Predicting heart attack risk in individuals based on physiological indicators

Arif Abdulshakour Othman

1 Introduction

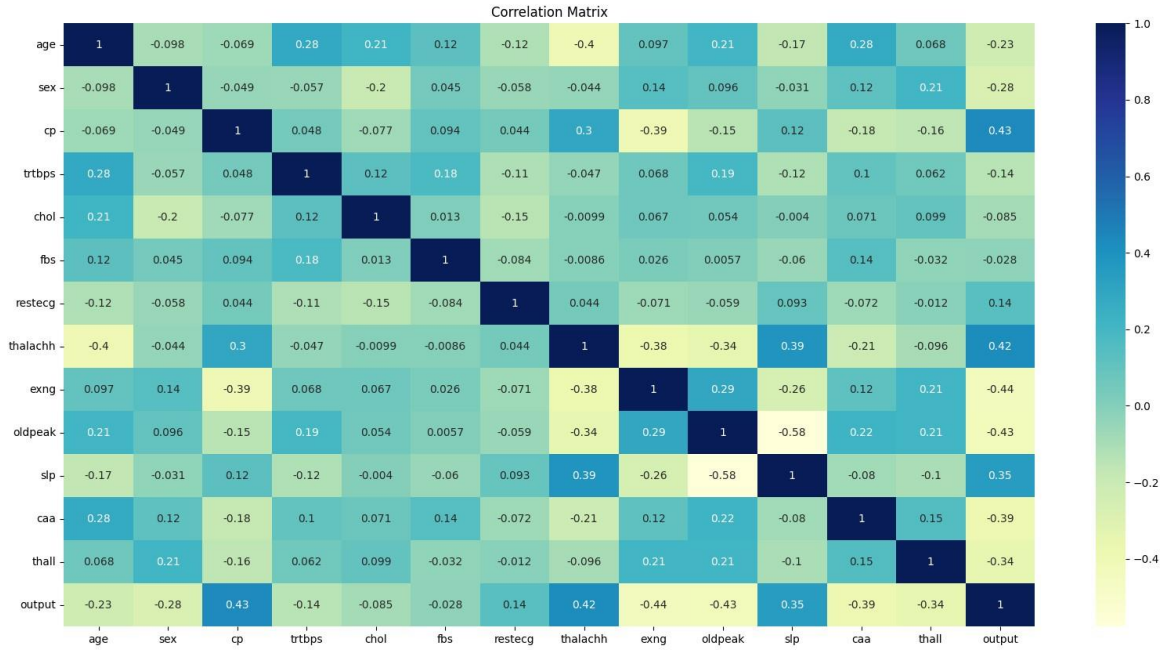
Cardiovascular diseases are the leading global cause of death [3], necessitating a comprehensive preventive healthcare approach. This study addresses the need for early intervention in myocardial infarction risk prediction. In response to evolving healthcare dynamics, there's a demand for innovative, data-driven strategies to enhance diagnostic accuracy. This study explores the potential of machine learning classification algorithms to precisely estimate heart attack risk using a dataset emphasising crucial physiological markers. The central question is whether these models can reliably detect individuals at risk, aligning with current healthcare challenges to reduce the global burden of heart attacks.

1.1 Dataset Overview

The dataset encompasses diverse characteristics capturing physiological and demographic factors related to heart health. The age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, output variable indicating the risk of a heart attack is all listed in the columns (See Appendix 1). Important physiological markers for the study include trestbps, chol, fbs, restecg, thalach, exang, slope. Together, continuous variables (age, trtbps, chol, thalachh, oldpeak) and categorical variables (sex, cp, fbs, restecg, exng, slp, caa, thall) help to provide a comprehensive comprehension of the dataset. This collection of variables serves as the basis for ensuring predictive modelling initiatives by enabling a sophisticated investigation of the intricate interaction between physiological parameters and heart attack risk.

2 Exploratory Data Analysis (EDA)

Utilising pie charts, box plots, pair plots, and heat maps, this study employs visualisations to gain deeper insights into the dataset. Further plot can be found in the img/eda directory.

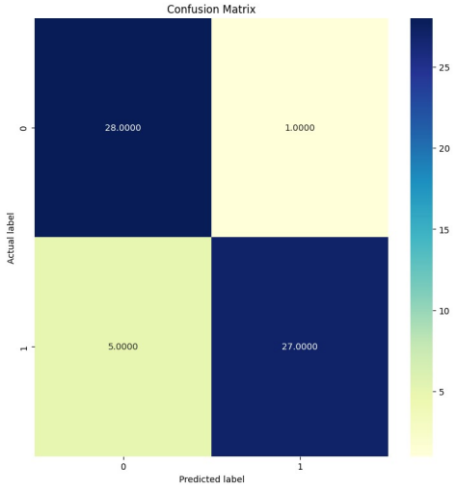


3 Machine Learning

Employing classification, a supervised learning technique, this paper uses labelled data to predict heart attack risk. Logistic Regression, SVM, and KNN are explored using feature engineering strategies like scaling and encoding. Model generalisation is evaluated via the holdout method using train_test_split, and hyper-parameter tuning is implemented using GridSearchCV for optimal performance by cross-validating over a parameter grid.

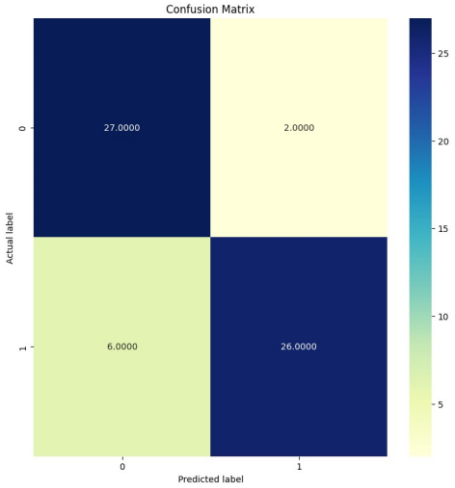
3.1 Logistic Regression

Logistic Regression is utilised for its suitability with binary outcome variables [7]. The model is fine-tuned with parameters such as C and penalty, achieving an accuracy of 90.2%, demonstrating its potential in healthcare applications.



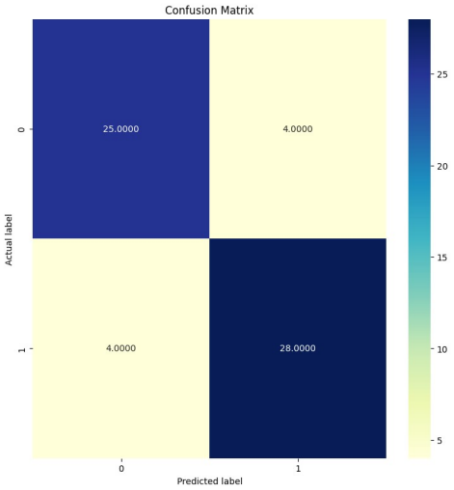
3.2 SVM

Support Vector Machines, fine-tuned with parameters like C and gamma [1], achieve an accuracy of 86.9%, showcasing their potential in healthcare applications.



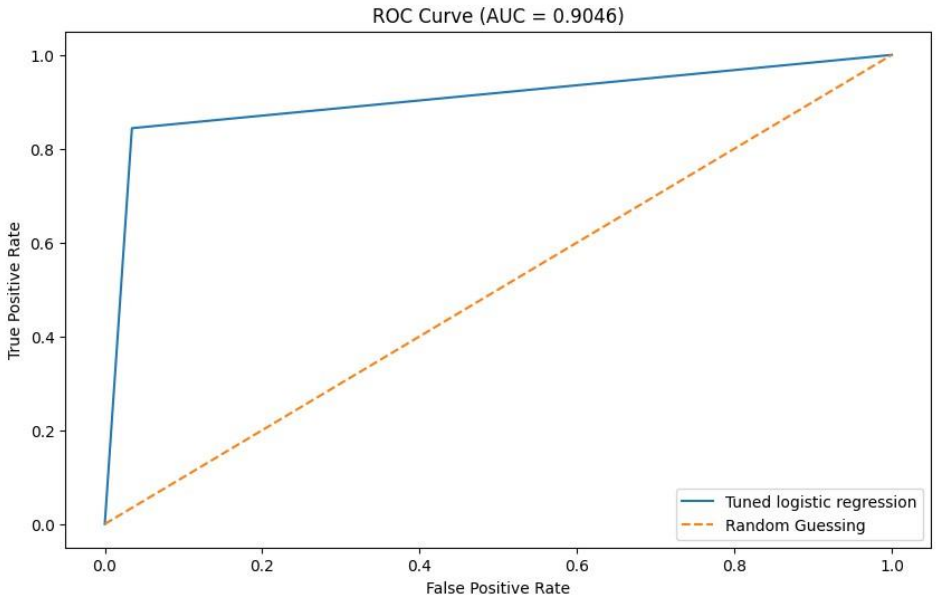
3.3 KNN

K-Nearest Neighbours, fine-tuned with parameters like n_neighbors and weights [4], also achieved an accuracy of 86.9%, highlighting their potential in healthcare applications.



4 Results

The findings of the study that used physiological parameters to predict the likelihood of a heart attack provide fascinating new information about how various categorisation algorithms function. The results show that Logistic Regression is the best-performing model, with an astounding accuracy of 90.2%. But at 84.6%, the cross-validated accuracy score (a crucial indicator of model generalisation) is a little bit lower. It's important to remember that an accuracy score might not give a complete picture of how well a model performs.



A more sophisticated knowledge of discriminating capacities is provided by the area under the curve (AUC), especially when referring to the receiver operating characteristic (ROC) curve. When the AUC is less than 0.5, the performance is comparable to random guessing; when the AUC is more than 0.9, the discrimination is regarded as excellent. Analysing the ROC curve offers a more thorough understanding of each model's prediction ability to determine heart attack risk based on physiological signs by providing a graphical depiction of the trade-off between true positive and false positive rates.

5 Conclusion

In conclusion, this paper emphasises the critical role that classification algorithms play in precisely estimating an individual's risk of having a heart attack based on physiological parameters. The study confirms how crucial early detection is to reducing the effects of cardiovascular illnesses, which are a major cause of death worldwide. With an astounding accuracy score of 90.2%, Logistic Regression stands out as the best-performing model among the ones examined. This result indicates that it is robust in identifying patterns in physiological markers linked to the risk of a heart attack. These prediction models' effects go beyond the realm of research, proving their usefulness in actual healthcare environments. These algorithms can detect people who are in danger of having a heart attack, which could save lives by facilitating early interventions and preventive measures. As a result, the study offers practical insights with real-world consequences for patient treatment and public health, in addition to furthering our scientific understanding of heart health prediction.

6 Limitations and Considerations

Although estimating the risk of a heart attack based on physiological parameters has yielded useful insights, some limitations and constraints require careful attention. The dependability of the results could be jeopardised by probable flaws in the dataset, which is a necessity for the findings' accuracy. Additionally, the dataset's small number of records may limit its usefulness in healthcare applications; hence, a larger and more varied dataset will be required for improved generalisation. Specifically, the sex-based data skewness raises questions regarding the model's generalisability across the binary sex demographics. In addition, the research might gain from using non-psychological markers that have been shown in the literature to be heart attack risk factors [6]. Increasing the dataset's size to include a larger range of records and adding other characteristics, such as ethnicity [2] and geography [5] (country/ postcode), could improve the model's prediction power even further. However, the inclusion of such private data necessitates strict data security protocols, highlighting the privacy and ethical issues that are fundamental to healthcare data analysis. All of these drawbacks point to the necessity of expanding and improving datasets to strengthen the predictive models' inclusivity and robustness in the field of cardiovascular health.

References

[1] Support vector machine (svm) algorithm, Jun 2023.

[2] How race and ethnicity impact heart disease, n.d.

[3] British Heart Foundation. Risk factors, n.d.

[4] GeeksforGeeks. K-nearest neighbor(knn) algorithm, Nov 2023.

[5] Growth360Partners. The role of geography and race in cardiovascular disease, May 2023.

[6] Quinn Mattingly. Cardiovascular diseases, n.d.

[7] Amiya Ranjan Rout. Advantages and disadvantages of logistic regression, Jan 2023.

Appendix

Appendix 1

age - age in years

sex - sex (1 = male; 0 = female)

cp - chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 0 = asymptomatic)

trestbps - resting blood pressure (in mm Hg on admission to the hospital)

chol - serum cholesterol in mg/dl

fbs - fasting blood sugar > 120 mg/dl (1 = true; 0 = false)

restecg - resting electrocardiographic results (1 = normal; 2 = having ST-T wave abnormality; 0 = hypertrophy)

thalach - maximum heart rate achieved.

exang - exercise-induced angina (1 = yes; 0 = no)

oldpeak - ST depression induced by exercise relative to rest.

slope - the slope of the peak exercise ST segment (2 = upsloping; 1 = flat; 0 = down sloping)

ca - number of major vessels (0-3) colored by fluoroscopy.

thal - 2 = normal; 1 = fixed defect; 3 = reversible defect

output - the predicted attribute – 1 is a high risk of heart attack, 0 is low risk of heart attack