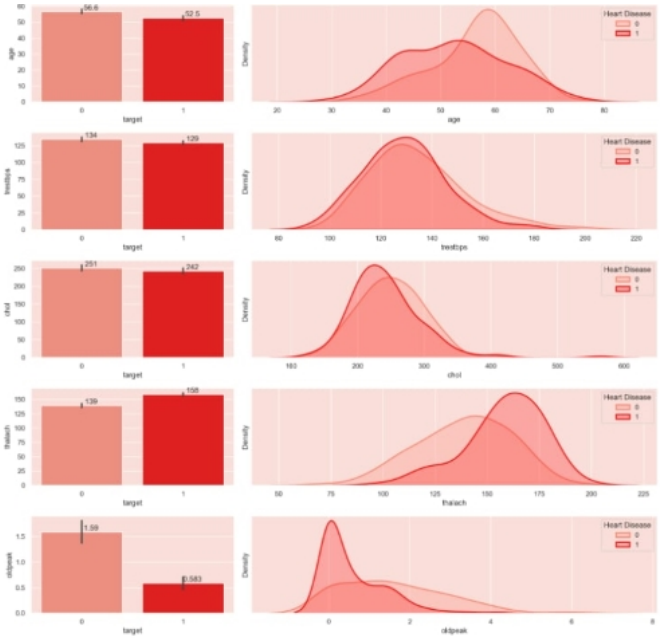

Fig. 4. Continuous Features vs Target Distribution

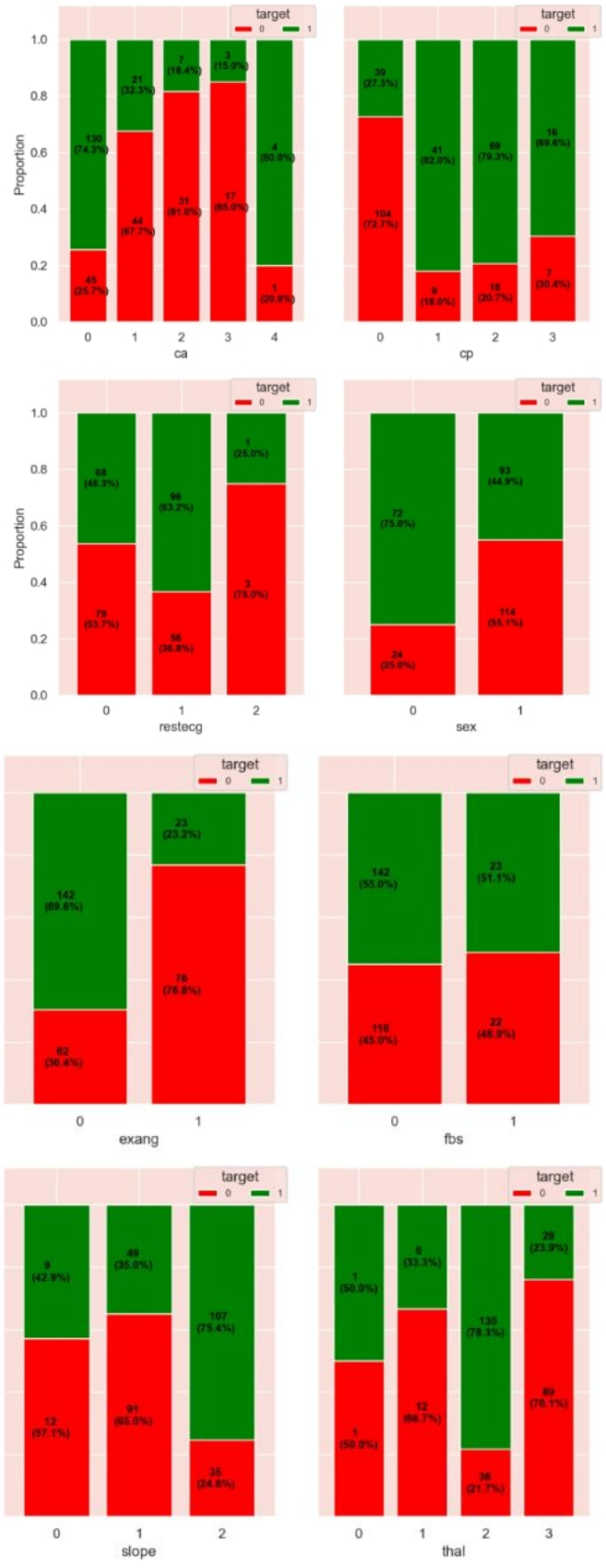

Fig. 5. Categorical Features vs Target Stacked Barplots

V. Methodology

The methodology involves the following steps:

1) Missing Value Treatment: Upon our inspection, it is obvious that there are no missing values in our dataset. This is ideal as it means we don't have to make decisions about imputation or removal, which can introduce bias or reduce our already limited dataset size.
2) Outlier Treatment: To check for outliers we use the IQR method for the continuous features. Upon identifying outliers for the specified continuous features, we found the following:
   a) trestbps: 9 outliers
   b) chol: 5 outliers
   c) thalach: 1 outlier
   d) oldpeak: 5 outliers
   e) age: No outliers

   Given the nature of the algorithms (especially SVM and KNN) and the small size of our dataset, direct removal of outliers might not be the best approach. Instead, we'll focus on applying transformations like Box-Cox in the subsequent steps to reduce the impact of outliers and make the data more suitable for modeling.
3) Categorical Features Encoding: Based on the feature descriptions, we decide to do one-hot encoding:
   a) Nominal Variables: These are variables with no inherent order. They should be one-hot encoded because using them as numbers might introduce an unintended order to the model.
   b) Ordinal Variables: These variables have an inherent order. They don't necessarily need to be one-hot encoded since their order can provide meaningful information to the model.

   Summary:
   - Need One-Hot Encoding: cp, restecg, thal
   - Don't Need One-Hot Encoding: sex, fbs, exang, slope, ca
4) Transforming Skewed Features: Box-Cox transformation is a powerful method to stabilize variance and make the data more normal-distribution-like. It's particularly useful when you're unsure about the exact nature of the distribution you're dealing with, as it can adapt itself to the best power transformation. However, the Box-Cox transformation only works for positive data, so one must be cautious when applying it to features that contain zeros or negative values. The Box-Cox transformation requires all data to be strictly positive. To transform the oldpeak feature using Box-Cox, we can add a small constant (e.g., 0.001) to ensure all values are positive.

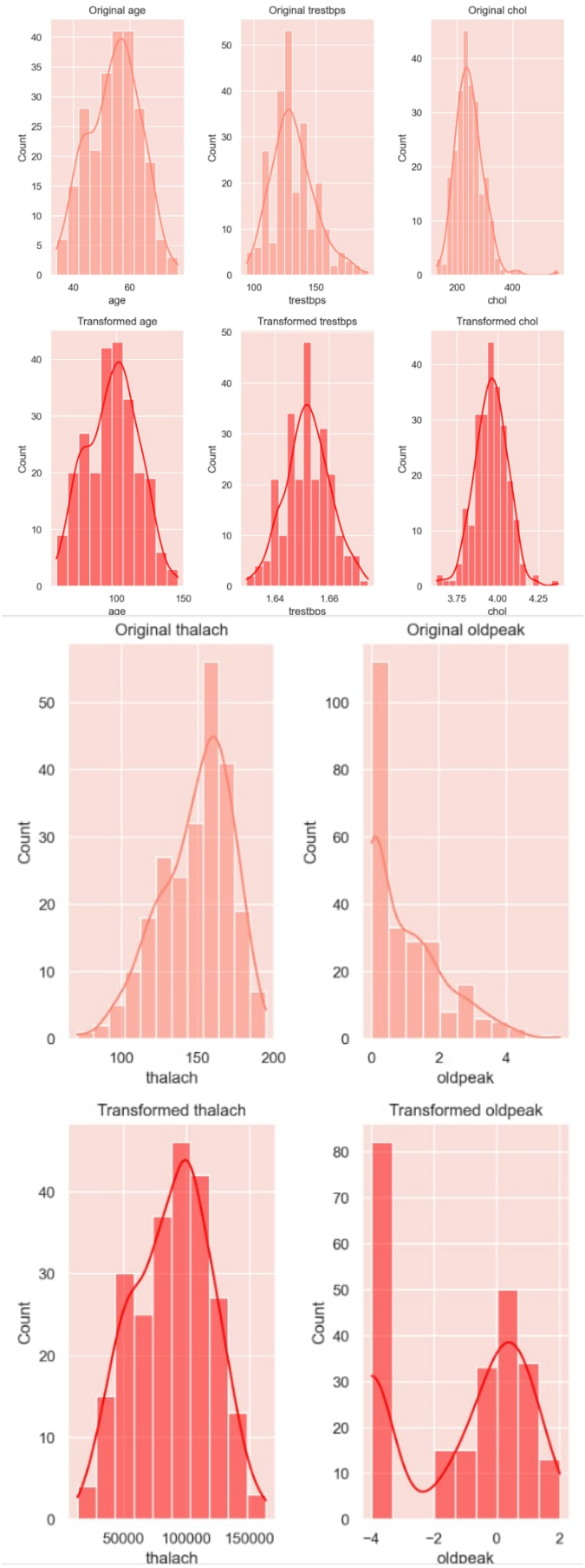

Fig. 6. Transforming Skewed Features

VI. Model Description

This section provides a brief description of each machine learning model used in the study:

A. Support Vector Machine (SVM)

A classification model that finds the hyperplane that best separates the data into classes. We evaluate our SVM model's performance on both the training and test datasets:

	precision	recall	f1-score	support
0	0.92	0.54	0.68	110
1	0.71	0.96	0.82	132
accuracy			0.77	242
macro avg	0.82	0.75	0.75	242
weighted avg	0.81	0.77	0.76	242

Fig. 7. SVM Model Evaluation on Training Data

	precision	recall	f1-score	support
0	0.94	0.57	0.71	28
1	0.73	0.97	0.83	33
accuracy			0.79	61
macro avg	0.83	0.77	0.77	61
weighted avg	0.83	0.79	0.78	61

Fig. 8. SVM Model Evaluation on Testing Data

B. Random Forest Model (RF)

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. We are setting up the hyperparameters grid and utilize the tune_clf_hyperparameters function to pinpoint the optimal hyperparameters for our RF model. We evaluate our RF model's performance on both the training and test datasets:

	precision	recall	f1-score	support
0	0.84	0.79	0.81	110
1	0.83	0.87	0.85	132
accuracy			0.83	242
macro avg	0.83	0.83	0.83	242
weighted avg	0.83	0.83	0.83	242

Fig. 9. RF Model Evaluation on Training Data

	precision	recall	f1-score	support
0	0.85	0.79	0.81	28
1	0.83	0.88	0.85	33
accuracy			0.84	61
macro avg	0.84	0.83	0.83	61
weighted avg	0.84	0.84	0.84	61

Fig. 10. RF Model Evaluation on Testing Data

C. Decision Trees

A tree-like model that splits the data based on feature values to make predictions. We evaluate our DT model's performance on both the training and test datasets:

	precision	recall	f1-score	support
0	0.73	0.75	0.74	110
1	0.78	0.77	0.78	132
accuracy			0.76	242
macro avg	0.76	0.76	0.76	242
weighted avg	0.76	0.76	0.76	242

Fig. 11. DT Model Evaluation on Training Data

	precision	recall	f1-score	support
0	0.80	0.71	0.75	28
1	0.78	0.85	0.81	33
accuracy			0.79	61
macro avg	0.79	0.78	0.78	61
weighted avg	0.79	0.79	0.79	61

Fig. 12. DT Model Evaluation on Testing Data

D. k-nearest neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values. During the training phase, the KNN algorithm stores the entire training dataset as a reference. We set up the hyperparameters grid and utilize the tune_clf_hyperparameters function to pinpoint the optimal hyperparameters for our KNN pipeline. We evaluate our KNN model's performance on both the training and test datasets:

	precision	recall	f1-score	support
0	0.80	0.79	0.79	110
1	0.83	0.83	0.83	132
accuracy			0.81	242
macro avg	0.81	0.81	0.81	242
weighted avg	0.81	0.81	0.81	242

Fig. 13. KNN Model Evaluation on Training Data

	precision	recall	f1-score	support
0	0.82	0.82	0.82	28
1	0.85	0.85	0.85	33
accuracy			0.84	61
macro avg	0.83	0.83	0.83	61
weighted avg	0.84	0.84	0.84	61

Fig. 14. KNN Model Evaluation on Testing Data

E. Logistic Regression Model

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. We evaluate our Logistic regression model's performance on both the training and test datasets:

	precision	recall	f1-score	support
0	0.81	0.83	0.82	110
1	0.85	0.84	0.85	132
accuracy			0.83	242
macro avg	0.83	0.83	0.83	242
weighted avg	0.84	0.83	0.83	242

Fig. 15. Logistic Regression Model Evaluation on Training Data

	precision	recall	f1-score	support
0	0.82	0.82	0.82	28
1	0.85	0.85	0.85	33
accuracy			0.84	61
macro avg	0.83	0.83	0.83	61
weighted avg	0.84	0.84	0.84	61

Fig. 16. Logistic Regression Model Evaluation on Testing Data

VII. Comparison of Models

The performance of each model is compared using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

Tables and graphs are used to illustrate the comparative performance of the models.

	precision_0	precision_1	recall_0	recall_1	f1_0	f1_1	macro_avg_precision	macro_avg_recall	macro_avg_f1	accuracy
SVM	0.94	0.73	0.57	0.97	0.71	0.83	0.83	0.77	0.77	0.79
RF	0.85	0.83	0.79	0.88	0.81	0.85	0.84	0.83	0.83	0.84
DT	0.80	0.78	0.71	0.85	0.75	0.81	0.79	0.78	0.78	0.79
KNN	0.82	0.85	0.82	0.85	0.82	0.85	0.83	0.83	0.83	0.84
Logistic Regression	0.82	0.85	0.82	0.85	0.82	0.85	0.83	0.83	0.83	0.84

Fig. 17. Comparison of Metrics of Different Classifiers

VIII. Results and Analysis

Metrics that have been used are accuracy, precision, recall/sensitivity/ F1 Score. Data was pre-processed using one-hot encoding which makes the data more usable and expressive and it can be rescaled easily. The SVM model demonstrates a commendable capability in recognizing potential heart patients. With a recall of 0.97 for class 1, it's evident that almost all patients with heart disease are correctly identified. This is of paramount importance in a medical setting. However, the model's balanced performance ensures that while aiming for high recall, it doesn't compromise on precision, thereby not overburdening the system with unnecessary alerts.

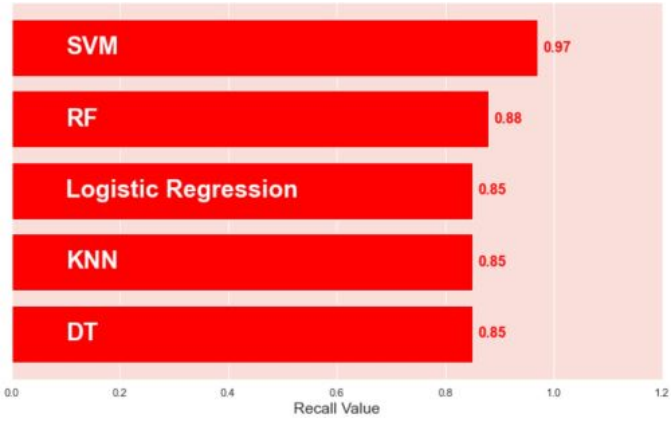

Fig. 18. Recall For Positive Class Across Model

IX. Conclusion

The goal of this research was to examine the performance of different supervised machine learning algorithms for predicting heart disease. which were "Logistic Regression" (LR), "Decision Tree" (DT), "Random Forest" (RF), "Support Vector Machine" (SVM), and "k-Nearest Neighbor" (kNN). Many prior researches on the same topic were analyzed. Data was preprocessed using feature selection, one-hot encoding and then split into testing and training data. Models were trained and various metrics were drawn. In this study it was found out that Support Vector Machine algorithm performed the best with 97% recall rate after using hyper-parameter tuning.

X. References

[1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clinical epidemiology. 2011;3:67.

[2] Ali M, Khan MD, Imran MA, Siddiki M. Heart disease prediction using machine learning algorithms (Doctoral dissertation, BRAC University).

[3] Kondababu A, Siddhartha V, Kumar BB, Penumutchi B. A comparative study on machine learning based heart disease prediction. Materials Today: Proceedings. 2021 Feb 19.

[4] Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience. 2021 Jul 1;2021.

[5] Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience. 2021 Jul 1;2021.

[6] Riyaz L, Butt MA, Zaman M, Ayob O. Heart Disease Prediction Using Machine Learning Techniques: A Quantitative Review. InInternational Conference on Innovative Computing and Communications 2022 (pp. 81-94). Springer, Singapore.

[7] Gao XY, Amin Ali A, Shaban Hassan H, Anwar EM. Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. Complexity. 2021 Feb 10;2021.

[8] Abdullah AS, Rajalaxmi R. A data mining model for predicting the coronary heart disease using random forest classifier. InInternational Conference in Recent Trends in Computational Methods, Communication and Controls 2012 Apr (pp. 22-25).

[9] Lafta R, Zhang J, Tao X, Li Y, Tseng VS. An intelligent recommender system based on short-term risk prediction for heart disease patients. In2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 2015 Dec 6 (Vol. 3, pp. 102-105). IEEE.

[10] Hasan R. Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction. InITM Web of Conferences 2021 (Vol. 40, p. 03007). EDP Sciences.

[11] Shah D. Heart Disease Prediction using Machine Learning Techniques Springer Nature Singapore Pte Ltd, 2020.

[12] Singh B, Prabhakar Tiwari SN, Singh RP, Vishwakarma M, Patel DK, Kumar A, Pratap A, Singh SP, Mishra S, Raj R, Lohia P. SN Paper ID. InInternational Conference on Electrical and Electronics Engineering (ICE3) 2020 Feb (Vol. 14, p. 15).

[13] Heart Disease Dataset: Heart Disease Dataset — Kaggle