Ahliya Journal of Allied Medico-Technology Science
[Published by Palestine Ahliya University]

Vol. 01 Issue 01 (2024)
ISSN: 3007-9489

Predicting the Risk of Myocardial Infarction (MI) using Machine Learning (ML)

Murad Zeer[1], Mutaz Rasmi Abu Sara[1], Asma Sbeih[1] and Khaled Sabarna[2]
[1]IT Department, Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)
✉ muradzeer@paluniv.edu.ps
✉ moutaz.a@paluniv.edu.ps
✉ Asma_sbeih@paluniv.edu.ps
[2]Nursing Department, Faculty of Allied Medical Sciences, Palestine Ahliya University (Palestine)
✉ k.sabarna@paluniv.edu.ps

Received: 25/04/2024 Accepted:22/05/2024 Published:30/06/2024

Abstract: *The purpose of this study was to use machine learning to forecast the likelihood of a myocardial infarction and identify when one would occur. The study made use of a preprocessed dataset regarding cardiac attacks from Kaggle. The purpose of this study was to use machine learning to forecast the likelihood of a heart attack and identify when one would occur. The research used a preprocessed dataset from Kaggle related to heart attacks. K-nearest neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, Decision Tree, Naive Bayes, XGBoost, Random Forest, and Gradient Boosting were the eight techniques used in the study. According to the findings, the models that predicted heart attacks in our dataset the best were Decision Tree and Gradient Boosting. These models showed excellent precision, recall, and F1-score balance, among other important criteria. They can effectively reduce overfitting and generalize well to new data thanks to their ability to handle complicated, non-linear interactions and their use of regularization and ensemble learning techniques. Decision Tree and Gradient Boosting are the most reliable options for this predictive task because of their all-encompassing strengths, even if models like XGBoost and Random Forest also performed well while Logistic Regression and SVC produced strong results.*

Keywords: *Heart Attack, Atherosclerosis, Myocardial Infarction, Machine Learning, Classification.*

1. Introduction

Heart conditions are common all around the world and are the leading causes of death. Their symptoms often coincide with those of other illnesses, making a fast and correct diagnosis extremely difficult. Delayed detection of heart attacks can worsen health outcomes for patients [1][2][3], as heart attacks result from interrupted blood flow to the heart, leading to tissue damage [4]. With abundant data on heart attacks available through technological advancements and storage, leveraging this data with artificial intelligence can help understand the underlying causes of heart attacks resulting from complete coronary artery occlusion [5]. Despite the difficulty in diagnosing heart attacks, machine learning has made it possible [6]. A heart attack is also called a myocardial infarction. it happens when a significant reduction in or interruption of the blood flow supplies the heart muscle with oxygen-rich blood. This is because the accumulation of fat, cholesterol, and other materials (plaque) narrows the coronary arteries. We refer to this process as atherosclerosis. Several reasons are thought to induce atherosclerosis, including hereditary and environmental influences. It takes several decades for clinical problems to manifest in humans. Several reasons are thought to induce atherosclerosis, including hereditary and environmental influences. It takes several decades for clinical problems to manifest in humans [8]. encompassing both environmental and genetic variables. It takes several decades for clinical problems to manifest in humans. Although there are numerous established risk factors for atherosclerosis, such as smoking, high blood pressure, diabetes, and hypercholesterolemia, it is generally accepted that atherosclerosis is a chronic inflammation of the blood vessels brought on by the interaction of these risk factors with arterial wall cells. encompassing both environmental and genetic variables. It takes several decades for clinical problems to manifest in humans. Although there are numerous established risk factors for atherosclerosis, such as smoking, high blood pressure, diabetes, and hypercholesterolemia, it is generally accepted that atherosclerosis is a chronic inflammation of the blood vessels brought on by the interaction of these risk factors with arterial wall cells [8]. Chest discomfort that radiates from the left arm to the neck, shortness of breath, perspiration, nausea, vomiting, irregular heartbeat, anxiety, exhaustion, weakness, stress, depression, and other symptoms are some of the signs and symptoms of MI [9]. Women are more likely than males to require hospitalization again within the first year following discharge; however, very few studies have examined the daily variations in readmission risk by sex during this entire year and how these variations relate to the mortality risk. Women are more likely than males to require hospitalization again within the first year following discharge; however, very few studies have examined the daily variations in readmission risk by sex during this entire year and how these variations relate to the mortality risk [10]. Finally, it should be noted that there are gender differences in the rates of unanticipated rehospitalization following PCI, with over 10% of women undergoing PCI experiencing readmission within 30 days. While significant differences were seen for cardiovascular causes, gender disparities were not seen for non-cardiac readmission causes [11].

Machine learning is a method for processing data and extracting information, involving various supervised and unsupervised learning classifiers extensively used in prediction tasks [7]. This makes it suitable for predicting heart attacks, as demonstrated by previous studies [1][2][4][5][6], among others, which utilized machine learning to predict and identify heart attacks.

This study aims to predict heart attacks using machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, Decision Tree, Naive Bayes, XGBoost, Random Forest, and Gradient Boosting.

2. Previous Studies:

Study [1] aimed to predict heart diseases using machine learning, employing Random Forest, Decision Tree, SVM, Bayes, and KNN with particle swarm optimization (PSO) for feature selection. The proposed SVM model achieved an accuracy of 94.3%, while other algorithms achieved accuracies between 85% and 90%.

Study [2] utilized Decision Tree, Random Forest, Logistic Regression, and Gradient Boosting, finding that Gradient Boosting achieved the highest accuracy at 84.7%, with Logistic Regression achieving the lowest at 69%.

Study [3] used Logistic Regression, Decision Tree, and SVM, with Logistic Regression achieving the highest accuracy at 87%.

Study [4] employed SVM, Logistic Regression, XGBoost, and Naive Bayes, finding that XGBoost achieved the highest accuracy at 94%, followed by Logistic Regression at 92%, and SVM the lowest at 75%.

Study [5] used SVM with Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), achieving the highest accuracies of 91.8% with SVM using LDA, linear, and RBF kernels.

Study [6] utilized Decision Tree, Logistic Regression, Naive Bayes, SVM, and KNN, finding that SVM achieved the highest accuracy at 91%.

Study [7] employed KNN, Logistic Regression, and Random Forest Classifier, with KNN achieving the highest accuracy at 88.52%.

Study [8] Atherosclerosis is thought to be caused by several factors, including environmental and genetic factors. In humans, it takes several decades for clinical issues to appear.

Study [9] Some of the indications and symptoms of MI include shortness of breath, sweating, nausea, vomiting, irregular heartbeat, anxiety, weariness, weakness, stress, depression, and chest tightness that extends from the left arm to the neck.

Study [10] Within the first year after discharge, women are more likely than men to need to be readmitted to the hospital. However, very few studies have looked at the daily fluctuations in readmission risk by sex over this full year, and how these variations relate to the mortality risk.

Study [11] women who get PCI wind up being readmitted within 30 days. Gender discrepancies were not observed for non-cardiac readmission, despite considerable differences being observed for cardiovascular reasons.

3. Methodology

3.1 Dataset

The study relied on data from heart attacks published on Kaggle (https://www.kaggle.com/), which consisted of 1319 rows and 8 inputs, with one output (Classification: presence or absence of a heart attack). The following table describes the dataset used in this study.

Table 1: Dataset Description

	Description
Source	Kaggle
Shape	(1319 rows × 8 Columns)
Input	Age, gender, impulse, pressurehight, pressurelow, glucose, kcm, troponin
Output	Class

3.2 Data Preprocessing

Data preprocessing is a crucial step to ensure the quality of data used in predictive and analytical models. This process involves several stages aimed at cleaning and preparing data for use in machine learning models [12]. In this study, data preprocessing included checking for missing values, removing outliers, and selecting appropriate features for analysis.

3.3 Missing Values

One fundamental aspect of data preprocessing is verifying the presence of any missing values, as these can negatively impact model performance. In the dataset used in this study, it was ensured that there were no missing values. This verification enhances data reliability and ensures that all models will deal with complete data, thereby contributing to result accuracy and avoiding bias.

3.4 Outliers

Outliers are data points that lie outside the normal range and significantly affect analysis results. To remove these outliers, the Interquartile Range (IQR) method was employed. This method identifies outliers by determining values that fall outside 1.5 times the interquartile range beyond the first and third quartiles. Removing outliers helps reduce the influence of these abnormal values and ensures that models are not affected by large deviations in the data.

3.5 Feature Selection

To identify the most impactful features in the data, a Correlation Matrix was used to examine the relationship between inputs and outputs. This matrix helps determine the strength and direction of relationships between variables. The results of the correlation matrix are shown in Figure (1).

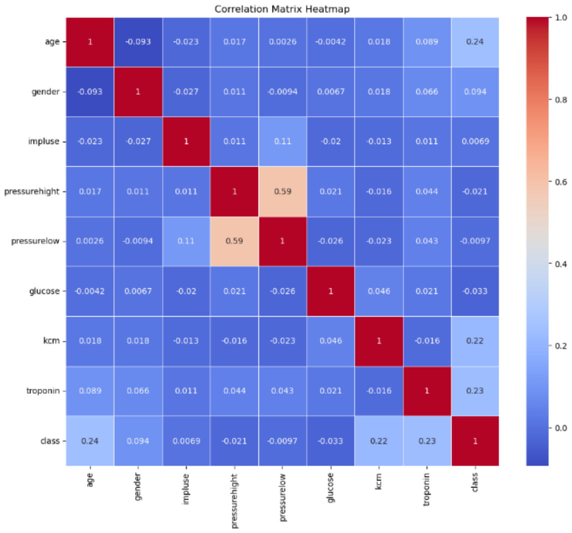

Figure (1): Correlation Matrix

Based on the results in the figure, it was found that the variables "impulse" and "pressurelow" had the least impact on the data, showing weak correlations with the outputs. Consequently, these variables were removed during the data preprocessing stage. This step helps simplify the model and reduce its complexity, increasing analysis efficiency and minimizing the potential bias resulting from noise in the data. After removing non-influential variables, the dataset was updated to better align with analytical objectives, contributing to the development of more accurate and efficient models.

3.6 Study Model

Following the preprocessing of the dataset, the inputs (Age, gender, pressurehight, glucose, kcm, troponin) were adopted, with "Class" serving as the categorical output for heart attacks (heart attack present, no heart attack present).

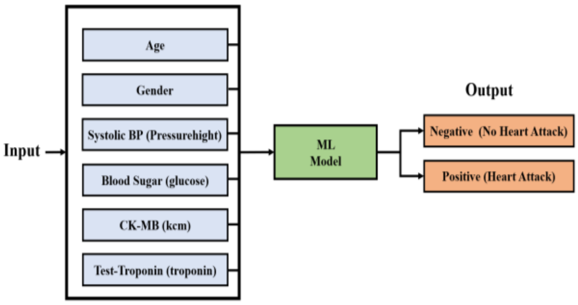

Figure (2): Study Model

3.7 Algorithms

In this study, a diverse set of algorithms was employed to analyze data and make predictions, chosen for their robustness and performance across various problems. The algorithms include K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Logistic Regression, Decision Tree, Naive Bayes, XGBoost, Random Forest, and Gradient Boosting. For KNN, the parameter k was set to 13, balancing computational efficiency and accuracy. The SVC was configured with a regularization parameter C of 100 and a gamma value of 0.1 to enhance the decision boundary's precision. Logistic Regression utilized a regularization parameter C of 100, optimizing the model's fit. The Decision Tree algorithm was fine-tuned with a maximum depth of 5, and a minimum of 4 samples required to split a node. Naive Bayes, known for its effectiveness in high-dimensional data, was included for its simplicity and performance in text classification. XGBoost, a powerful boosting algorithm, was selected for its speed and flexibility in handling various data complexities. The Random Forest model was optimized with parameters including bootstrap set to True, no maximum depth restriction, a minimum of 1 sample per leaf, a minimum split of 2 samples, and 100 estimators, enhancing the model's stability and accuracy. Lastly, Gradient Boosting was configured with a learning rate of 0.01, a maximum depth of 5, a minimum of 1 sample per leaf, a minimum split of 10 samples, and 50 estimators, focusing on iterative correction of prediction errors to improve overall model performance. These algorithms were chosen and tuned to maximize prediction accuracy and address different aspects of the data analysis challenges, contributing to a comprehensive and reliable predictive model.

4. Results:

After using algorithms to predict the classification of heart attacks, researchers obtained accuracy, precision, recall, and F1-score, as shown in Table (1) below:

Table (1): Results of Accuracy of Algorithms Used in the Study

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
KNN (k= 13)	84.3%	86.9%	69.7%	77.3%
SVC (C= 100, gamma= 0.1)	93.4%	90.9%	92.1%	91.5%
LogisticRegression (C=100)	93.4%	89.8%	93.4%	91.6%
Decision Tree	98.4%	97.4%	98.6%	98.03%
Bayes	94.4%	94.5%	90.7%	92.6%
XGBoost	97.9%	96.1%	98.6%	97.4%
Random Forest	97.4%	97.3%	96.05%	96.6%
Gradient Boosting	98.4%	97.4%	98.6%	98.03%

Taking a closer look at the data at the table, we see that Decision Tree and Gradient Boosting are the champions when it comes to predicting heart attacks. Both reach a staggering 98.4% accuracy, along with very high precision, recall, and F1-score. This suggests they're the most powerful tools for this particular dataset. XGBoost and Random Forest are also top contenders, achieving impressive accuracies close to 98% and mirroring the strong performance in other metrics. This indicates their ability to strike a good balance between catching true positives and avoiding false alarms. Naive Bayes comes in a strong third place with a 94.4% accuracy and well-rounded scores across the board. Logistic Regression and SVC are decent performers, both reaching 93.4% accuracy. However, Logistic Regression falls a bit behind SVC in terms of precision and recall. K-Nearest

Neighbors (KNN) lags behind the others, with an accuracy of only 84.3%. Its lower precision, recall, and F1-score suggest it's not as effective for this specific task. In conclusion, Decision Tree and Gradient Boosting are the clear winners, demonstrating exceptional accuracy and a balanced performance across all metrics.

5. Discussion:

Out of all the models tested, Decision Trees and Gradient Boosting shine brightest at predicting heart attacks in our data. Because they can handle the tricky, non-obvious connections between different factors. Decision Trees, when adjusted just right, strike a balance between complexity and overfitting the data. This sweet spot leads to high accuracy and reliable performance. Gradient Boosting is like a team effort, constantly learning from its mistakes and adding new "mini-predictions" to get closer to the truth. It also throws in some safeguards to prevent overfitting, making its predictions more generalizable. XGBoost, a particular type of Gradient Boosting, and Random Forest, which combines predictions from many Decision Trees, also perform very well thanks to their strong team approaches. On the other hand, some models like K-Nearest Neighbors struggle when dealing with a lot of data points. Support Vector Machines (SVC) heavily depend on specific settings, and Logistic Regression might miss some important connections because it assumes things are always straightforward. In the end, Decision Trees and Gradient Boosting win because they're masters at handling complex relationships between factors, have built-in safeguards, and leverage the power of teamwork.

6. Conclusion:

In conclusion, the Decision Tree and Gradient Boosting models emerged as the best performers for predicting heart attacks in our dataset, demonstrating superior accuracy and balance across key metrics such as precision, recall, and F1-score. Their ability to handle complex, non-linear relationships, coupled with effective regularization and ensemble learning techniques, allows these models to generalize well to new data and minimize overfitting. While other models like XGBoost and Random Forest also performed well, and Logistic Regression and SVC provided strong results, the comprehensive strengths of Decision Tree and Gradient Boosting make them the most reliable choices for this predictive task.

References

[1] M. Alshraideh, N. Alshraideh, A. Alshraideh, Y. Alkayed, Y. Al Trabsheh, and B. Alshraideh, "Enhancing Heart Attack Prediction with Machine Learning: A Study at Jordan University Hospital," *Applied Computational Intelligence and Soft Computing*, 2024.

[2] S. K. Gupta, A. Shrivastava, S. P. Upadhyay, and P. K. Chaurasia, "A machine learning approach for heart attack prediction," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 10, no. 6, pp. 1-11, 2021.

[3] Y. N. V. S. Prakash, B. Sathyam, M. S. Venkat, B. P. M. Rao, M. M. Naik, and S. Panchikkil, "Heart Attack Detection using Machine Learning," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, March 2024, pp. 1-6.

[4] N. Nandal, L. Goel, and R. TANWAR, "Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis," *F1000Research*, vol. 11, p. 1126, 2022.

[5] S. Vamshi Kumar, T. V. Rajinikanth, and S. Viswanadha Raju, "Heart Attack Classification Using SVM with LDA and PCA Linear Transformation Techniques," in *Machine Learning Technologies and Applications: Proceedings of ICACECS 2020*, Springer Singapore, 2021, pp. 99-112.

[6] M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik, and A. U. Zaman, "Heart disease prognosis using machine learning classification techniques," in *2021 6th International Conference for Convergence in Technology (I2CT)*, April 2021, pp. 1-6.

[7] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "heart disease prediction using machine learning algorithms," in *IOP conference series: materials science and engineering*, vol. 1022, no. 1, 2021, p. 012072.

[8] J. Fan and T. Watanabe, "Atherosclerosis: Known and unknown," Pathology International, vol. 72, no. 3, pp. 151-160, 2022.

[9] L. Lu, M. Liu, R. Sun, Y. Zheng, and P. Zhang, "Myocardial Infarction: Symptoms and Treatments," Cell Biochemistry and Biophysics, vol. 72, no. 3, pp. 865-867, Jul. 2015, doi: 10.1007/s12013-015-0553-4.

[10] R. P. Dreyer, K. Dharmarajan, A. F. Hsieh, J. Welsh, L. Qin, and H. M. Krumholz, "Sex Differences in Trajectories of Risk After Rehospitalization for Heart Failure, Acute Myocardial Infarction, or Pneumonia," Circulation: Cardiovascular Quality and Outcomes, vol. 10, no. 5, p. e003271, May 2017, doi: 10.1161/CIRCOUTCOMES.116.003271.

[11] C. S. Kwok et al., "Effect of Gender on Unplanned Readmissions After Percutaneous Coronary Intervention (from the Nationwide Readmissions Database)," American Journal of Cardiology, vol. 121, no. 7, pp. 810-817, Apr. 2018, doi: 10.1016/j.amjcard.2017.12.032.

[12] H. N. Zalloum, S. Al Zeer, A. Manassra, M. R. Abu Sara, and J. H. Alkhateeb, "Breast Cancer Grading using Machine Learning Approach Algorithms," Journal of Computer Science, vol. 18, no. 12, pp. 1213-1218, 2022. doi: 10.3844/jcssp.2022.1213.1218.