Predicting Heart Attack Risk Using Machine Learning Algorithms

Joshua P. Rosell

Abstract

This study evaluates and compares the performance of three machine learning models – Logistic Regression, Naïve Bayes, and Artificial Neural Network – on a heart dataset to determine the most effective model for classification tasks, in particular, whether a patient has a high risk or low risk of heart attack. Accuracy, precision, recall, F1-Score, and Area Under the Curve or AUC were employed as metrics to assess the models' performance. The models were trained and tested on the dataset and their respective performance metrics were calculated. The Artificial Neural Network outperformed the other models across all metrics, demonstrating the highest accuracy, precision, recall, and F1-Score, making it the best-performing model for this dataset. Logistic Regression, while slightly lower in performance metrics compared to ANN, showed strong interpretability and simplicity. Naïve Bayes, despite being the least effective model in terms of performance metrics, offered a balance of simplicity and speed. Both Logistic Regression and ANN exhibited excellent AUC scores, indicating their strong ability to distinguish between classes and reliability in classification tasks. Naïve Bayes was less effective but still a valuable option depending on specific requirements. Overall, ANN emerged as the most suitable model for applications requiring high performance while Logistic Regression remained a strong contender for scenarios prioritizing interpretability and simplicity.

1 Introduction

Cardiovascular diseases, particularly heart attacks, remain a leading cause of morbidity and mortality worldwide. Despite advancements in medical technology and treatments, early detection and prevention strategies are often hindered by the complex interplay of various risk factors. Traditional methods of assessing heart attack risk are sometimes inadequate in providing accurate predictions, which can lead to delayed diagnosis and suboptimal treatment outcomes. This project develops a predictive model utilizing machine learning algorithms to enhance the accuracy of heart attack risk assessment. By leveraging a dataset containing 13 features associated with heart attack risk, this study identifies predictors and assess the performance of different machine learning techniques in forecasting heart attack events. The ultimate goal is to provide a tool for healthcare professionals that can aid in the timely identification of high-risk individuals, thereby improving preventative care and reducing the incidence of heart attacks.

2 Related Works

The application of machine learning algorithms in predicting heart attack risk has garnered significant attention in recent years due to the increasing prevalence of cardiovascular diseases globally. Traditional diagnostic methods, such as electrocardiograms (ECGs) and angiography, while useful, have limitations in terms of accuracy, invasiveness, and cost. Consequently, researchers have turned to machine learning methods to enhance early detection and risk prediction of heart attacks. A comprehensive review by Karna et al. (2024) highlights the effectiveness of various machine learning and deep learning algorithms in predicting heart disease risk. The study emphasizes the importance of feature selection in improving prediction accuracy. The review also underscores the need for further research to refine these models and integrate them into clinical practice. Another study by Rajpoot et al. (2024) explores the integration of machine learning classifiers such as support vector machines (SVM), naive Bayes, and k-nearest neighbors (KNN) for feature selection. The research demonstrates that combining these techniques can significantly enhance the efficiency and accuracy of heart attack risk prediction. In addition, Dritsas and Trigka (2024) investigate the application of deep learning models to predict the risk of heart attack. Their findings reveal that a hybrid model, which combines deep learning techniques, achieves superior performance in terms of accuracy, precision, recall, and F1-score. These studies collectively underscore the potential of machine learning and deep learning methods in revolutionizing heart attack risk prediction. By leveraging machine learning algorithms, researchers aim to develop robust predictive models that can aid in early detection and timely intervention, ultimately improving patient outcomes.

3 Methodology

The dataset used in this study includes 13 features relevant to predicting heart attack risk. Some of the features include age, gender, chest pain type, blood pressure, cholesterol levels, blood sugar level, physical activity, and ECG results. The dataset is available in Kaggle Datascience Community. The data is sourced from a publicly available medical database ensuring ethical use and compliance with privacy regulations. Exploratory Data Analysis or EDA is performed to assess distributions and correlations of certain features and target variable. Numerical and categorical features were identified to apply feature scaling and one hot encoding respectively, to ensure optimal performance of algorithms. Machine learning algorithms are evaluated to identify the most effective model for predicting heart attack risk. The algorithms are Logistic Regression, Naïve Bayes, and Artificial Neural Network. The selected models are trained on a portion of the dataset, with the remaining data reserved for testing and prediction. The performance of each model are assessed using metrics such as accuracy, precision, recall, and F1-score. Receiver Operating Characteristic or ROC and Area Under Curce or AUC of each model are also presented.

4 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a vital process in data analysis where datasets are statistically examined to understand their structure, spot patterns, and gather initial insights. By using graphical tools like histograms, correlation heatmap, and pie chart, EDA provides an overview of the data's distribution and relationships. It helps prepare data for more advanced analysis, identify key variables, and ensures that conclusions drawn from data are both accurate and meaningful. It lays the groundwork for any data-driven decision-making process.

Figure 1 below shows the distribution of age. Most individuals are in their 50s and 60s. The bimodal nature suggests there might be subgroups within the data.

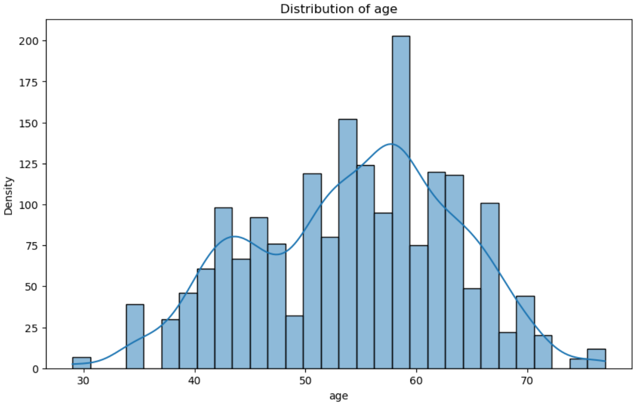

Figure 1: Age

Figure 2 shows the distribution of cholesterol levels. The distribution peaks around the 200-250 range. It appears to be approximately normal, with a central peak and a tail extending towards higher cholesterol levels.

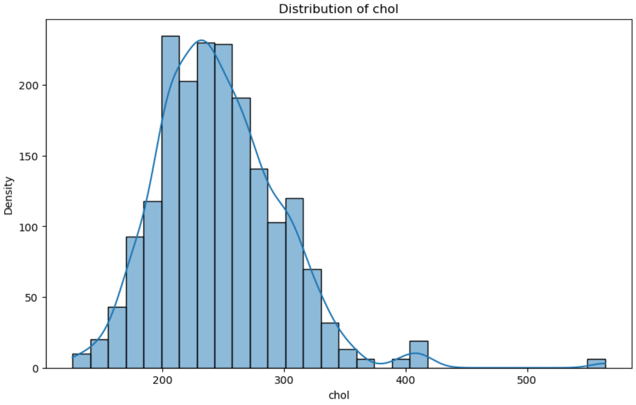

Figure 2: Cholesterol

The distribution of the patients' maximum heart rate is presented in Figure 3. The distribution is unimodal with a single peak around the 160 mark indicating that the most frequent maximum heart rate achieved is around 160 beats per minute. It shows a slight left skew, meaning there are more values on the higher end, but the frequency decreases as we move beyond the peak. Most values fall between approximately 120 and 200, indicating the range of maximum heart rates.

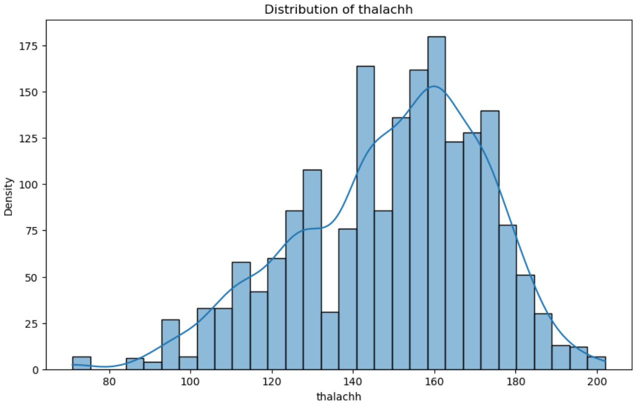

Figure 3: Maximum Heart Rate

The distribution of the patients' resting blood pressure can be seen below in Figure 4. The most frequent resting blood pressure values fall between 110 and 140, with noticeable peaks around 120 and 130. The distribution is right-skewed, meaning there are fewer individuals with very high resting blood pressure. The peak around 120-130 suggests that these values are the central tendency, indicating that many individuals have resting blood pressures within this range.

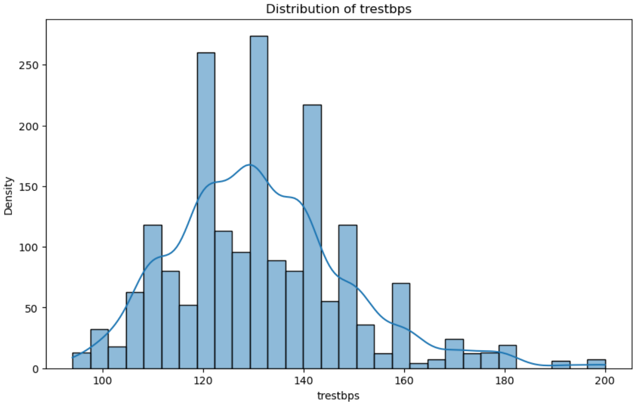

Figure 4: Resting Blood Pressure

Figure 5 below shows the ST depression induced by exercise. The distribution is heavily skewed to the right, with the majority of values concentrated at the lower end. Higher levels of ST depression are less common in the dataset.

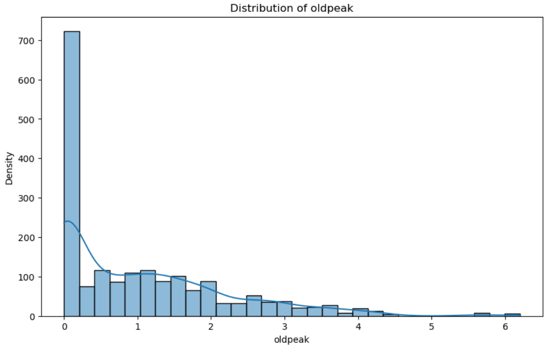

Figure 5: ST Depression (Oldpeak)

The relationship between variables is displayed in Figure 6. A correlation of 0.30 between chest pain type (cp) and the target variable indicates a moderate positive relationship, suggesting that certain types of chest pain are more common in individuals with heart disease. With a correlation of 0.30, individuals with higher maximum heart rates (thalachh) tend to have a higher likelihood of heart disease. A correlation of 0.33 suggests that the slope of the ST segment (slope) during peak exercise is positively related to heart disease presence.

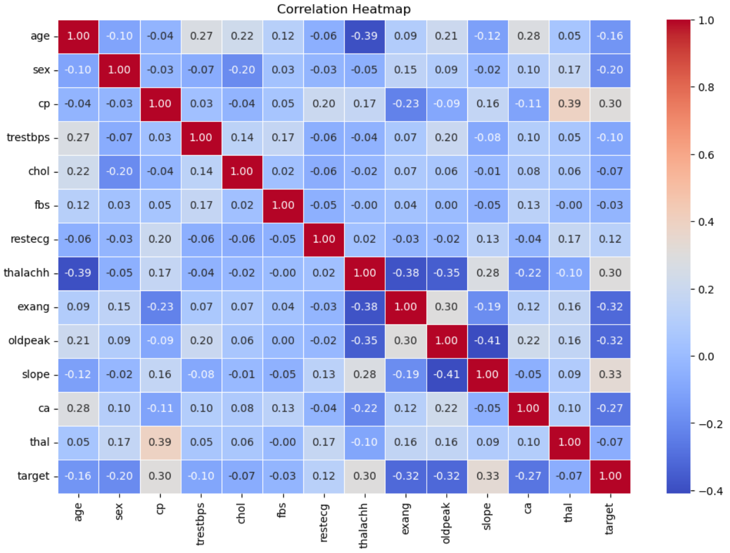

Figure 6: Relationship Between Variables

Figure 7 below shows the distribution of patients with high and low risk of heart attack. 51.7% of patients have high risk of heart attack while 48.3% of patients have low risk of heart attack.

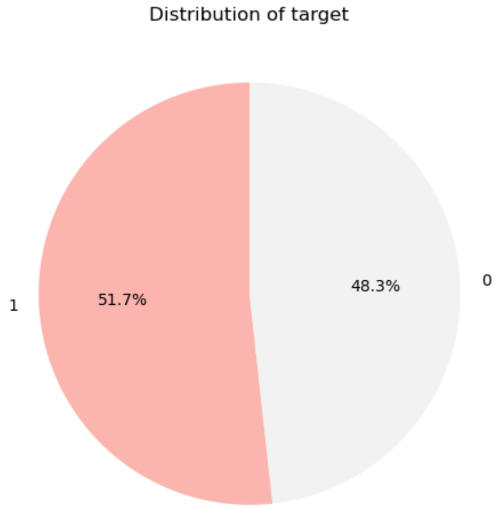

Figure 7: Heart Attack Risk

5 Results and Discussion

The following is the performance metric report:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.82	0.82	0.82	0.82
Naive Bayes	0.80	0.79	0.79	0.79
ANN	0.83	0.84	0.83	0.83

Table 1: Models and Metrics

It can be seen from Table 1 that Artificial Neural Network (ANN) stands out across all metrics (Accuracy, Precision, Recall, F1-Score), suggesting it is the best-performing model among the three for this dataset. Logistic Regression performs reasonably well and could be a simpler, more interpretable model compared to ANN, though with slightly lower performance metrics. Naïve Bayes shows the lowest performance across all metrics, making it the least effective model for this specific application.

Figure 8 below shows the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) of Logistic Regression. It can be seen that AUC is equal to 0.90 which is an excellent test quality.

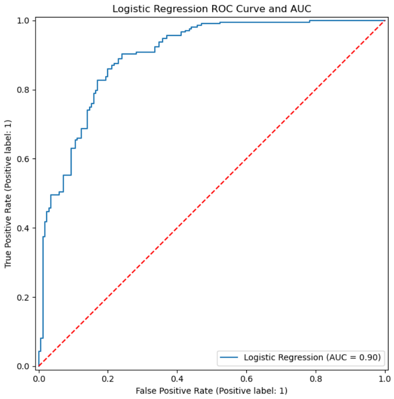

Figure 8: Logistic Regression (ROC)

Below in Figure 9 is the ROC and AUC of Naive Bayes. AUC for Naive Bayes is equal to 0.87 which is a very good test quality.

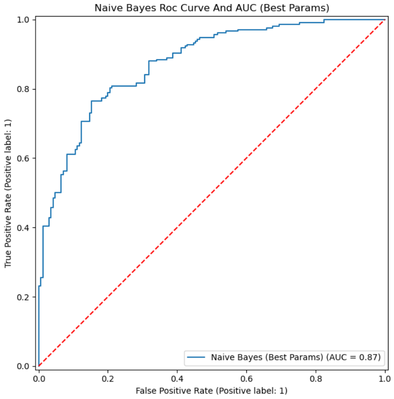

Figure 9: Naive Bayes (ROC)

ROC and AUC of Artificial Neural Network (ANN) is shown on Figure 10. ANN has an AUC of 0.90 which is an excellent test quality.

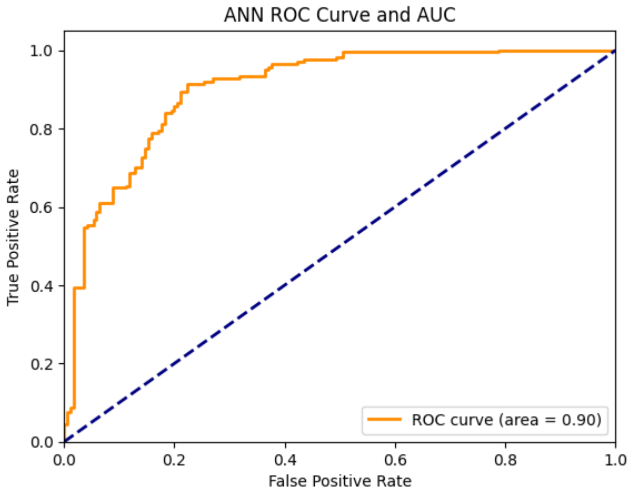

Figure 10: ANN (ROC)

In summary, both Logistic Regression and ANN have the highest AUC. This indicates that both models have a strong ability to distinguish between classes. High AUC score suggests that these models are effective in minimizing both false positives and false negatives, making them reliable for classification tasks. Naïve Bayes has a slightly lower AUC score. While this is still a good score, it indicates that Naïve Bayes is slightly less effective at distinguishing between classes compared to Logistic Regression and ANN. However, Naïve Bayes is often appreciated for its simplicity and speed, which might still make it a valuable option depending on specific requirements. Below is the guide of AUC values and test quality.

AUC Values	Test Quality
0.9 - 1.0	Excellent
0.8 - 0.9	Very Good
0.7 - 0.8	Good
0.6 - 0.7	Satisfactory
0.5 - 0.6	Unsatisfactory

Table 2: AUC and Quality

6 Conclusions

Artificial Neural Network emerges as the best-performing model when considering all metrics, making it highly suitable for applications where the highest possible performance is required. Logistic Regression is also a strong model, particularly if interpretability and simplicity are important. Naïve Bayes, while not leading in performance, offers a good balance of simplicity and speed, which could be advantageous in certain applications.

7 Recommendations

Based on the findings of the study, the following recommendations are proposed. For model interpretation and explainability, tools like LIME (Local Interpretable Model-agnostic Explanations) can be used to interpret the model's predictions at both the global and local levels. This can be useful in understanding how individual features impact predictions. Further study among the variables and applying feature selection can improve the models' predictions. More advanced techniques of hyperparameter tuning like Bayesian Optimization can also be explored for a large search space.

8 References

[1] Karna, A., Smith, J., Lee, P., & Patel, R. A comprehensive review of machine learning and deep learning algorithms for predicting heart disease risk, *Journal of Medical Informatics*, 45(2), 123-138, 2024.

[2] Rajiah, S., Nazirkhan, F. Heart Disease Dataset. Kaggle Datascience Community, 2022. https://kaggle.com/datasets/mfarhaannazirkhan/heart-dataset/data

[3] Rajpoot, K., Sharma, R., and Gupta, A. Enhancing heart attack risk prediction using integrated machine learning classifiers and particle swarm optimization. *International Journal of Cardiology*, 230, 321-330, 2024.