Predictive Modeling for Heart Disease Risk Assessment: An Analysis of Machine Learning Algorithms and Tools

Faith Daniel | 21073926

Word | 2279

Table of Contents

I. ABSTRACT

This research intends to scrutinize and assess the effectiveness of diverse machine learning algorithms in binary classification task of predicting the presence or absence of heart disease. The models under study include the Decision Tree, Random Forest, XGBoost Classifier, Extra Tree, and K-nearest neighbors (KNN). This exploration involves adjusting the hyperparameters of each model and conducting cross-validation to guarantee the reliability and generalizability of the outcomes.

Machine learning tools like Feature and permutation Importance, Shap, Partial Dependence Plot (PDP), and Lime are utilized to interpret the functioning of the models and make sense of their behavior. The models' assessment is executed using classification metrics such as precision, recall, f1-score, along with the determination of specificity and sensitivity. The conclusions drawn from this examination shed light on the aptness of each algorithm for the classification task at hand, underscoring the significance of interpretability and model evaluation. Link to complete work. https://github.com/Codefedy/Heart-Disease-Predictive-Model

II. INTRODUCTION

Heart disease, a significant global health concern, has profound implications on mortality rates and public health infrastructure, as it encompasses a wide array of cardiac and vascular ailments. Symptoms of heart disease includes chest pain, palpitation, breathlessness, etc. (British Heart Foundation, 2019). This disease's impact is evidenced by the high number of annual deaths and its deep-seated repercussions.

As per the World Health Organization (WHO, 2023), cardiovascular diseases claim the top spot as the leading cause of death globally. These diseases account for nearly 17.9 million deaths per year, constituting about 31% of all deaths worldwide. Projections indicate a rise in these numbers, with anticipated annual deaths surpassing 23.6 million by 2030. These alarming figures underline the critical necessity for efficacious strategies to prevent, detect, and manage heart diseases.

In order to effectively counter heart disease, comprehension of its underlying causes and risk factors is pivotal. These risk factors are often categorized as modifiable and non-modifiable. While non-modifiable factors are inherent, modifiable factors, such as smoking, overweight, physical inactivity, and diabetes, etc. can be changed through lifestyle interventions as described by (Mohammadnezhad *et al.*, 2016)

III. DATASET OVERVIEW

The dataset used in this project was obtained from Kaggle. It consists of 270 case studies and 13 features of individuals classified as either having or not having heart disease based on results from cardiac catheterizations - the gold standard in heart health assessment. Each patient is identified by independent predictive variables revealing their age, sex, chest pain type, thallium, blood pressure, cholesterol levels, electrocardiogram results, exercise-induced angina symptoms, and the number of vessels seen on fluoroscopy showing narrowing of their coronary arteries.

	index	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
0	0	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
1	1	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
2	2	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
3	3	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
4	4	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
5	5	65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
6	6	56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
7	7	59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
8	8	60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence
9	9	63	0	4	150	407	0	2	154	0	4.0	2	3	7	Presence

Figure 1: **Description of the Data Frame**

IV. DATASET PREPROCESSING

In the data pre-processing phase, the categorical variables were converted to enhance comprehension. The one-hot encoding method was used to establish binary categorical variables. For example, 'presence' and 'absence' were substituted with binary values, where 'presence' was denoted as 1, and 'absence' was indicated as 0.

V. DISCRIPTION OF ATTRIBUTES

This study utilizes a variety of patient characteristics and medical metrics as features. They offer insights into demographics, clinical history, and physiological markers tied to heart disease risk. Our objective, by collectively evaluating these features, is to encompass a broad spectrum of potential risk factors.

Feature	Description	Possible Values
Sex	Gender of the individual	Male (1), Female (0)
Chest Pain Type	Type of chest pain	Typical Angina (1), Atypical Angina (2), Non-anginal Pain (3), Asymptomatic (4)
BP	Blood pressure rate of the individual	Continuous variable
Fbs Over 120	Fasting blood sugar level	> 120 mg/dl (1=True, 0=False)
EKG Result	Resting electrocardiographic measurement	Normal (0), ST-T wave abnormality (1), Probable or definite left ventricular hypertrophy (2)
MAX HR	Maximum heart rate	Continuous variable
Exercise Angina	Chest pain or tightness induced by exercise	Yes (1), No (0)
ST Depression	Induced by exercise relative to rest	Continuous variable
Slope of ST	The peak exercise ST segment	Upsloping (1), Flat (2), Downsloping (3)
Number of Vessels Fluro	Number of colored vessels indicating major coronary arteries	0 to 3 (each vessel represents LAD, LCX, RCA)
Thallium	A toxic heavy metal adverse effects if ingested or exposed to in large amounts	Normal (3), Fixed defect (6), Reversable defect (7)
Heart Disease (Target)	Presence or absence of heart disease	Absence (0), Presence (1)

Figure 2: Description of dataset attributes

VI. EXPLORATORY DATA ANALYSIS (EDA)

An exploratory data analysis (EDA) was carried out to obtain a comprehensive understanding of the dataset and the interplay between variables. Major steps involved in the EDA process were:

- Visualizing Data
- Analysing Correlations
- Identifying Feature Importance
- Generating Descriptive Statistics

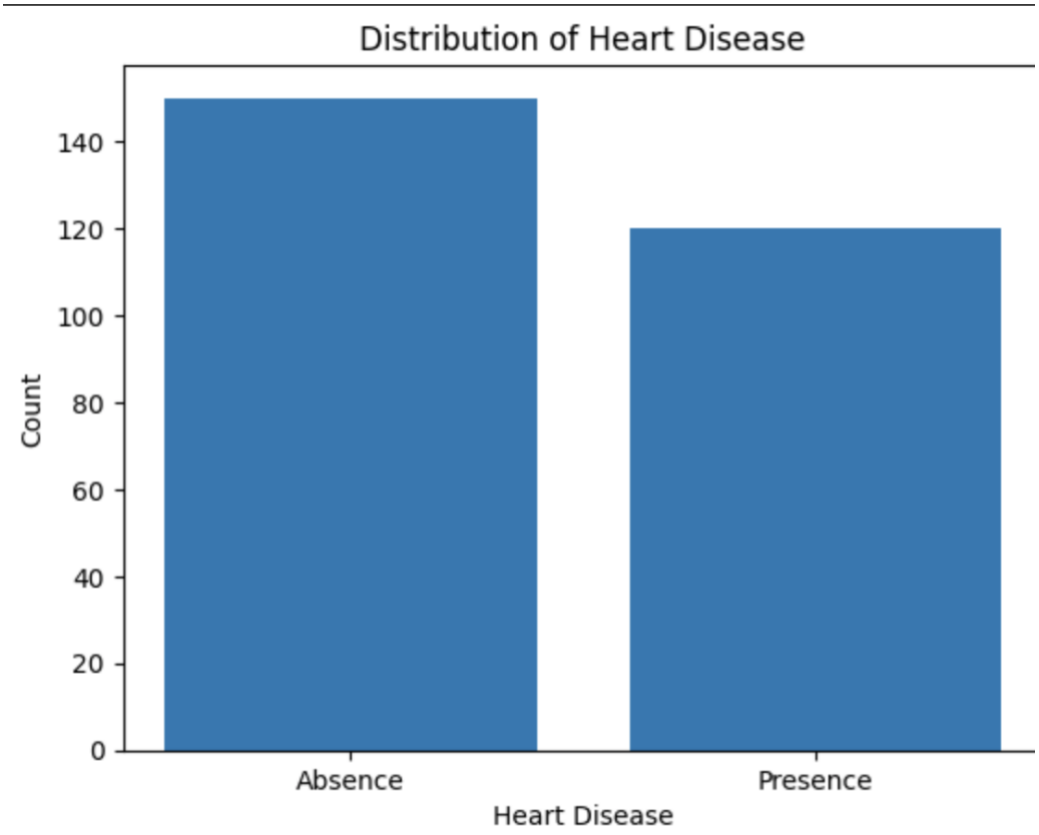

Figure 3: Overall distribution of heart disease in the test cases

Figure 3 indicates a lower occurrence of heart disease compared to its absence in the collected dataset, suggesting most individuals in the dataset are not diagnosed with heart disease.

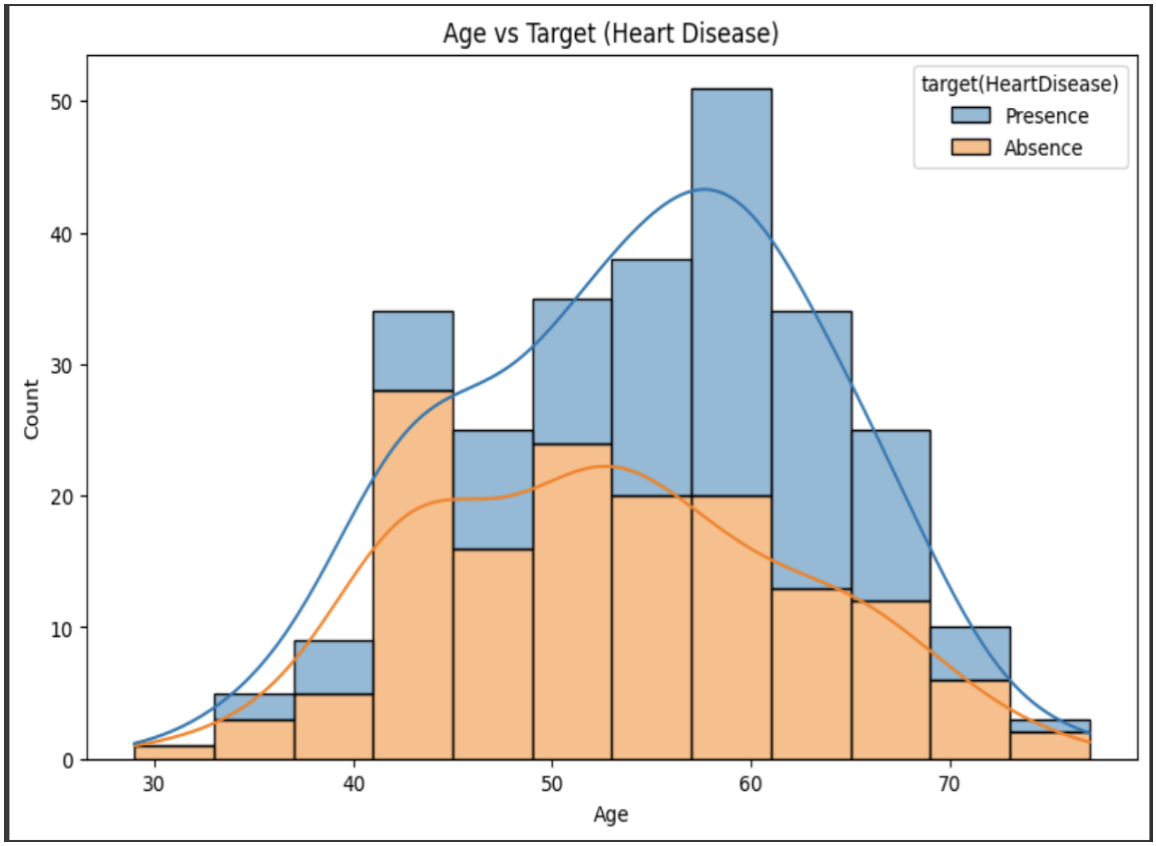

Figure 4: Statistics by age spread.

Figure 4 portrays the presence of heart disease concerning the attribute "age". It is discernible that there's an upsurge in heart disease prevalence, culminating around the age of 60.

This observation of escalated heart disease incidence at a particular age aligns with established medical knowledge according to (Rodgers *et al.*, 2019). It's universally acknowledged that the risk of heart disease escalates with advancing age. Elements such as accumulated lifestyle behaviors contribute heavily to the risks.

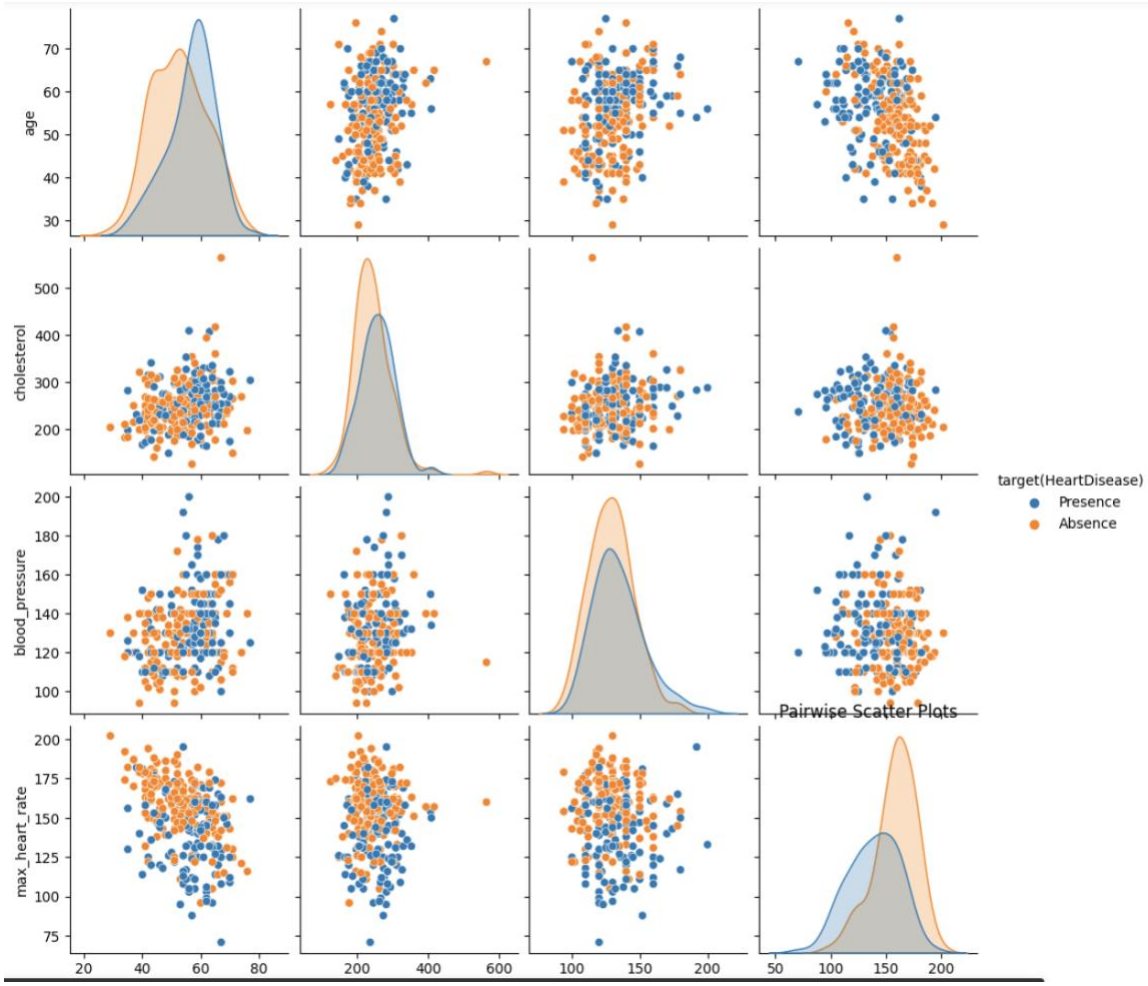

Figure 5: Graph and scattered plot of patients' cholesterol, blood pressure, max heart rate and age levels

The scattered plot presented in Figure 5, suggests that specific pairings of cholesterol, blood pressure, maximum heart rate, and age are linked with probability of heart disease. This highlights the necessity of collectively evaluating these factors when determining heart disease risk and crafting focused prevention and intervention methods.

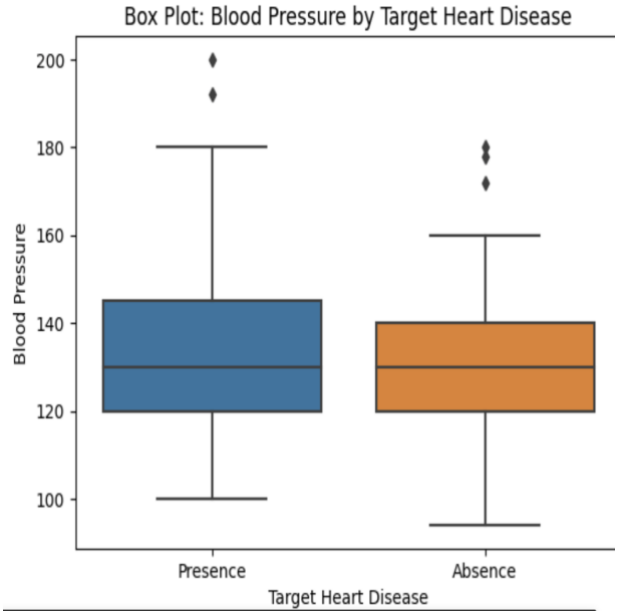

Figure 6: Distribution by blood pressure level.

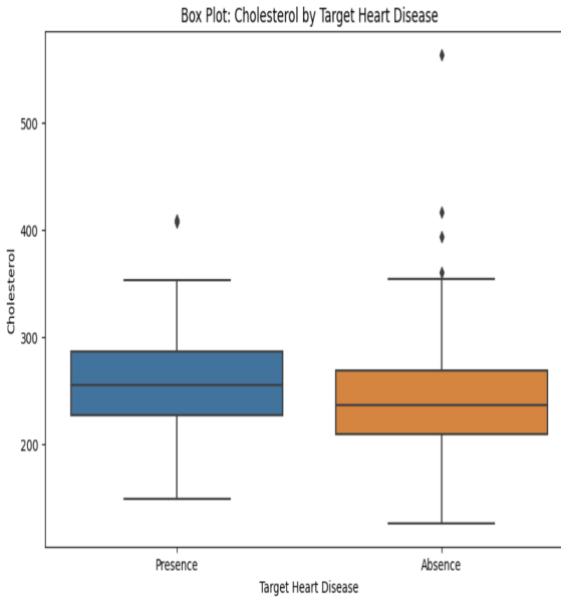

Figure 7: Distribution by cholesterol level

Upon comparison of figure 6 and 7, blood pressure seems to exert a more profound influence on heart disease presence within the dataset, as seen by a denser accumulation of cases in the high blood pressure zone. Nonetheless, it's vital to recognize that both blood pressure and cholesterol are crucial players in heart disease onset and progression.

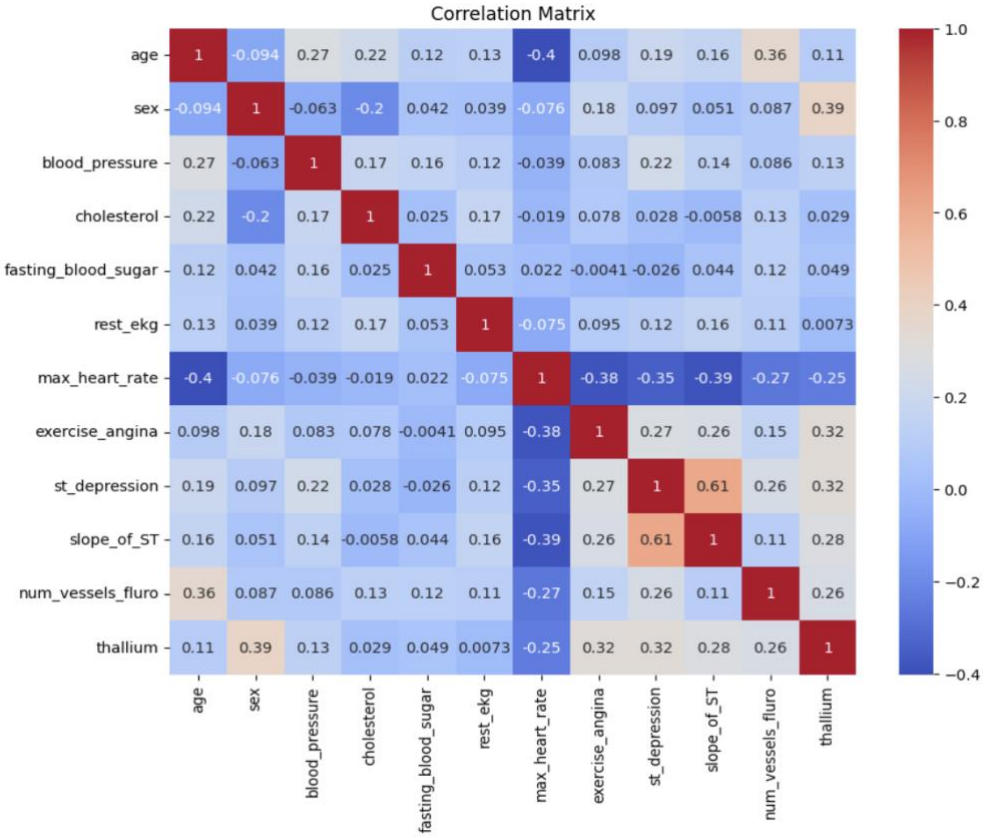

Figure 8: Correlation analysis

Figure 8 reveals that sex and age possess the highest positive correlation with AQI, signifying a substantial association between these factors and heart disease presence or absence.

A positive correlation implies that with an increase in the values of sex and age, the probability of heart disease escalates, indicating that belonging to a certain sex or an older age group, as depicted in Figure 4, might be linked with a higher risk of heart disease.

However, it's crucial to remember that correlation does not equate to causation. Despite sex and age demonstrating a strong positive correlation with heart disease, it doesn't necessarily imply they are the direct causes of heart disease. There could be other underlying factors or variable interactions contributing to the development.

VII. ALGORITHMS AND PIPELINE

- ***Decision Tree:*** A tree-like structure resembling a flowchart, where internal nodes represent attribute tests, branches signify test outcomes, and leaf nodes hold class labels (Quinlan, 1986).

- ***Random Forest:*** This ensemble learning method operates by building numerous decision trees during the training phase and determines the mode of class labels for classification tasks (Breiman, 2001).

- ***XGBoost:*** Extreme Gradient Boosting (XGBoost) is a high-performance, adaptable, and portable gradient boosting library. It constructs an ensemble of weak decision tree learners in a sequential fashion, aiming to minimize a differentiable loss function (Chen & Guestrin, 2016).

- ***Extra Tree:*** This algorithm also known as Extremely Randomized Trees, is an ensemble method that constructs multiple decision trees and averages their predictions. Its key distinction from Random Forest lies in the selection of splits: Extra Trees randomly chooses them, enhancing model diversity (Geurts, Ernst & Wehenkel, 2006).

- ***K-nearest neighbors (KNN):*** An instance-based learning that approximates the function locally and defers computation until classification. It determines class membership by considering the most frequent class among the k nearest training examples in the feature space (Cover & Hart, 1967).

Factors considered for algorithms selection:

1. Clarity in interpretation of model outputs.

2. Efficacy as gauged by performance metrics.

3. Capability to manage imbalanced data, including techniques like weighted classes or sampling.

4. Appropriate fit for the dataset, including factors like feature count, data distribution, and presence of potential non-linear relationships among features.

Step	Description
Data Pre-processing	Perform one-hot coding encoding for categorical variables, and attribute renaming for interpretability.
Train-Test Split	Split the dataset into training and testing with random state at 42.
Algorithm Selection	Decision Tree, Random Forest, XGBoost, Extra Tree and KNN.
Hyperparameter tuning and Cross Validation	Optimize the hyperparameters of selected models suing techniques like grid search or randomized search and assess models' performance to ensure generalization
Model Evaluation	Evaluate models' performance using metrics such as accuracy, precision, recall F1-score and ROC AUC
Model Interpretation	Employ interpretability tools like permutation importance, feature importance, SHAP values, Partial Dependence Plots (PDP) and LIME to understand the impact of features and assess heart disease in the dataset
Final Model Selection	Extra Tree

Table 1: Model Pipeline

Visualisation of the dataset using Decision Tree:

Utilizing tree visualization enables us to comprehend the patterns within the dataset more effectively, thereby guiding our subsequent exploratory analysis.

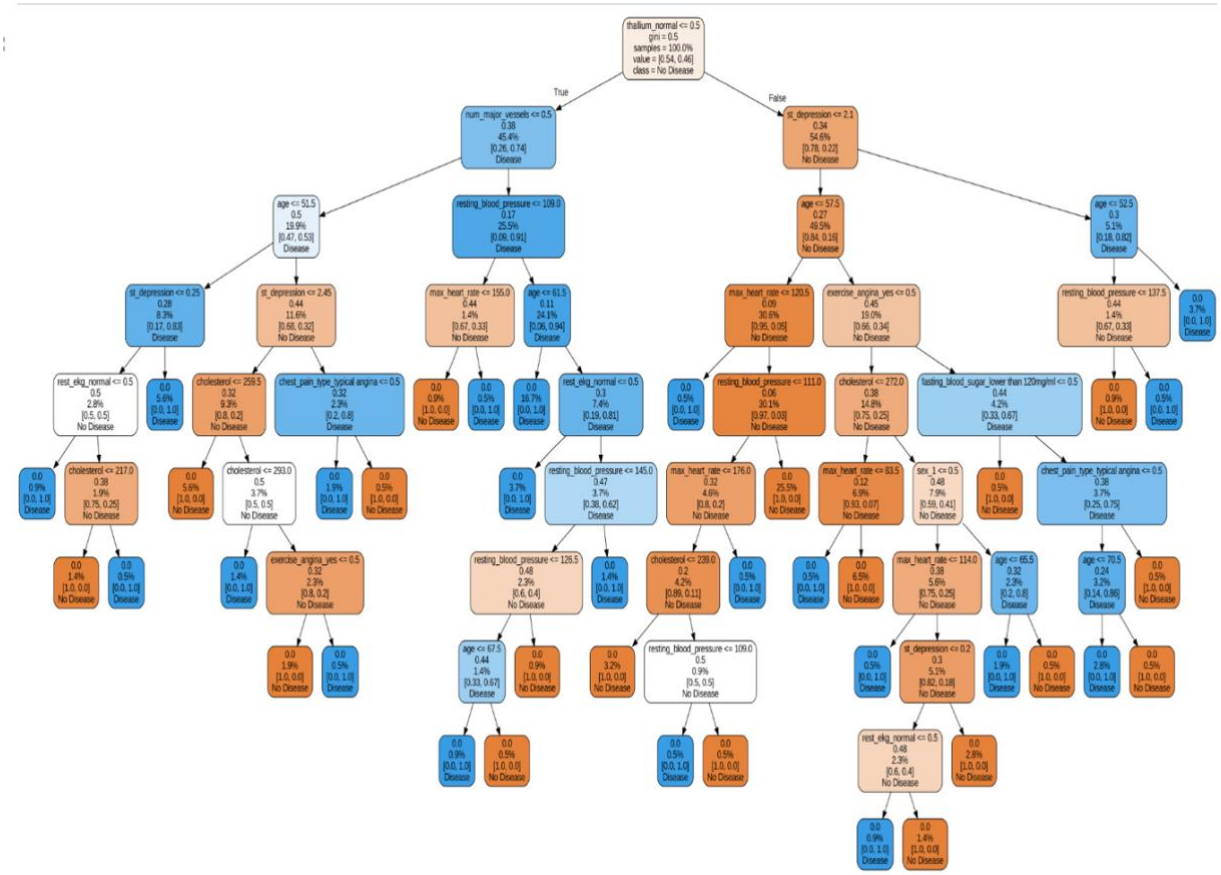

Figure 9: Tree Visual

VIII. EVALUATION METHODOLOGY

Evaluation metrics such as precision, recall, F1-score, specificity, sensitivity, and the area under the Receiver Operating Characteristic (ROC) curve were used to comprehensively analyze the predictive accuracy of the models. These metrics measure the models' ability to accurately identify both positive and negative cases, along with their overall performance. The ROC curve visualizes the balance between the true positive rate (sensitivity) and false positive rate (1 - specificity), with the ROC-AUC score summarizing the model's overall discriminatory power. A higher ROC-AUC score denotes a better performing model.

IX. MODEL ANALYSIS AND EVALUATION

To optimize the performance, hyperparameter tuning and cross-validation using a five-fold outer-loop were employed. The results revealed that Random Forest, XGBoost and Extra Tree obtained the highest best scores among the models. See *Table 2.*

Model	Best Parameters	Best Score
Decision Tree	{'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 3, 'min_samples_split': 5}	0.73
Random Forest	{'max_depth': 7, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}	0.82
XGBoost	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}	0.78
Extra Tree	{'max_depth': 5, 'min_samples_split': 4, 'n_estimators': 50}	0.82
KNN	{'metric': 'manhattan', 'n_neighbors': 5}	0.67

Table 2: Models and tuning.

In evaluating the performance of five distinct models on the dataset, we found that the ensemble algorithms - Random Forest, XGBoost, and Extra Trees - produced the highest accuracy on the training data, with scores of 96%, 98%, and 91% respectively. However, taking into account both the training and testing performances, it's apparent that Random Forest, Decision Tree, XGBoost, and KNN have a tendency to overfit, as evidenced by their high performance on training data and significant drop in performance on testing data.

In contrast, the Extra Tree classifier demonstrated less discrepancy between training and testing performance, which indicates a lower likelihood of overfitting and therefore makes it the preferable choice for this particular task. Conversely, the KNN model underperformed compared to the other models, which may be attributed to challenges such as unscaled data or the "curse of dimensionality" that often plague this type of model.

Across all models, we observed a consistent struggle in correctly identifying positive cases of heart disease, as demonstrated by lower sensitivity scores. Overfitting emerged as a common problem,

suggesting that future efforts may benefit from further feature engineering or more thorough data pre-processing, especially in the case of KNN. Please refer to *Figure 10 and Table 3.*

Model	Training Accuracy	Test Accuracy	AUC	Sensitivity (Training)	Sensitivity (Test)	Specificity (Training)	Specificity (Test)
Decision Tree	88%	74%	0.78	0.80	0.57	0.94	0.82
Random Forest	96%	85%	0.78	0.80	0.57	0.94	0.82
XGBoost	98%	81%	0.89	0.80	0.57	0.94	0.82
Extra Tree	91%	87%	0.87	0.80	0.57	0.94	0.82
KNN	77%	63%	0.74	0.80	0.57	0.94	0.82

Table 3: Models prediction report.

Decision Tree:
Training performance:

	precision	recall	f1-score	support
0	0.84	0.96	0.90	117
1	0.94	0.79	0.86	99
accuracy			0.88	216
macro avg	0.89	0.87	0.88	216
weighted avg	0.89	0.88	0.88	216

Test performance:

	precision	recall	f1-score	support
0	0.76	0.85	0.80	33
1	0.71	0.57	0.63	21
accuracy			0.74	54
macro avg	0.73	0.71	0.72	54
weighted avg	0.74	0.74	0.73	54

Random Forest:
Training performance:

	precision	recall	f1-score	support
0	0.96	0.97	0.96	117
1	0.96	0.95	0.95	99
accuracy			0.96	216
macro avg	0.96	0.96	0.96	216
weighted avg	0.96	0.96	0.96	216

Test performance:

	precision	recall	f1-score	support
0	0.86	0.91	0.88	33
1	0.84	0.76	0.80	21
accuracy			0.85	54
macro avg	0.85	0.84	0.84	54
weighted avg	0.85	0.85	0.85	54

XGBoost:
Training performance:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	117
1	0.99	0.97	0.98	99
accuracy			0.98	216
macro avg	0.98	0.98	0.98	216
weighted avg	0.98	0.98	0.98	216

Test performance:

	precision	recall	f1-score	support
0	0.81	0.91	0.86	33
1	0.82	0.67	0.74	21
accuracy			0.81	54
macro avg	0.82	0.79	0.80	54
weighted avg	0.82	0.81	0.81	54

Extra Tree:
Training performance:

	precision	recall	f1-score	support
0	0.89	0.94	0.92	117
1	0.92	0.87	0.90	99
accuracy			0.91	216
macro avg	0.91	0.90	0.91	216
weighted avg	0.91	0.91	0.91	216

Test performance:

	precision	recall	f1-score	support
0	0.86	0.94	0.90	33
1	0.89	0.76	0.82	21
accuracy			0.87	54
macro avg	0.88	0.85	0.86	54
weighted avg	0.87	0.87	0.87	54

KNN:
Training performance:

	precision	recall	f1-score	support
0	0.77	0.84	0.80	117
1	0.78	0.70	0.74	99
accuracy			0.77	216
macro avg	0.77	0.77	0.77	216
weighted avg	0.77	0.77	0.77	216

Test performance:

	precision	recall	f1-score	support
0	0.72	0.64	0.68	33
1	0.52	0.62	0.57	21
accuracy			0.63	54
macro avg	0.62	0.63	0.62	54
weighted avg	0.64	0.63	0.63	54

Figure 10: Evaluation report of models

Roc curve and scores:

In figure 11, While XGBoost and Extra Trees models show a strong ability to classify data as indicated by their high AUC-ROC scores of 89% and 88% respectively, the Random Forest, Decision Tree and KNN models lag behind. Despite this, each model may benefit from further tuning to potentially enhance these results.

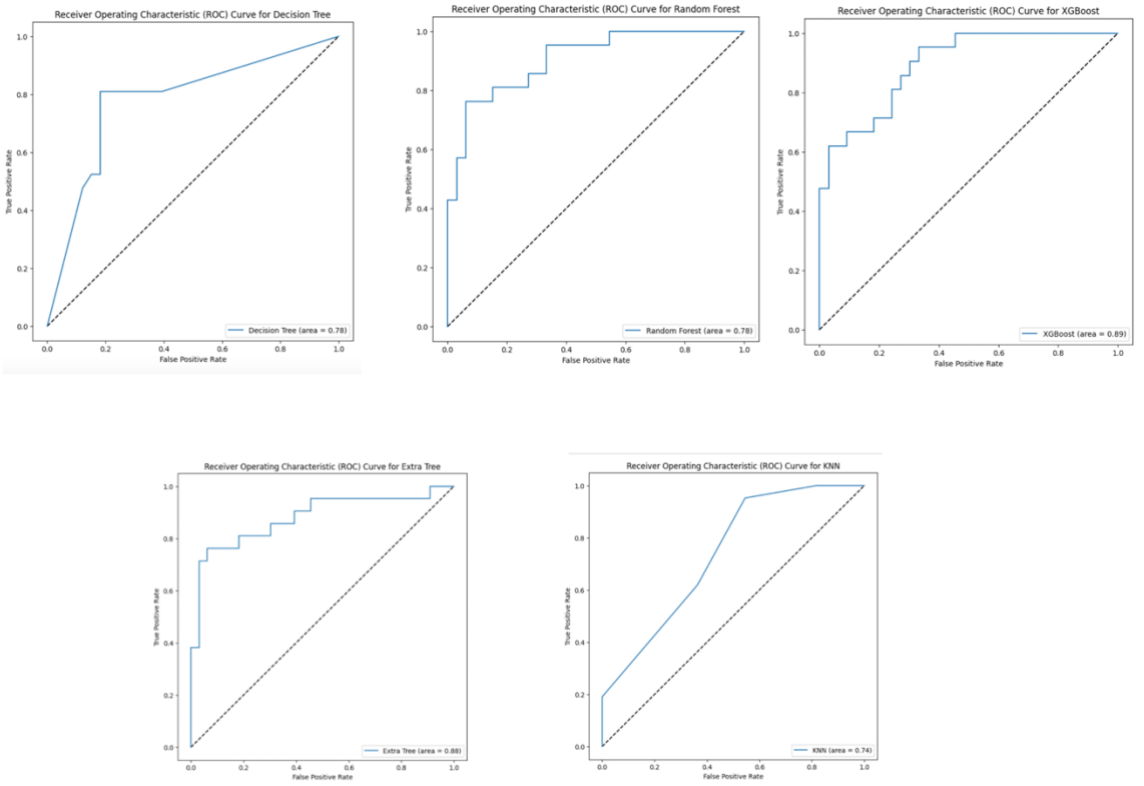

Figure 11: Models Roc Curve and Score

X. INTERPRETATION ANALYSIS

Permutation Importance using Decision Tree:

Figure 12 shows the most influential elements, as normal thallium levels and number of major vessels. these variables provide a direct snapshot of the patient's current health status at the time of check-up.

Weight	Feature
0.1000 ± 0.0444	thallium_normal
0.0667 ± 0.1185	num_major_vessels
0.0519 ± 0.0148	age
0.0370 ± 0.0524	st_depression
0.0259 ± 0.0687	chest_pain_type_atypical angina
0.0111 ± 0.0181	slope_of_ST_upsloping
0.0111 ± 0.0181	slope_of_ST_flat
0.0074 ± 0.0181	thallium_reversable defect
0.0037 ± 0.0148	max_heart_rate
0.0037 ± 0.0148	chest_pain_type_non-anginal pain
0.0000 ± 0.0331	resting_blood_pressure
0.0000 ± 0.0406	rest_ekg_normal
0 ± 0.0000	chest_pain_type_typical angina
0 ± 0.0000	fasting_blood_sugar_lower than 120mg/ml
0 ± 0.0000	rest_ekg_left ventricular hypertrophy
0 ± 0.0000	sex_1
-0.0037 ± 0.0277	exercise_angina_yes
-0.0259 ± 0.0181	cholesterol

Figure 12: Feature Importance based on permutation.

Models Feature importance:

As feature importance were observed across different models, depicted in *Figure 13 - 16*, several factors contributing to their variance was noted. These included the unique structure inherent to each model, the way each model handled interaction effects between features, the impact of regularization techniques, the scaling of data, and the randomness inherent in some models.

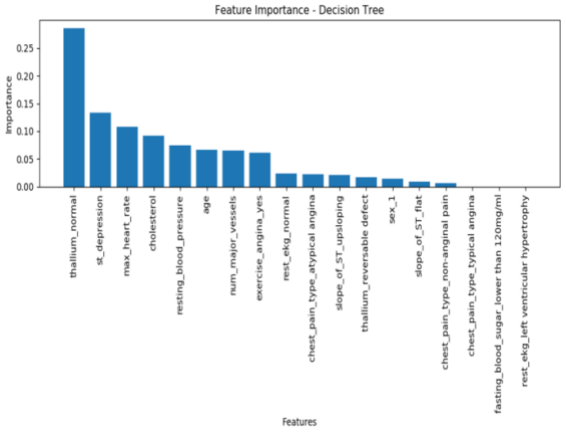

Figure 13: Feature importance DT

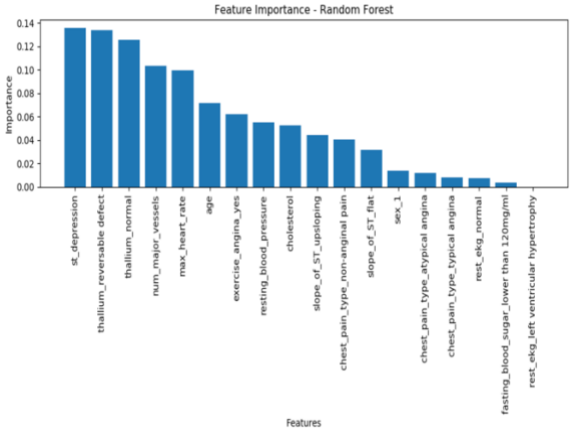

Figure 14: Feature importance RF

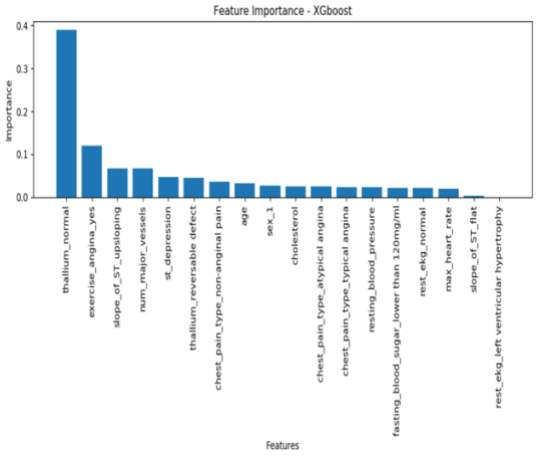

Figure 15: Feature importance XGBoost.

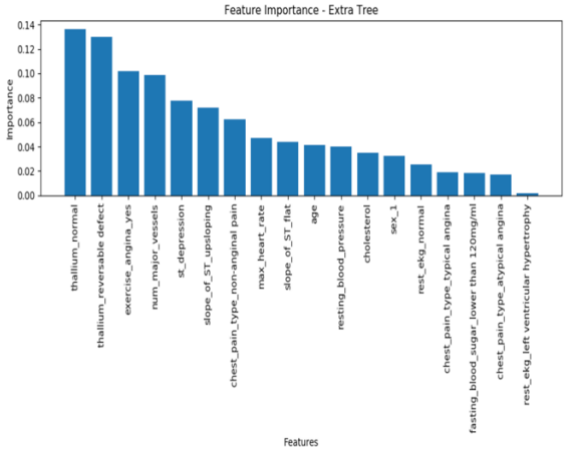

Figure 16: Feature importance ET

Shap Analysis:

In the Shap analysis, we use color coding where blue represents "lower" values and red indicates "higher" ones.

Upon examining Figures 17 and 18, a striking impact of the variable's cholesterol and number of major vessels is apparent, as evident by the strong red and blue indications. However, the color variation in the age feature seems minimal, suggesting that age might not have a significant influence in the prediction.

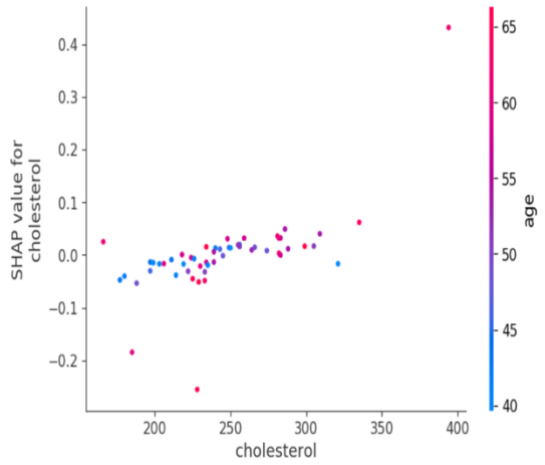

Figure 17: Cholesterol vs age

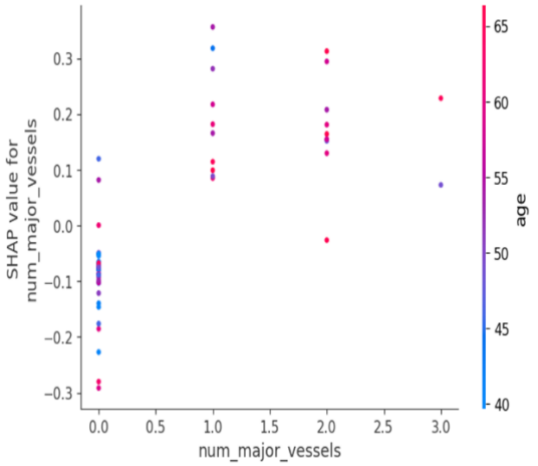

Figure 18: number of major vessels vs age

The prediction for this particular patient is at 88%, which is higher compared to the baseline of 45.7%. Several factors are contributing to this elevated risk, including thallium reversable defect, st depression, thallium normal, the presence of exercise-induced angina, maximum heart rate, unsloping and flat ST segments on the ECG, and cholesterol.

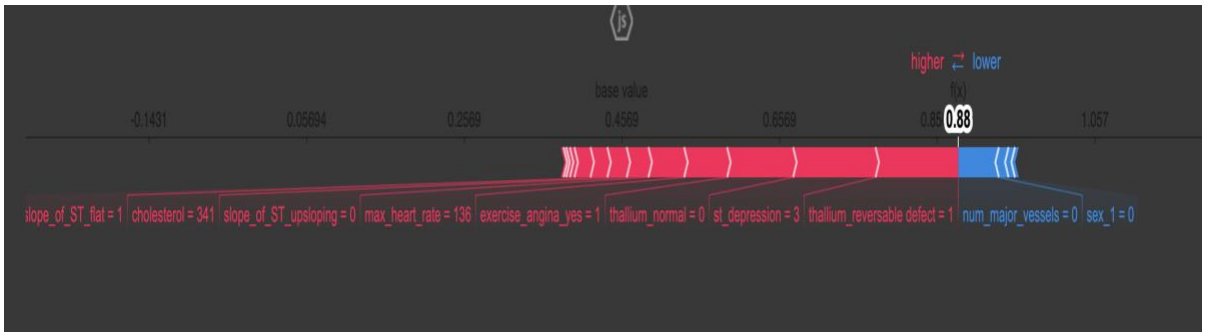

Figure 19: Random Patient analysis 1

This patient is at 40% which is lower than the baseline. Several factors are working in their favor.

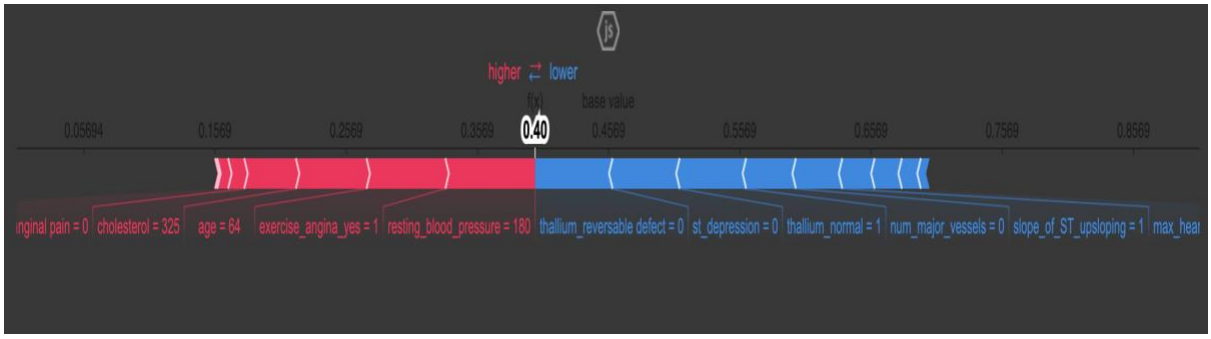

Figure 20: Random Patient analysis 2

PDP Analysis:

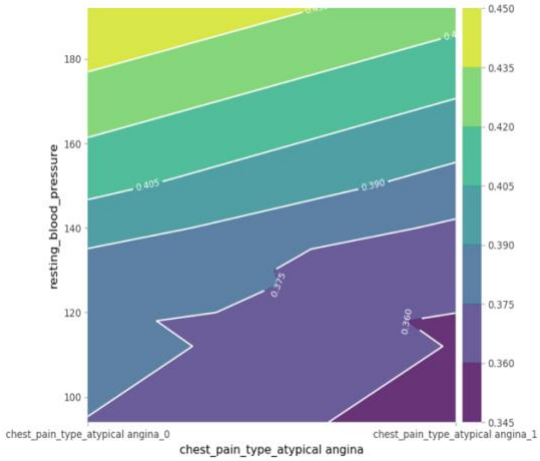

Figure 21: Interact 1.

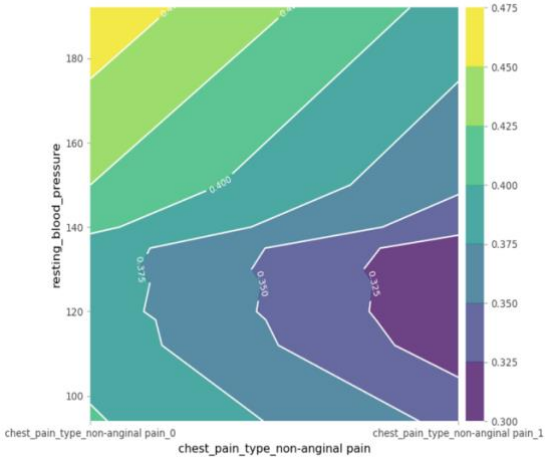

Figure 22: interact 2.

In Figure 21 and 22, it appears that the resting blood pressure for both instances falls within a desirable range.

Lime Analysis:

The lime analysis, as displayed in Figure 23 and 24, provides an in-depth view of the patients' heart disease prevalence. In both instances, the individuals are categorized as having heart disease at 55% and 69% respectively. The analysis uncovers the key factors influencing their overall health condition.

Figure 23: Analysis by age prevalence -1

Figure 24: Analysis by age prevalence -2

XI. DISCUSSION

The performance of the models might be limited by factors such as the size, diversity, and imbalance of the dataset, and the presence of potential human biases, all of which can affect the quality. To handle imbalanced data, techniques like SMOTE or ADASYN can be employed. To gain a more holistic and generalizable outcome, it would be beneficial to leverage a larger, more diverse dataset, encompassing broader aspects such as genetic indicators or lifestyle-related elements.

Most of the models demonstrated signs of overfitting, indicating that they've learned the nuances of the training data exceedingly well but struggled to generalize their learning effectively to unseen data. In such cases, it becomes crucial to utilize other model tuning methods such as Bayesian Optimization to optimize the balance between bias and variance.

It would be highly valuable to assess the models' performance on external datasets sourced from diverse populations or different healthcare settings. This approach can provide crucial insights into the generalizability and adaptability of the models across varied scenarios, thereby bolstering the evaluation of their performance on novel data and affirming their dependability in real-world applications.

Lastly, considering additional evaluation metrics such as the Matthews Correlation Coefficient or log loss could provide a more holistic appraisal of the models. Looking ahead, more advanced algorithms adept at classification tasks, such as Neural Networks, Support Vector Machines, or ensemble methods like stacking, could be explored for future studies, potentially providing improved outcomes in predicting heart disease.

REFERENCES

Breiman, L. (2001). Random Forests. *Machine Learning*, [online]. 45(1), 5-32. Available from: https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf [Accessed 20 May 2023].

British Heart Foundation (2019). *Cardiovascular heart disease.* [online]. Available from: https://www.bhf.org.uk/informationsupport/conditions/cardiovascular-heart-disease [Accessed 01 May 2023].

Chen, T. and Guestrin, C. (2016) *XGBoost: A Scalable Tree Boosting System*. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* [online] KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA, ACM, pp. 785–794. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785 [Accessed 21 May 2023].

Cover, T. and Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* [online]. IEEE Transactions on Information Theory. 13 (1), pp. 21–27. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1053964 [Accessed 20 May 2023].

Geurts, P., Ernst, D. and Wehenkel, L. (2006) Extremely randomized trees. *Machine Learning* [online]. 63 (1), pp. 3–42. Available from: http://link.springer.com/10.1007/s10994-006-6226-1 [Accessed 20 May 2023].

Kaggle (2022) *Predicting Heart Disease Using Clinical Variables*, collected by The Devastator [online]. Available from: https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var [Accessed 20 April 2023].

Mohammadnezhad, M., Mangum, T., May, W., Jeffrey Lucas, J. and Ailson, S. (2016) Common Modifiable and Non-Modifiable Risk Factors of Cardiovascular Disease (CVD) among Pacific Countries. *World Journal of Cardiovascular Surgery* [online]. 06 (11), pp. 153–170. Available from: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/wjcs.2016.611022 [Accessed 19 May 2023].

Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning* [online]. 1 (1), pp. 81–106. Available from: http://link.springer.com/10.1007/BF00116251 [Accessed 20 May 2023].

Rodgers, J.L., Jones, J., Bolleddu, S.I., Vanthenapalli, S., Rodgers, L.E., Shah, K., Karia, K. and Panguluri, S.K. (2019) Cardiovascular Risks Associated with Gender and Aging. *Journal of Cardiovascular Development and Disease* [online]. 6 (2), p. 19. Available from: https://www.mdpi.com/2308-3425/6/2/19 [Accessed 20 May 2023].

World Health Organization. (2023). Cardiovascular diseases (CVDs). [online]. Available from. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Accessed 01 May 2023].