Research on Humanities and Social Sciences
ISSN 2224-5766 (Paper) ISSN 2225-0484 (Online)
Vol.14, No.6, 2024

Predictive Modeling of Heart Failure Using Health Parameters and Machine Learning Techniques

Victor Moisés Silveira Santos[1*] Erika Carlos Medeiros[1] Patrícia Cristina Moser[1]
Jorge Cavalcanti Barbosa Fonsêca[1] Rômulo César Dias de Andrade[1]
Fernando Ferreira de Carvalho[1,2,3] Fernando Pontual de Souza Leão Junior[1]
Marco Antônio de Oliveira Domingues[2]

1. Universidade de Pernambuco, Caruaru, PE, Brazil
2. Instituto Federal de Ciência e Tecnologia de Pernambuco, Recife, PE, Brazil
3. Cesar School, Recife, PE, Brasil

*E-mail do autor correspondente: erika.medeiros@upe.br

Abstract

This study conducts a comprehensive analysis of machine learning models' potential in predicting heart failure using a dataset compiled from multiple sources across various locations. Through data preprocessing and analysis, significant correlations were identified between lifestyle characteristics and heart failure incidence. Several machine learning models, including Logistic Regression, Support Vector Machine, Random Forest, K-nearest neighbors, Extra trees, Gradient Boosting, and CatBoost, were developed, trained, and evaluated using performance metrics such as accuracy, feature importance, confusion matrix, and the ROC curve. The Random Forest model exhibited superior performance, emphasizing its robustness and effectiveness in heart failure prediction. This research underscores the significance of applying machine learning to enhance predictive accuracy and provides key insights for future applications in clinical decision support systems, suggesting directions for further research in expanding the models to encompass a broader range of cardiovascular conditions according to individual lifestyle.

Keywords: Heart Failure Prediction, Machine Learning Models, Lifestyle Characteristics, Clinical Decision Support Systems.

DOI: 10.7176/RHSS/14-6-01

Publication date: June 30th 2024

1. Introdution

Amidst the continuous evolution of the global health landscape, the ongoing rise in concerns related to heart disease and heart failure stands out as a pressing demand within the medical community (Groenewegen *et al*., 2020). Heart Failure, a condition that significantly affects the heart's ability to pump blood effectively throughout the body, poses a substantial challenge for both healthcare systems and patients' quality of life. Data reveals a growing prevalence of these conditions, underscoring the urgent need for effective prediction methods to enable early diagnoses and informed medical decisions. In 2019, heart failure alongside other cardiovascular diseases led as the primary cause of mortality worldwide, comprising approximately 85% (Who, 2019).

The application of artificial intelligence (AI) in the context of heart failure prediction emerges as a promising approach (Yu *et al*., 2018), leveraging significant technological advances and the availability of extensive medical datasets. Recent statistics indicate a sharp increase in the quantity of available clinical information, allowing advanced machine learning algorithms to thoroughly analyze these data for relevant patterns. Furthermore, with the continuous application of artificial intelligence in various health-related studies, specifically in disease prediction, proposing its use for heart failure prediction becomes a promising and relevant proposition. However, it is crucial to recognize that this advancement is not without challenges, and ethical issues such as algorithmic bias and data privacy protection require careful consideration and heightened reliability, as we are dealing with the prediction of a disease.

This scenario highlights the pressing need to understand the capabilities and limitations (Hickman *et al*., 2021) of AI tools, adapting them insightfully to clinical needs. The effective integration of these technologies will not only provide tangible benefits for heart failure prediction but also substantially contribute to advancements in

cardiac health.

This study aims to develop machine learning models for predicting heart failure based on lifestyle-related characteristics using data from Kaggle (Fedesoriano, 2021). Statistical techniques and data mining methods will be employed to identify meaningful correlations between lifestyle factors and heart failure. Machine learning models will be trained and evaluated based on performance metrics to select the most effective model for forecasting heart failure. This introduction outlines the methodology employed, with specific objectives assigned to achieve the research objective.

1. Preprocess and analyze the data collected from Kaggle, identifying meaningful correlations between lifestyle characteristics and the occurrence of Heart Failure, employing statistical techniques and data mining methods.

2. Develop and train machine learning models to predict the probability of heart failure occurrence based on the most common lifestyle characteristics identified in the previous stage.

3. Identify and compare machine learning models based on performance metrics, including accuracy, permutation feature, importance plot, confusion matrix (CM) (Zeng, 2020), and receiver operating characteristic (ROC) (Hoo *et al*., 2017) curve.

4. Select the model that exhibits the highest performance metrics as the chosen model for predicting heart failure, based on common lifestyle characteristics in each of the cases.

This research consists of four main sections. After the introduction, the "Related Work" section reviews the literature supporting this study and its connections to another research within the field. Subsequently, the "Methodology" section outlines the theoretical framework and methods utilized, encompassing initial data analysis, model development, and evaluation of machine learning models. The "Results Obtained" section focuses on data analysis, model interpretation, and evaluation using specific metrics. "Conclusions" summarizes key insights, discusses limitations, and suggests avenues for future research. Finally, the study concludes with "Bibliographic References".

Some studies have explored heart disease prediction through machine learning techniques, each providing distinct perspectives to enhance predictive accuracy and address challenges linked to cardiovascular diseases.

The study proposed by Hashi and Md Shahid Uz Zaman (2020), it was a study in the healthcare sector employed machine learning techniques to predict heart disease, comparing a traditional system to a newly proposed model utilizing Logistic Regression (LR) (Lavalley, 2008), K-Nearest Neighbor (KNN) (Dhanabal and Chandramatih, 2011), Support Vector Machine (SVM) (Mammone *et al*., 2009), Decision Tree (DT) (Liang *et al*., 2021), and Random Forest (RF) (Cutler *et al*., 2012). The proposed model, enhanced through hyperparameter tuning, outperformed the traditional system, achieving a notable peak accuracy of 91.80%. This highlighted the effectiveness of the proposed approach in accurately predicting heart disease.

The 2021 research focuses on predicting heart disease based on medical attributes using machine learning algorithms such as LR and KNN. The exploration of different algorithms to increase prediction accuracy corresponds to our objective of identifying and comparing machine learning models based on performance metrics, which is what the research carried out by Jindal *et al*. (2021) consists of. It is worth noting that the best result reported in this study achieved an accuracy of 87.5%.

In 2022, the study of Shaker *et al*. (2022) emphasized heart health's significance, utilizing machine learning and deep learning (DL) (Lecun *et al*., 2015) techniques, especially RF, for heart failure prediction. This aligns with our criteria for selecting the most effective predictive model the research employed an extensive range of machine learning algorithms, including DL and RF, demonstrating their effectiveness in predicting heart diseases. The best accuracy achieved was 78.767%, utilizing DL.

In a significant medical breakthrough in 2023, the research highlighted the importance of accurate detection and prediction of cardiovascular diseases, critical for correct treatment by cardiologists. The application of machine learning in the diagnosis of cardiovascular diseases has shown promising potential, particularly in reducing misdiagnoses. This innovative study developed a model using k-modes clustering with initialization to enhance classification accuracy. Various models including RF, DT, Multilayer Perceptron (MLP), and Extreme Gradient Boost (XGB) (Zhao *et al*., 2020) Classifier were implemented and optimized through GridSearchCV. Tested on an extensive dataset of 70.000 instances from Kaggle, the standout model was the Multilayer Perceptron with cross-validation, achieving an impressive accuracy of 87.28%. This result represents the pinnacle of the study desenvolved by Bhatt *et al*. (2023), demonstrating the efficacy of the Multilayer Perceptron in predicting cardiovascular diseases with the highest accuracy achieved.

In the upcoming section, we will delve into the goals and objectives of our research, providing a comprehensive

overview of the specific aims we aimed to achieve throughout the study.

2. Methodology

In this section, we will address the systematic approach adopted in our study. The selected methodology is the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Schröer *et al*., 2021), a robust framework that guides us through the stages of data exploration, understanding of the chosen problem, preprocessing, modeling, and evaluation. This structured method provides a clear and well-defined roadmap for our research, ensuring a comprehensive and effective analysis of the collected data on heart failure prediction based on lifestyle characteristics.

a) Understanding the Studied Issue

The primary objective of this research is to address heart failure based on the characteristics outlined in the dataset. Our main goal is to develop a reliable and accurate machine learning-based system capable of predicting and classifying the presence of heart failure using specific information. Identifying these patterns early through automated approaches can have a substantial impact on healthcare, providing timely interventions and improving outcomes.

b) Data Understanding

The dataset for heart failure prediction comprises 918 samples, obtained after excluding duplicate cases, originating from five distinct and independent cardiac datasets. Data collection took place at five specific locations: Cleveland, with 303 observations, Hungarian, with 294 observations, Switzerland, with 123 observations, Long Beach VA, with 200 observations, and Stalog, with 270 observations.

c) Data Preparation

The dataset is available on Kaggle (Bojer and Meldgaard, 2021) and comprises 11 attributes: Age, Gender, Chest Pain Type, Resting Blood Pressure, Cholesterol, Fasting Blood Sugar, Resting Electrocardiogram Results, Maximum Heart Rate Achieved, Exercise-Induced Angina, Oldpeak, ST Segment Slope, and Heart Disease. The target variable indicates the presence or absence of heart failure. In the following topics, the variables that make up the data set studied will be applied. The following variables are explained regarding their meaning:

- Age: Provides information about the patient's age in years, facilitating exploration of the relationship between age and the probability of cardiovascular diseases, considering age as a cardiovascular risk factor.
- Gender: Attribute indicates the patient's gender, with "M" for Male and "F" for Female. Including this information is relevant for investigating potential differences in heart conditions between men and women, contributing to a more comprehensive understanding of cardiovascular health.
- Chest pain type: ChestPainType categorized into four classes: Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP), and Asymptomatic (ASY). This classification provides insights into symptoms associated with heart disease, contributing to more predictions that are accurate.
- RestingBP (Resting Blood Pressure): RestingBP represents the patient's blood pressure in millimeters of mercury (mm Hg) and is crucial for assessing the risk of heart disease, as elevated blood pressure can indicate potential cardiac issues.
- Cholesterol: The (Cholesterol) attribute indicates serum cholesterol levels measured in milligrams per deciliter (mg/dl). This parameter plays an important role in cardiovascular health, and its analysis helps identify patterns and correlations with the risk of heart disease.
- FastingBS (Fasting Blood Sugar): FastingBS categorizes fasting glucose levels as one if above 120 mg/dl and zero otherwise. This information provides insights into the patient's metabolic health, relating to the risk of heart disease, especially in cases of elevated fasting sugar.
- RestingECG (Resting Electrocardiogram): Results RestingECG represent the condition of the electrocardiogram, categorized as Normal, ST (indicating abnormalities in ST-T wave), and LVH (indicating probable or definite left ventricular hypertrophy). This classification contributes to the detection of anomalies and potential cardiac conditions.
- MaxHR (Maximum Heart Rate Achieved): Indicates the maximum heart rate during activity, providing information about the patient's cardiovascular fitness.
- ExerciseAngina (Exercise-Induced Angina): Categorizes whether the patient experiences angina during exercise, with "Y" indicating yes and "N" indicating no. Information on angina during exercise helps identify patients who may experience chest pain or discomfort during physical activity, aiding in the assessment of exercise-related cardiac stress.

- ST_Slope (ST Segment Slope): Characterizes the slope of the ST segment during peak exercise, contributing to the evaluation of electrocardiographic changes associated with heart disease.
- Oldpeak: Represents the depression in the ST segment of the electrocardiogram during exercise, providing data on the extent of this depression as an indicator of potential cardiac issues.
- Heart Disease: The output class, HeartDisease, is binary, indicating the presence (1) or absence (0) of heart disease. This variable is essential for predicting and classifying cardiac conditions in the dataset.

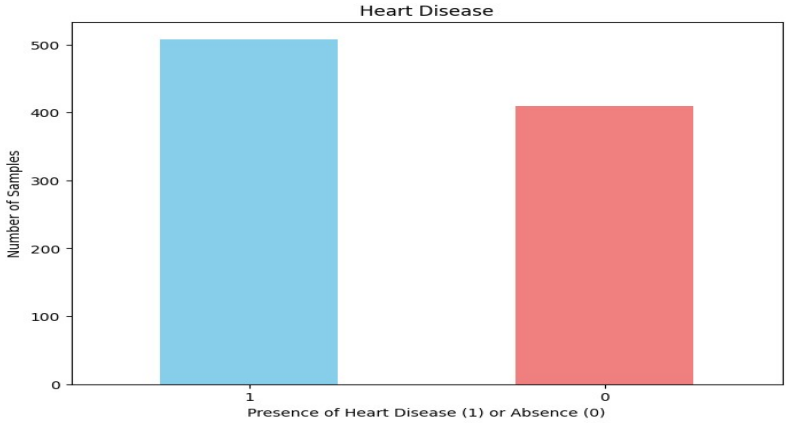

Figure 1. Class Distribution

By examining the distribution of classes in this variable, we observe 508 instances indicating the presence of heart disease and 410 instances indicating its absence, as shown in Figure 1. This division provides a precise overview of how the samples are distributed and balanced between the two categories, crucial for assessing the dataset's balance in terms of the presence or absence of heart disease. This information plays an important role in conducting subsequent analyses and developing machine learning models, ensuring an informed and effective approach to predicting heart conditions. The pivotal output variable in our study is "HeartDisease," possessing a binary nature that signifies the presence or absence of heart disease. In this representation, a value of "1" indicates the presence of heart disease, while a value of "0" denotes its absence. This variable stands as the central focus of our objectives, serving as the primary factor for predicting and classifying heart failure within our dataset.

Proper data preparation is essential to ensure the quality and reliability of results in any study. In the case of the dataset used in our research, made available on the Kaggle platform, the original authors have already performed significant pre-processing. Specifically, the balancing of the dataset was achieved by removing duplicate entries. This step is fundamental in mitigating the risk of bias in the developed models, helping to avoid the challenges commonly associated with class imbalances. Such a measure ensures a more equitable basis, as demonstrated in Figure 1, which shows an instance of 508 cases where heart disease was present and another 410 instances where heart disease was absent. Therefore, it can be observed that there is an increase in the reliability of the information generated from the study.

The dataset was divided into 80% for training and 20% for testing, following a common practice in the literature (Agarap, 2018). However, other proportions were also explored, such as 70% and 30%, and 75% and 15%. After testing and analysis, it was found that the division of 80% and 20% yielded the best results in terms of model performance metrics. This data splitting approach proved to be most effective for our specific dataset, ensuring an adequate distribution for training and testing the models.

The following balancing techniques were tested, such as SMOTE (Synthetic Minority Over-sampling Technique) (Ileberi *et al*., 2021), UnderSampling, and OverSampling techniques (Mohammed *et al*., 2020), despite the low imbalance between classes However, even with the application of these techniques, there was no improvement in the performance metrics of the models. These approaches are common for handling imbalanced datasets, where the number of examples in one class is significantly lower than in another. SMOTE creates synthetic examples of the minority class, UnderSampling reduces the number of examples in the majority class, while OverSampling increases the number of examples in the minority class. It is worth noting that these balancing techniques were applied only to the training data. Through the application of these techniques, the aim is to achieve a more balanced distribution of data, contributing to the robustness and effectiveness of prediction models.

Subsequently, the ordinal coding technique was applied to convert nominal variables into categorical ones

(Rosario *et al*., 2004), providing a more suitable representation for subsequent analyses and facilitating model analysis. In addition, we included data normalization using the MinMaxScaler (Patro and Sahu, 2015) function to standardize the scales of the variables and enhance the performance of the models during the training phase. Normalization is important because it helps prevent features with vastly different magnitudes from dominating model training, ensuring faster and more stable convergence. Additionally, we tested the StandartScaler (Aldi *et al*., 2023) function; however, due to the smaller size of the dataset, MinMaxScaler yielded better results in terms of performance metrics.

The correlation matrix analysis between the selected variables and the target variable “HeartDisease” reveals significant insights into the existing relationships. The variable "Age" shows a moderate positive correlation (0.28) with the presence of heart disease, indicating that age may slightly influence the likelihood of the condition. The feature "Sex" presents a positive correlation of (0.31), suggesting a slightly stronger influence of gender on the predisposition to heart disease.

“ChestPainType” demonstrates a notable negative correlation (-0.39), indicating an inverse relationship with the presence of heart disease. This pattern may reflect different chest pain manifestations in patients with and without cardiac conditions. The variable “MaxHR” exhibits a substantial negative correlation (-0.40), suggesting an inverse relationship between the maximum heart rate reached during activity and the presence of heart disease.

The presence of exercise-induced angina “ExerciseAngina” reveals a significant positive correlation (0.49), indicating that the occurrence of angina during exercise is associated with a higher likelihood of heart disease. Additionally, the depression in the ST segment during exercise Oldpeak shows a strong positive correlation (0.40), suggesting that greater depression may be related to the presence of heart disease.

The variable ST_Slope exhibits a considerable negative correlation (-0.56), indicating an inverse relationship between the ST segment slope during peak exercise and the presence of heart disease. This finding underscores the importance of these parameters in assessing potential cardiac conditions.

The correlation matrix, as shown in Figure 2, is an essential tool in statistical analysis, providing a systematic view of linear relationships between variables. Its interpretation, ranging from -1 to 1, allows us to understand the strength and direction of relationships between attributes and the target variable, which in our case is the presence or absence of Heart Disease. Thus, it plays an elementary role in identifying patterns and relationships among the dataset attributes, directly influencing decisions related to predictive modeling (Bun *et al*., 2017).

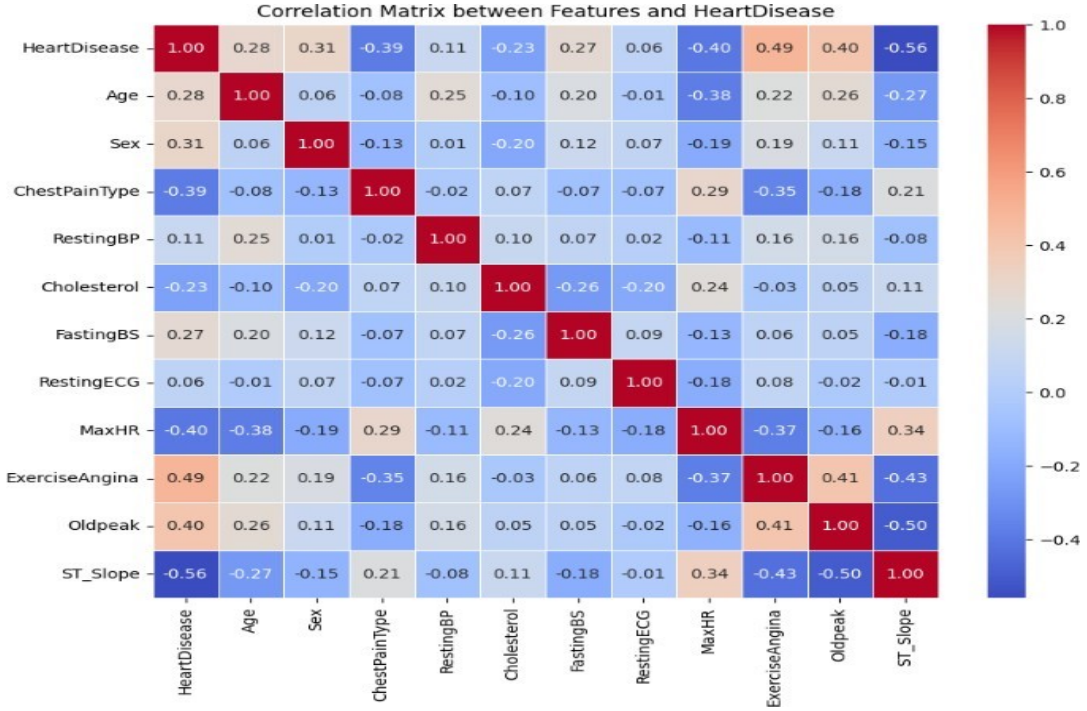

Figure 2. Class Correlation

As insights from the correlation matrix analysis, depicted in Figure 2, provide a strong foundation for further exploration and enhancement of predictive modeling techniques. By examining correlation coefficients, we can

identify variables strongly linked to the target variable, indicating their significant influence on heart failure occurrence. Leveraging this information, we can prioritize influential variables during feature selection, focusing model training efforts on attributes with the most substantial impact. Understanding interrelationships between predictor variables can guide feature engineering efforts, enabling the creation of new informative features derived from correlated attributes. Identifying variables with low correlations with the target variable allows filtering out irrelevant features, simplifying model complexity and potentially improving generalization performance. The correlation matrix serves as a valuable resource for refining predictive models, offering insights into variable importance, feature selection, and feature engineering strategies to enhance model performance and predictive accuracy.

d) Data Modeling

In this study, we utilized a variety of prediction models, including LR, SVM, ET, Cat Boost (CTB) (Saber *et al*., 2022), RF, Gradient Boost Classifier (Dorogush *et al*., 2018), and KNN. The choice of these models for the work stems from them being some of the classic models (Faouzi and Colliot, 2023) used for problem solving in the healthcare domain, as well as being models with which we had a greater understanding of their functioning. Initially, 11 models were tested, and in this study, only the top seven based on the analyzed performance metrics were considered.

In this study, we utilized a variety of prediction models, including LR, SVM, ET, Cat Boost (CTB) (Saber *et al*., 2022), RF, Gradient Boost Classifier (Dorogush *et al*., 2018), and KNN. We comprehensively evaluated these models, employing techniques such as modifying the default values set in the scikit-learn library (Hao and Ho, 2019). Furthermore, we explored the use of different sets of hyperparameters (Weerts *et al*., 2020) through the GridSearch technique. GridSearch is a method used to optimize the hyperparameters of a machine learning model, characterized as an exhaustive search method that systematically evaluates various combinations of hyperparameters, the betters hyperparameters explored in the search were represented in Table 1.

In the hyperparameter tuning process for the machine learning models, an empirical approach was employed, where various values were iteratively tested and refined. This method combined systematic experimentation with certain identified values, such as increasing or decreasing values, and more exploratory exploration of ranges near zero or specific values. The final selection of hyperparameters was primarily based on observed improvements in model accuracy during validation. While the methodology did not follow a strictly deterministic pattern, the search for continuous improvements in model accuracy guided the decisions throughout the experimentation process.

The aim is to find the configuration that yields the best performance of the model. This detailed strategy seeks to significantly improve accuracy during both the training and testing phases of the models.

e) Evaluation

In the model evaluation stage, we analyzed various criteria to determine the effectiveness in predicting heart failure. The metrics employed for evaluation included training and testing accuracy, confusion matrix, and ROC curve. These metrics provide a more detailed view of the models' performance, assessing their ability to correctly classify instances and discriminate between classes.

In the model evaluation stage, we analyzed various criteria to determine the effectiveness in predicting heart failure. The metrics employed for evaluation included training and testing accuracy, confusion matrix, and ROC curve. These metrics provide a more detailed view of the models' performance, assessing their ability to correctly classify instances and discriminate between classes.

The role of training and testing accuracy is crucial in this context. Training accuracy reflects the model's performance on the data used for training, while testing accuracy estimates of how well the model generalizes to data not used during training. both metrics are necessary to evaluate the model's ability to effectively learn patterns in the data and apply this knowledge to new observations. Therefore, when interpreting the results, it is essential to consider not only training accuracy but also testing accuracy, ensuring a comprehensive and more reliable assessment of the model's performance and avoiding potential issues such as overfitting (Ying, 2019).

A detailed analysis of the confusion matrix provided valuable insights into the models' performance in different scenarios, displaying their ability to predict true positives and true negatives, a key metric in model selection.

Additionally, when evaluating the ROC curve, considering sensitivity, the learning curve, and specificity, the curve demonstrated a favorable balance between true and false positives, highlighting its efficiency in discriminating between classes.

In the next section, we will present the results regarding the models and their evaluation criteria, which are of

fundamental importance to ensure reliability related to the model.

Table 1. Models with the best hyperparameters

Model	Hyperparameter	Tested Values	Best Value
Logistic Regression	Regularization value	0.01,0.1,1,10,100	100
	Penalty	L1,L2	L2
	Cross validation	5,6,7,8,9,10	5
Support Vector Machine	Regularization value	0.1, 1, 10	10
Extra Trees	N_estimators	100, 200, 300	100
	Max_deph	None, 100, 200, 258,	None
	Min_samples_splits	2, 5, 10	5
	Min_samples_leaf	1, 2, 4	2
	Max_features	Sqrt, Log2, 0.2	Sqrt
Random Forest	N_estimators	100, 200, 300	100
	Max_deph	None, 100, 200, 258,	None
	Min_samples_splits	2, 5, 10	5
	Min_samples_leaf	1, 2, 4	2
	Max_features	Sqrt, Log2, 0.2	Sqrt
Cat Boost	Learning_rate	0.01, 0.1, 0.2	0.2
	Deph	4, 6, 8	6
	N_estimators	100, 200, 300	100
	Sub_sample	0.8, 0.9, 1.0	1
	Cross validation	5,6,7,8,9,10	5
Gradient Boost	Learning_rate	0.01, 0.1, 0.2	0.1
	Max_deph	6, 8, 10	6
	N_estimators	50, 100, 200	50, 100, 200
K-Nearest Neighbors	N_neighbors	2, 5, 7, 9	9
	weights	Uniform, Distance	Distance
	metric	Euclidean, Manhattan	Manhattan

3. Discussion and Results

In this section, we proceed with an analysis of the performance of machine learning models, starting with the evaluation of relevant metrics such as accuracy on test and training data, as well as confusion matrices and the learning rate of the models. This approach allowed us to gain insights into the efficiency with which the models adapt and generalize from the provided data. The performance of these models was further improved through hyperparameter optimization, which fine-tuned their settings to achieve optimal results. We continued with a review of the effectiveness of a variety of models, including ensemble models like RF and GB, which achieved excellent accuracy in model testing, and tree-based models like DT and ET, focusing on their application in predicting heart failure. We observed that models such as KNN and GB excelled, achieving notable training accuracies, and the analysis was further enriched by considering models like CB and SVM, providing a broad perspective on the available options. This detailed analysis underlines the importance of a careful selection of models, aiming to enhance precision in predicting heart failure, thereby extending the scope of our study.

An analysis of Table 2 reveals the training and test accuracies of each tested model. None of the models exhibited higher training accuracy than test accuracy, effectively avoiding overfitting. This demonstrates that the models can generalize their predictive capabilities to unseen data, a critical factor in their overall performance.

A deeper analysis of the Table 2 highlights the models' performance, particularly focusing on the test accuracies. CB and RF models stand out with the highest test accuracies of 91.848% and 90.217%, respectively, indicating their superior ability to generalize and make accurate predictions on unseen data. On the lower end, LR and GB show the lowest test accuracies of 84.239% and 87.500%, suggesting these models, while still effective, might benefit from further tuning or may inherently be less suited to the dataset's specific challenges compared to their counterparts

Table 2. Model Performance Comparison

MODEL	TRAIN ACCURACY	TEST ACCURACY
Logistic Regression	86.104%	84.239%
Support Vector Machine	90.736%	88.587%
Extra Trees	94.112%	88.043%
Random Forest	96.458%	90.217%
Cat Boost	98.043%	91.848%
Gradient Boost	100.000%	87.500%
K-Nearest Neighbors	100.000%	88.043%

In a confusion matrix, "true positives" (TP) represent cases that are correctly identified as positive, while "true negatives" (TN) denote cases that are accurately identified as negative. Conversely, "false positives" (FP) occur when negative cases are incorrectly labeled as positive, and "false negatives" (FN) arise when positive cases are mistakenly labeled as negative. TP and TN reflect the model's accurate predictions, serving as indicators of its effectiveness in identifying each class correctly. On the other hand, FP and FN represent errors in prediction, highlighting instances where the model fails to distinguish accurately between classes. These metrics are crucial for evaluating a model's performance, as they provide insight into not only the model's ability to identify true cases of each class but also its propensity to misclassify them. A model's performance is considered more robust and reliable if it achieves a high number of TP and TN with minimal FP and FN, indicating a strong capability in classifying cases correctly across both positive and negative classes. This detailed breakdown helps in understanding the model's strengths and weaknesses, guiding further improvements and adjustments to enhance its predictive accuracy and reliability in distinguishing between different classes.

In Figure 3, the confusion matrix of the LR model in binary classification contexts offers an insightful analysis of its performance. The matrix shows 68 instances accurately classified as TN, indicating the model's strong capability in correctly identifying negative instances. Despite the presence of 9 FN and 20 FP, which highlight errors in misclassifying positive instances as negative and failing to recognize some negative instances respectively, these numbers do not drastically affect the model's overall effectiveness. The 87 TP further demonstrate the model's competence in accurately identifying positive instances. These outcomes emphasize the LR model's ability to effectively differentiate between classes, underlining its resilience and precision in categorizing various groups.

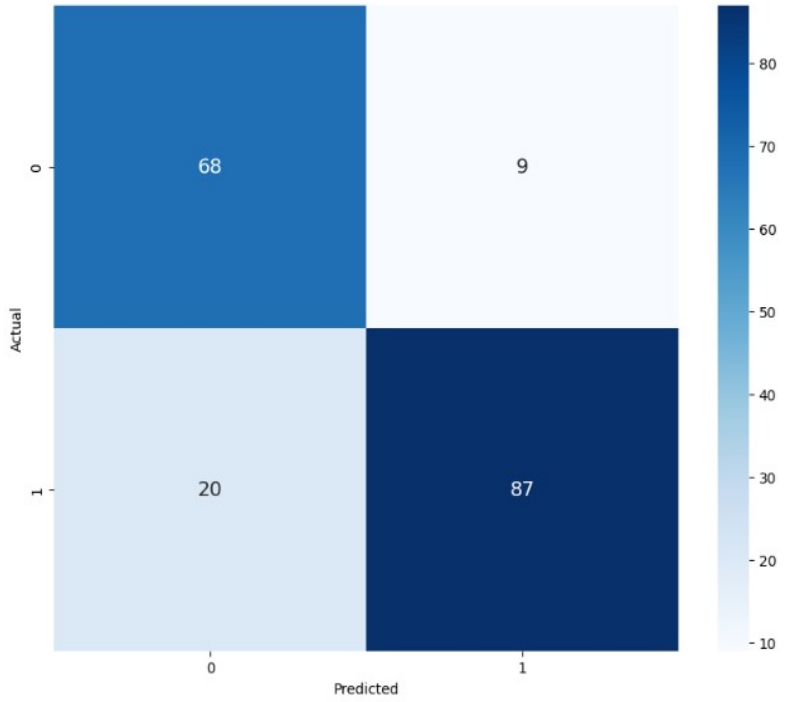

Figure 3. Confusion Matrix LR

The Figure 4 depicts the permutation importance graph for the LR model, indicating that the “ST_slope” feature, which corresponds to the slope of the ST segment on an electrocardiogram during exercise, is the most influential in the model's accuracy, standing out as the most important characteristic. In contrast, “Age, representing the patients' age, shows the least impact, being the least significant feature. This implies that, within the context of your model, variations in the ST slope are crucial for outcome prediction, while variations in patient age are comparatively less significant.

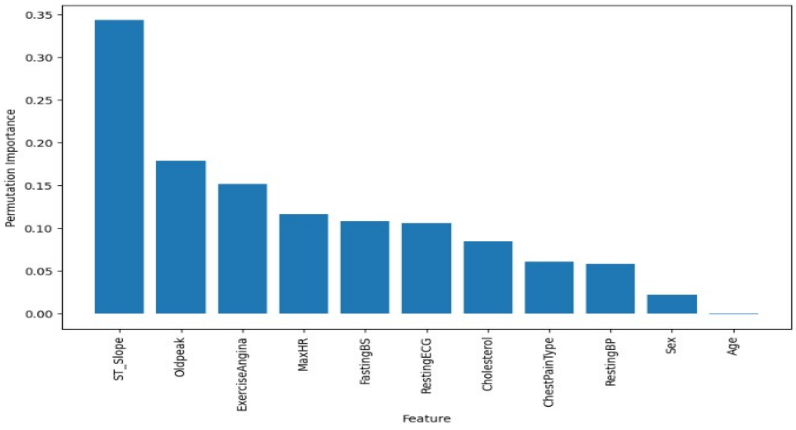

Figure 4. Permutation Importance LR

In Figure 5, the confusion matrix of the SVM model in binary classification contexts presents a comprehensive evaluation of its performance. With 70 instances accurately classified as TN, the model exhibits exceptional effectiveness in correctly identifying negative instances. Despite encountering 7 FN and 14 FP, indicating errors in misclassifying positive instances as negative and missing some negative instances respectively, these discrepancies do not markedly detract from the model's overall performance. The presence of 93 TP further affirms the model's proficiency in accurately identifying positive instances. These findings underscore the SVM model's adeptness at effectively differentiating between classes, highlighting its durability and precision in the prediction of diverse categories.

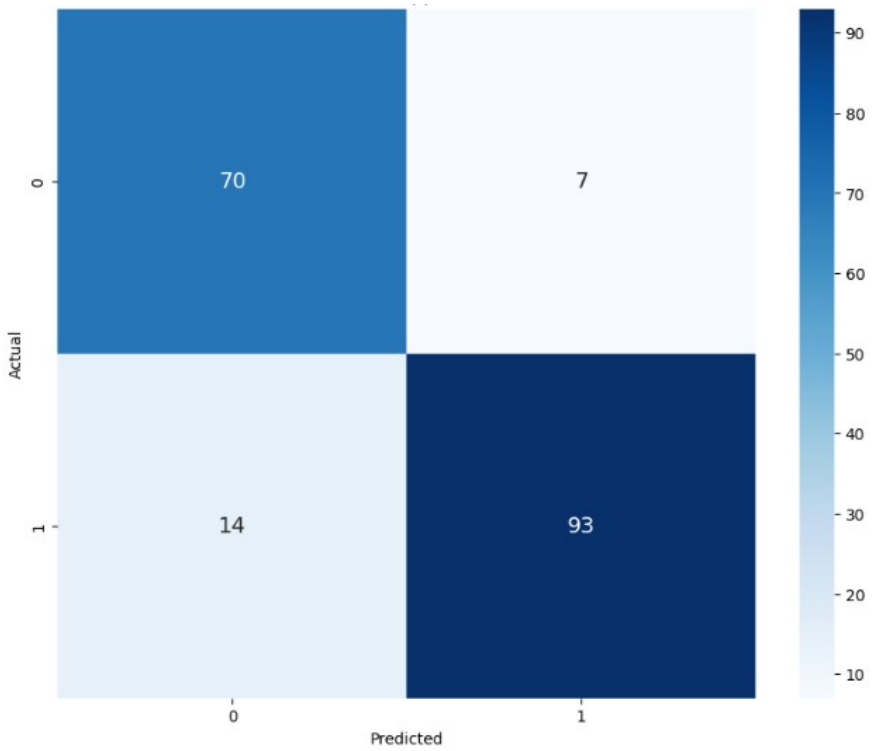

Figure 5. Confusion Matrix SVM

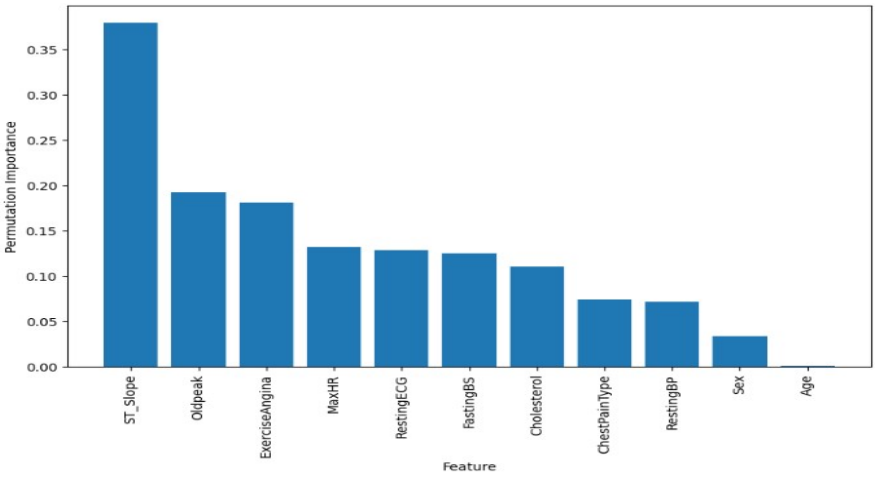

Figure 6. Permutation Importance SVM

The Figure 6 presents the permutation importance graph for the Support Vector Machine (SVM) model. This graph indicates that the ST_slope feature, denoting the slope of the ST segment on an electrocardiogram during exercise, holds the highest importance in the model's predictive accuracy. It stands out as the critical characteristic for the model's performance. On the other end of the spectrum, Age, which represents the patients' age, appears to have the least influence on the model's accuracy, making it the least significant feature. The implication here is that in the SVM model's predictions, the variations in the ST segment's slope during exercise are essential for accurate outcome prediction, while patient age has a minimal effect on the model's performance.

In Figure 7, the confusion matrix of the ET model in binary classification contexts provides a detailed view of its performance. With 68 instances correctly classified as TN, the model demonstrates high effectiveness in accurately identifying negative instances. Although there are 9 FN and 13 FP, representing errors in classifying negative instances as negative and vice versa, these figures do not significantly compromise the overall efficiency of the model. This is further evidenced by the 94 TP, showcasing the model's ability to correctly

identify positive instances. Such results highlight the potential of the ET model to distinguish between classes effectively, underscoring its robustness and accuracy in predicting different categories.

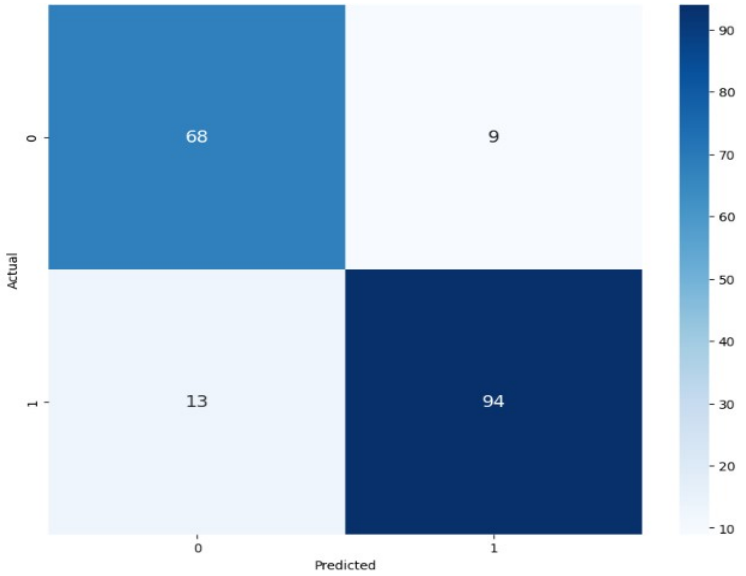

Figure 7. Confusion Matrix ET

Figure 8 corresponds to the permutation importance graph provided for the ET model. It shows that the "ST_slope" feature, representing the slope of the ST segment on an electrocardiogram during exercise, is the most influential in the model's accuracy. This feature emerges as the most critical factor for the model's predictions. In contrast, "Age", which denotes the patient's age, has the least influence on the model's accuracy, marking it as the least significant feature. This suggests that within the context of the ET model, changes in the ST slope are vital for accurate outcome prediction, while the age of the patients does not significantly affect the model's performance.

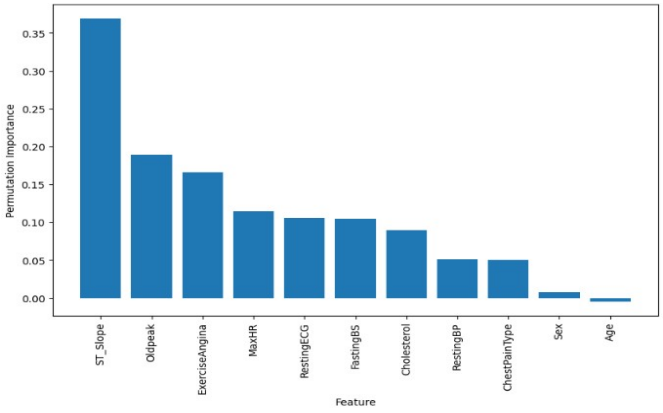

Figure 8. Permutation Importance ET

In Figure 9, the confusion matrix of the RF model in binary classification scenarios offers a nuanced understanding of its performance. With 68 instances accurately identified as TN, the model exhibits substantial efficacy in correctly identifying positive instances. Despite the presence of 9 FN and 11 FP, which indicate errors in mistakenly classifying positives instances as negative and overlooking some negative instances, respectively, these numbers do not drastically undermine the model's overall performance. The 96 TP further highlight the model's capability in accurately recognizing positive instances. These outcomes emphasize the RF model's competence in effectively differentiating between classes, illustrating its strength and precision in the prediction of diverse categories.

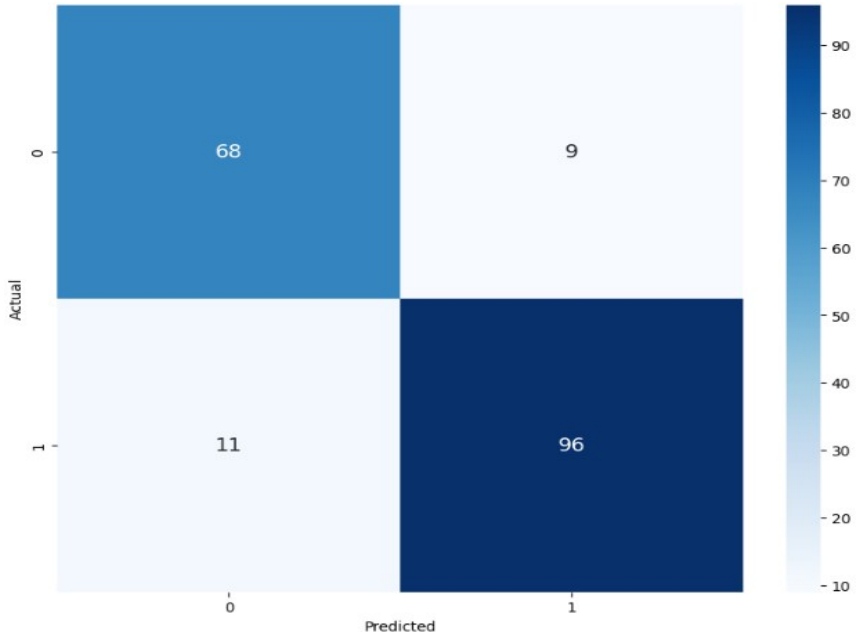

Figure 9. Confusion Matrix RF

Figure 10 presents the permutation importance graph for the RF model, illustrating that the "ST_slope" feature, which corresponds to the slope of the ST segment on an electrocardiogram during exercise, holds the highest importance in the model's accuracy. It is identified as the most critical characteristic for the model's predictions. On the other end of the scale, "Sex", which denotes the patients' age, has the least impact on the model's accuracy, rendering it the least significant feature. This indicates that in the RF model, variations in the ST slope are essential for accurate outcome prediction, whereas variations in patient age have a relatively minor influence.

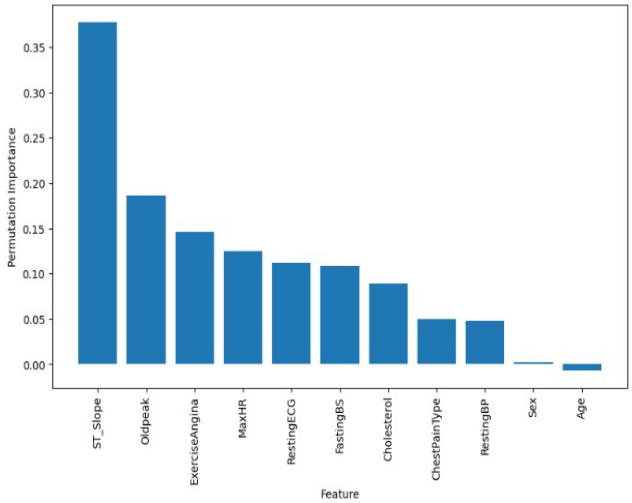

Figure 10. Permutation Importance RF

In Figure 11, the confusion matrix of the CB model in binary classification scenarios shows its performance. The matrix indicates that 65 instances were correctly classified as true negatives, demonstrating the model's effectiveness in identifying negative instances. However, there were 12 FN and 11 FP, indicating errors in classifying some positive instances as negative and some negative instances as positive, respectively. Despite these errors, the overall performance of the model remains solid, as evidenced by the 96 TP, which highlight the model's strong capability in correctly identifying positive instances. These results highlight the CB model's ability to effectively differentiate between classes, demonstrating its precision in predicting diverse categories.

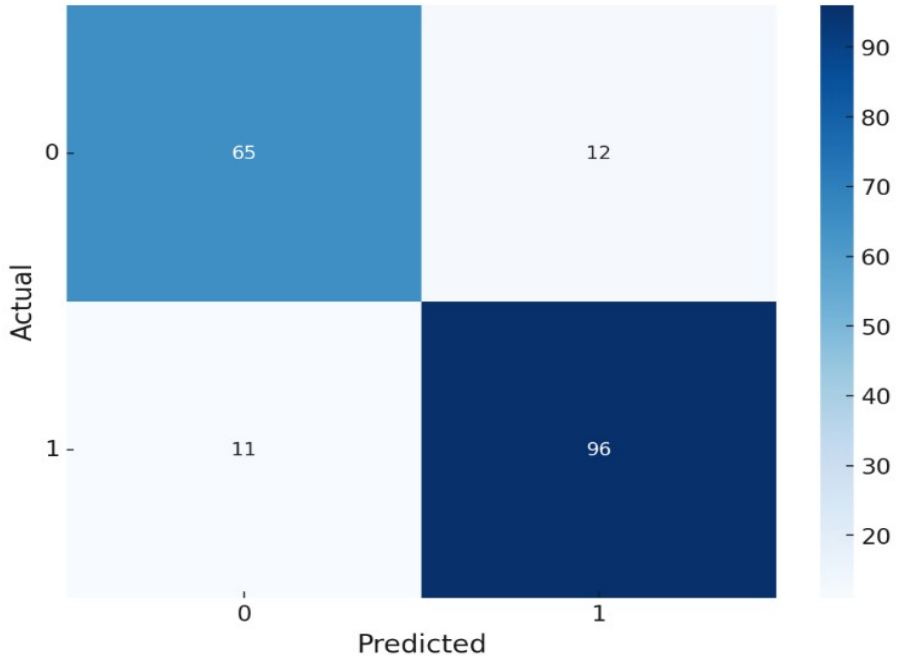

Figure 11. Confusion Matrix CB

Based on the permutation importance the Figure 12 graph for the CB model that you provided, it can be deduced that “ST_slope” emerges as the feature with the highest impact on the model's accuracy. This implies that the slope of the ST segment on an electrocardiogram during exercise is the most significant predictor within this model. Conversely, “Age” appears to have the least impact on the model's predictions, suggesting that patient age is the least important feature in the context of the CB model for outcome prediction. This distinction highlights the value of the ST segment's behavior during exercise as a critical factor in the CB model, overshadowing the relevance of age in this analysis.

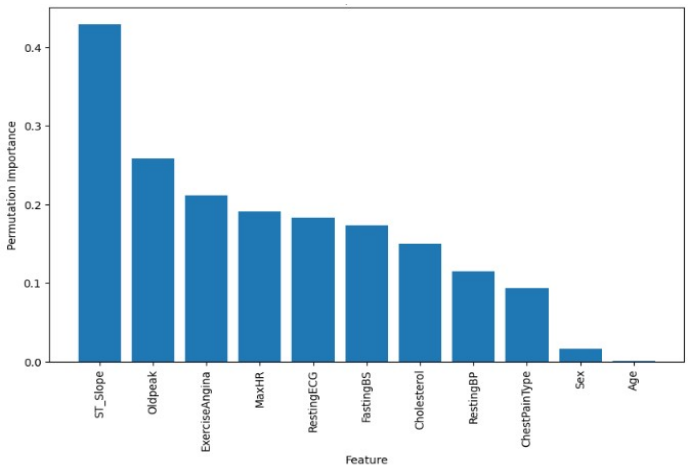

Figure 12. Permutation Importance CB

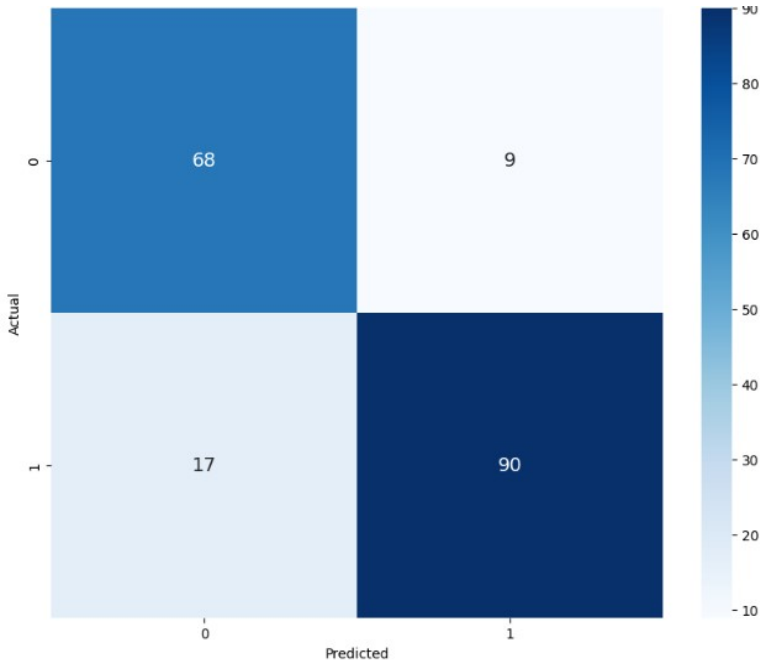

Figure 13. Confusion Matrix GB

In Figure 13, the confusion matrix of the GB model in binary classification scenarios is analyzed to assess its performance. The matrix reveals that 68 instances were correctly classified as TN, showcasing the model's effectiveness in identifying negative instances. However, there were 9 FN and 17 FP, indicating instances where positive cases were incorrectly marked as negative and negative cases as positive, respectively. Despite these misclassifications, the model's overall performance remains commendable, highlighted by the 90 TP which underscore the model's ability to correctly recognize positive instances. These results demonstrate the GB model's proficiency in distinguishing between classes accurately, showing its capability and precision in handling a variety of categories.

The permutation importance graph for the GB model, as illustrated in Figure 14, highlights that the “ST_slope” attribute is the most critical in determining the model's precision. This attribute corresponds to the slope of the ST segment observed in an electrocardiogram while a patient exercises and is identified as the predominant factor influencing the model's predictive power. In stark contrast, the “Sex” attribute, which denotes the age of the patients, is identified with minimal impact, ranking as the feature with the least significance. This contrast suggests that for the GB model's predictive accuracy, the variation in the ST segment's slope is imperative, whereas the age of the patients plays a minimal role in the predictions.

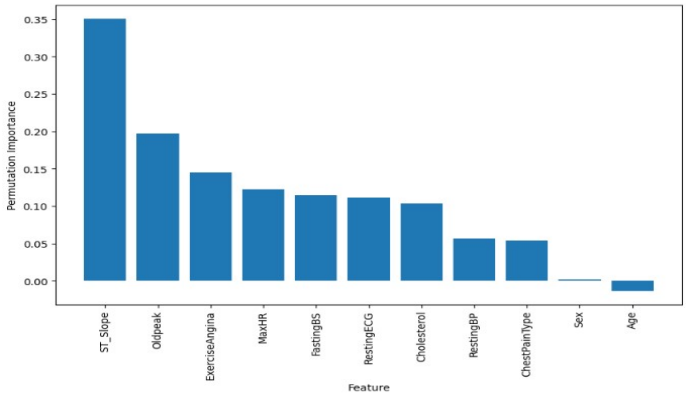

Figure 14. Permutation Importance GB

In Figure 15, the confusion matrix of the KNN model in binary classification scenarios presents a clear view of its effectiveness. The matrix shows that 68 instances were correctly identified as TN, indicating the model's

efficiency in recognizing negative instances accurately. However, it also includes 9 FN and 13 FP, reflecting mistakes in classifying some positive instances as negative and some negative instances as positive, respectively. Despite these classification errors, the model's overall performance is strong, as evidenced by the 94 TP, which demonstrate the model's skill in correctly identifying positive instances. These results validate the KNN model's ability to accurately differentiate between classes, showcasing its robustness and precision in categorizing diverse groups.

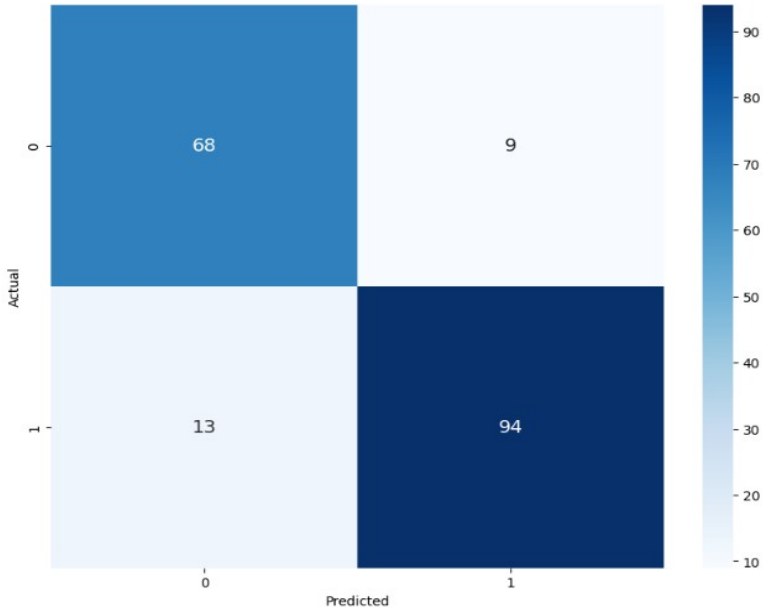

Figure 15. Confusion Matrix KNN

The permutation importance graph for the KNN model, shown in Figure 16, demonstrates that the "ST_slope"feature is paramount in affecting the model's accuracy. This feature, indicative of the ST segment's slope on an electrocardiogram during physical exertion, is the most significant determinant for the model's predictions. In marked contrast, the 'Age' feature, indicative of the patients' age, exerts the smallest influence on the model's accuracy, establishing it as the feature with the least impact. These distinctions suggest that for the KNN model's ability to predict outcomes accurately, the fluctuation in the ST segment's slope is of considerable importance, whereas the patients' age has a comparatively negligible effect.

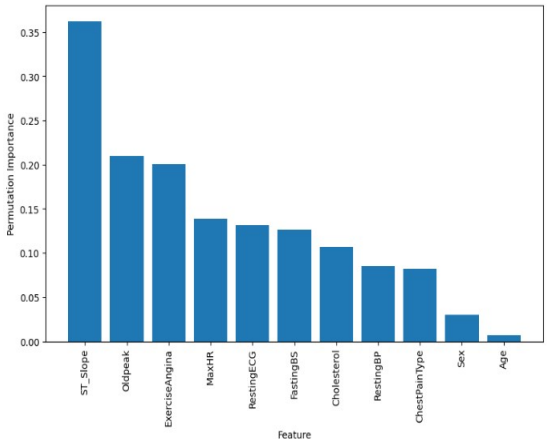

Figure 16. Permutation Importance KNN

The ROC curve represented in Figure 17 is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the True Positive Rate TPR against the FPR at various threshold settings. The Area under the Curve (AUC) (Marzban, 2004) provides a single measure of overall performance of the classifier; the closer the AUC is to 1, the better the model is at distinguishing between the two classes.

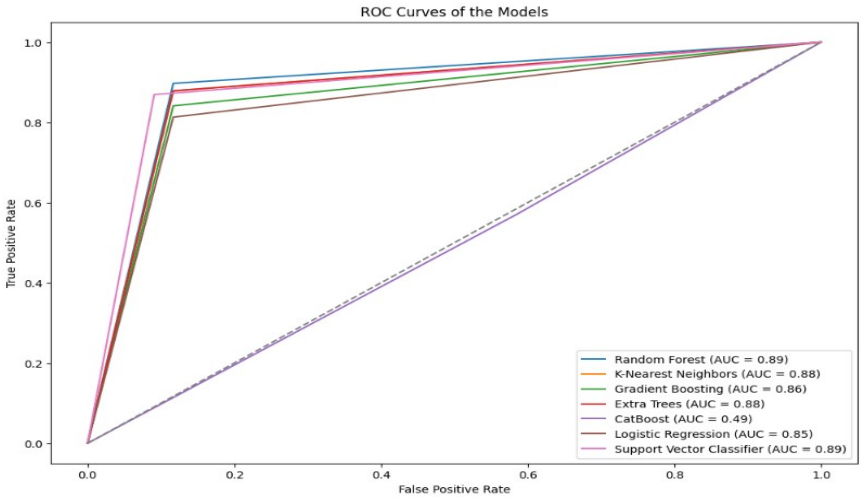

Figure 17. Curve Roc of Models

The ROC curve for the RF classifier shows an AUC of 0.89, which suggests that the model has a high capability in discriminating between the positive and negative classes. The RF curve is closer to the top-left corner of the plot, indicating a lower rate of false positives for any given true positive rate.

For the KNN model, the ROC curve displays an AUC of 0.88. This is slightly less than the RF model but still denotes a high discriminatory power. The KNN curve follows closely to the RF, suggesting similar performance characteristics, especially in the middle range of the FPR.

The GB model's ROC curve, with an AUC of 0.86, indicates that it has a good classification performance, albeit slightly lower than RF and KNN. The GB curve shows that the model performs well, particularly as the FPR begins to increase, which is indicative of its ability to maintain a relatively high TPR even as the number of false positives increases.

The ET classifier exhibits an ROC curve with an AUC of 0.88, on par with the KNN model. This indicates a high level of classification accuracy similar to that of KNN. The ET curve, much like the KNN has, maintains a close proximity to the top-left corner, which is desirable in an ROC curve.

The CB classifier stands out with an AUC of 0.49, which is significantly lower than the other models and suggests that the model is performing no better than random chance at distinguishing between the positive and negative classes. This is a critical point of concern, as it might indicate issues with model training, feature selection, or data quality.

The Logistic LR model has an ROC curve with an AUC of 0.85, which suggests a good predictive performance with a strong ability to classify the positive and negative instances correctly. The LR curve, while not as high as RF or KNN, still indicates a model that performs well at most threshold levels.

Lastly, the SVC ROC curve, with an AUC of 0.89, is indicative of excellent performance, equivalent to the RF model. The SVC curve rises sharply and maintains a high TPR across most levels of FPR, which shows that the model has a strong discriminative power.

Each of these ROC curves provides valuable insights into the performance of the models, allowing for a comparison not just of the overall accuracy via the AUC but also of the behavior of the models at different thresholds, which can be crucial for decision-making processes in a clinical setting.

The present study employed a comprehensive statistical analysis to assess the distributions of seven distinct populations in relation to 30 paired samples each. The results indicated a significant rejection of the null hypothesis (H0) that the CTB, RF, GB, ET, KNN, and LR populations follow a normal distribution ($p=0.000$), suggesting that not all populations are normal.

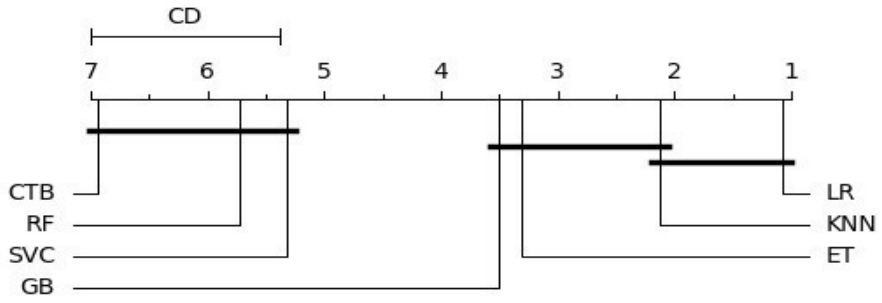

Figure 18. Statistical differences between the tested models.

Given the heterogeneity of the populations and the presence of non-normal distributions, the non-parametric Friedman test was chosen as an omnibus test to identify possible significant differences in the medians of the populations. Subsequently, the Nemenyi test was applied as a post-hoc test to determine which differences were statistically relevant. The results revealed a statistically significant difference between the medians of the populations (p=0.000).

We detailed the medians (MD), median absolute deviation (MAD), and mean rank (MR) for each population, providing specific information about CTB (MD=1.000+-0.050, MAD=0.000, MR=6.935), RF (MD=2.000+-0.500, MAD=0.500, MR=5.726), SVC (MD=3.000+-0.500, MAD=0.000, MR=5.323), GB (MD=4.500+-0.500, MAD=0.500, MR=3.500), ET (MD=4.700+-0.500, MAD=0.700, MR=3.306), KNN (MD=6.000+-0.000, MAD=0.000, MR=2.129), and LR (MD=7.000+-0.000, MAD=0.000, MR=1.081).

Based on the Nemenyi test (Pereira et al., 2015), we inferred that there are no significant differences within the CTB, RF, and SVC groups; GB, ET, and KNN groups; KNN and LR groups. However, all other comparisons between populations revealed statistically significant differences.

As shown in Figure 18, these results underscore the importance of the appropriate choice of statistical methods and indicate notable divergences in distributions among the studied populations. These findings contribute to a deeper understanding of variability between groups, providing valuable insights for future statistical analyses and clinical interpretations.

In the following section, we will enter the conclusion, where we will consolidate the results obtained regarding the models and their evaluation criteria presented in the results section.

In the following section, we will enter the conclusion, where we will consolidate the results obtained regarding the models and their evaluation criteria presented in the results section.

4. Conclusion

heart failure prediction using machine learning models. We delve into not only the results achieved by the various models employed but also essential considerations regarding limitations, ethical considerations, and future perspectives. This analysis aims to provide a comprehensive synthesis of the study, highlighting significant contributions, and outlining areas that warrant further attention in subsequent research.

In the culmination of this research, we successfully achieved the objectives and goals outlined at the inception of the study. The pursuit of developing machine learning models capable of predicting the occurrence of heart failure based on lifestyle-related characteristics has been fully realized. Initially, a detailed analysis of the Kaggle-collected dataset was conducted, unveiling significant correlations between lifestyle features and the incidence of heart failure through statistical techniques and data mining methods. Subsequently, we proceeded with the development and training of machine learning models, focusing on predicting the probability of heart failure occurrence based on the most common characteristics identified in the preceding stage. The meticulous identification and comparison of models, assessing performance metrics such as accuracy, permutation feature importance plot, Confusion Matrix, and ROC Curve, enabled us to select the model with the most promising results for heart failure prediction.

When comparing the results obtained in this study with the findings in related works, significant distinctiveness highlighting the uniqueness of this research is observed. In contrast to approaches employing neural networks, we opted for traditional models such as RF and LR, aiming for superior computational efficiency translated into reduced processing time and resource consumption. This strategic choice, besides enabling a more accessible implementation, did not compromise accuracy, as a test precision close to the best accuracies observed in related

works was achieved, despite the limited quantity of data and the distribution of variables across diverse locations. Additionally, a noteworthy difference lies in the execution of statistical tests to assess the presence of significant differences between models. This more detailed approach provided robust statistical validation to the conclusions, imparting greater reliability to the achieved results. Moreover, it is emphasized that this study encompassed a broader variety of machine learning models, expanding the scope of the research and contributing to a more comprehensive understanding of heart failure prediction compared to some previous studies in the field.

Despite the promising outcomes achieved by the Random Forest (RF) model, along with SVM and CTB, in predicting heart failure through the analysis of lifestyle-related characteristics, an inherent limitation of this study lies in its primarily theoretical nature. The models were developed, trained, and tested within a controlled, simulated environment, relying on datasets that, although comprehensive, do not fully encapsulate the complexities and variabilities of real-world clinical scenarios. This theoretical foundation, while essential for initial exploration and understanding, may not accurately predict the effectiveness and applicability of the models when deployed in actual healthcare settings. The controlled environment of the study does not account for the unpredictable nature of patient responses, the variability of clinical conditions, or the potential for unforeseen factors that could influence the predictive accuracy of the models. Therefore, the transition from a theoretical model to practical application in clinical settings necessitates further empirical research, including pilot studies and clinical trials, to validate the models' efficacy, adaptability, and reliability in real-world scenarios. This step is crucial for ensuring that the predictive models can truly enhance clinical decision-making, improve patient outcomes, and contribute effectively to the management and prevention of heart failure.

A notable limitation arises from the relatively small number of cases in our dataset. The limited quantity of instances may impact the robustness and generalizability of the models, highlighting the need for cautious interpretation and consideration of this constraint (Xu et al., 2023).

A consideration that serves as a limitation of this study is the increased computational resource usage stemming from hyperparameter optimization. When employing exhaustive search methods such as GridSearch, or stochastic methods like RandomSearch, to enhance model performance, multiple trainings with various hyperparameter sets are required. Each iteration demands intensive processing, raising the need for CPU time, memory, and potentially even GPU resources, depending on the complexity of the data and models. This requirement can lead to prolonged training processes and increased consumption of energy and other computational resources, which results in higher operational costs. Thus, while hyperparameter optimization is crucial for achieving desired accuracy, it also imposes significant considerations regarding computational efficiency and the sustainability of the utilized resources.

While there are common concerns regarding the reliability of datasets collected from online sources such as Kaggle (Miller et al., 2022), it is important to acknowledge the benefits of a dataset formed from the merging of databases from different locations. Contrary to the limitations of datasets originating from a single region, which can be highly biased and unrepresentative, the amalgamation of data from five distinct locations can indeed enrich the research, offering a more holistic view and mitigating regional biases. This enhances the models' generalizability to broader populations. Nevertheless, the predictive capability of the study may still be affected by the availability and accuracy of lifestyle-related data, which need to be comprehensive enough to encompass all relevant factors affecting heart failure.

The chosen machine learning models and their predictive performance are subject to the specific characteristics and patterns present in the dataset, potentially limiting the applicability of the models to diverse populations or datasets with distinct features.

Moreover, it is important to emphasize that the effectiveness of this study may not translate in the same manner to other diseases (Oakden-Rayner et al., 2020), even within the group of heart diseases, due to specific characteristics and distinct mechanisms of each condition. Therefore, the generalization of results to different clinical contexts should be approached with caution. Highlighting the need for careful interpretation and consideration of potential limitations.

The future applications of the machine learning models developed to predict heart failure can extend beyond the scope of the current study. One potential application involves integrating predictive models into clinical decision support systems. By incorporating these models into healthcare practices, medical professionals can benefit from automated risk assessments, aiding in early detection and personalized treatment strategies for individuals at risk of heart failure (Dalal, 2020).

Moreover, the models developed in this study could serve as a foundation for broader predictions of cardiovascular health. Expanding the scope to include a variety of cardiovascular conditions and related lifestyle factors can enhance the utility of predictive models, providing a comprehensive tool for assessing overall heart health.

Possible future research directions include investigating the impact of lifestyle factors and additional clinical variables on the predictive accuracy of the models. Exploring diverse datasets from different demographic groups and geographic regions would contribute to the generalization and robustness of the models, making them applicable to a broader population.

Furthermore, there is potential for collaborative efforts among data scientists, healthcare professionals, and policymakers to implement these predictive models in preventive health initiatives. By identifying individuals at high risk of heart failure, interventions and lifestyle modifications can be targeted, potentially reducing the overall burden of cardiovascular diseases on healthcare systems.

In conclusion, the developed machine learning models have the potential for practical applications in clinical settings and public health initiatives. Future efforts can focus on refining and expanding these models to address a wider range of cardiovascular conditions, increase predictive accuracy, and facilitate their integration into real-world healthcare practices.

In conclusion, the three primary models RF, SVM, and CTB achieved superior results in the evaluated metrics, encompassing testing accuracy, training accuracy, confusion matrix, and AUC. It is noteworthy that there exists no statistically significant difference between these top-performing models and the remaining ones in the assessment. Moreover, among the three leading models, RF exhibited a slightly superior performance, establishing itself as the most effective choice based on the comprehensive evaluation of the metrics.

Although the Random Forest RF model did not achieve the second-highest test accuracy, it excelled with one of the best AUC, the most favorable confusion matrix, and outstanding results in statistical tests. These aspects underscore the RF's robustness, not just in classification accuracy but also in its ability to balance sensitivity and specificity, as well as maintain consistent performance across different testing conditions.

References

Agarap, A. F. (2018). On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, 5–9. https://doi.org/10.1145/3184066.3184080

Aldi, F., Hadi, F., Rahmi, N. A., & Defit, S. (2023). Standardscaler's Potential in Enhancing Breast Cancer Accuracy Using Machine Learning. *Journal of Applied Engineering and Technological Science* (JAETS), 5(1), 401–413. https://doi.org/10.37385/jaets.v5i1.3080

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, 16(2), 88. https://doi.org/10.3390/a16020088

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2), 587–603. https://doi.org/10.1016/j.ijforecast.2020.07.007

Bun, J., Bouchaud, J.-P., & Potters, M. (2017). Cleaning large correlation matrices: Tools from Random Matrix Theory. *Physics Reports*, 666, 1–109. https://doi.org/10.1016/j.physrep.2016.10.005

Colliot, O. (Ed.). (2023). *Machine Learning for Brain Disorders* (Vol. 197). Springer US. https://doi.org/10.1007/978-1-0716-3195-9

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble machine learning: Methods and applications*, 157-175.

Dalal, K. R. (2020). Analysing the implementation of machine learning in healthcare. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE. https://doi.org/10.1109/ICESC48915.2020.9156061

Dhanabal, S., & Chandramathi, D. S. (n.d.). A Review of various k-Nearest Neighbor Query Processing Techniques. *International Journal of Computer Applications*, 31.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support (arXiv:1810.11363). arXiv. http://arxiv.org/abs/1810.11363

Fedesoriano. (2021, September). Heart failure prediction dataset. Kaggle. Retrieved from https://www.kaggle.com/fedesoriano/heart-failure-prediction

Groenewegen, A., Rutten, F. H., Mosterd, A., & Hoes, A. W. (2020). Epidemiology of heart failure. *European Journal of Heart Failure*, 22(8), 1342–1356. https://doi.org/10.1002/ejhf.1858

Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. *Journal of Educational and Behavioral Statistics*, 44(3), 348–361. https://doi.org/10.3102/1076998619832248

Hashi, E. K. & Md. Shahid Uz Zaman. (2020). Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. *Journal of Applied Science & Process Engineering*, 7(2), 631–647. https://doi.org/10.33736/jaspe.2639.2020

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, 34(6), 357–359. https://doi.org/10.1136/emermed-2017-206735

Hickman, S. E., Baxter, G. C., & Gilbert, F. J. (2021). Adoption of artificial intelligence in breast imaging: Evaluation, ethical constraints and limitations. *British Journal of Cancer*, 125(1), 15–22. https://doi.org/10.1038/s41416-021-01333-w

Ileberi, E., Sun, Y., & Wang, Z. (2021). Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost. *IEEE Access*, 9, 165286–165294. https://doi.org/10.1109/ACCESS.2021.3134330

Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, 1022(1), 012072. https://doi.org/10.1088/1757-899X/1022/1/012072

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning. Nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Liang, J., et al. (2021). Efficient and privacy-preserving decision tree classification for health monitoring systems. *IEEE Internet of Things Journal, 8*(16), 12528–12539. https://doi.org/10.1109/JIOT.2021.3066307

Mammone, A., Turchi, M., & Cristianini, N. (2009). Support vector machines. *WIREs Computational Statistics, 1*(3), 283–289. https://doi.org/10.1002/wics.49

Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Weather and Forecasting*, 19(6), 1106–1114. https://doi.org/10.1175/825.1

Miller, C., et al. (2022). Limitations of machine learning for building energy prediction: ASHRAE Great Energy Predictor III Kaggle competition error analysis. *Science and Technology for the Built Environment, 28*(5), 610–627. https://doi.org/10.1080/23744731.2022.2067466

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 11th International Conference on Information and Communication Systems (ICICS)* (pp. 243–248). IEEE. https://doi.org/10.1109/ICICS49469.2020.239556

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, 7, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Re, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 151–159. https://doi.org/10.1145/3368555.3384468

Patro, S. G. K., & Sahu, K. K. (2015). Normalization: A Preprocessing Stage. IARJSET, 20–22. https://doi.org/10.17148/IARJSET.2015.2305

Pereira, D. G., Afonso, A., & Medeiros, F. M. (2015). Overview of Friedman's Test and Post-hoc Analysis. *Communications in Statistics - Simulation and Computation*, 44(10), 2636–2653. https://doi.org/10.1080/03610918.2014.931971

Rosario, G. E., Rundensteiner, E. A., Brown, D. C., Ward, M. O., & Huang, S. (2004). Mapping Nominal Values to Numbers for Effective Visualization. *Information Visualization*, 3(2), 80–95. https://doi.org/10.1057/palgrave.ivs.9500072

Saber, M., Boulmaiz, T., Guermoui, M., Abdrabo, K. I., Kantoush, S. A., Sumi, T., Boutaghane, H., Nohara, D., & Mabrouk, E. (2022). Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. *Geocarto International*, 37(25), 7462–7487. https://doi.org/10.1080/10106049.2021.1974959

Shaker, C. R., Sidhartha, A., Praveena, A., Chrsity, A., & Bharati, B. (2022). An Analysis of Heart Disease Prediction using Machine Learning and Deep Learning Techniques. *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 1484–1491. https://doi.org/10.1109/ICOEI53556.2022.9776745

Weerts, H. J. P., Mueller, A. C., & Vanschoren, J. (2020). *Importance of Tuning Hyperparameters of Machine Learning Algorithms* (arXiv:2007.07588). arXiv. http://arxiv.org/abs/2007.07588

Shiba, N., Nochioka, K., Miura, M., Kohno, H., Shimokawa, H., & on behalf of the CHART-2 Investigators. (2011). Trend of Westernization of Etiology and Clinical Characteristics of Heart Failure Patients in Japan: – First Report From the CHART-2 Study –. *Circulation Journal*, 75(4), 823–833. https://doi.org/10.1253/circj.CJ-11-0135

Xu, P., Ji, X., Li, M., & Lu, W. (2023). Small data machine learning in materials science. *Npj Computational Materials*, 9(1), 42. https://doi.org/10.1038/s41524-023-01000-z

Ying, X. (2019). An Overview of Overfitting and its Solutions. J*ournal of Physics: Conference Series*, 1168, 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering, 2*(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics - Theory and Methods*, 49(9), 2080–2093. https://doi.org/10.1080/03610926.2019.1568485

Zhao, W. P., Li, J., Zhao, J., Zhao, D., Lu, J., & Wang, X. (2020). XGB Model: Research on Evaporation Duct Height Prediction Based on XGBoost Algorithm. *Radioengineering*, 29(1), 81–93. https://doi.org/10.13164/re.2020.0081