PREDICTING HEART ATTACK RISK IN INDIVIDUALS BASED ON PHYSIOLOGICAL INDICATORS

By Arif Abdulshakour Othman

Introduction

- Cardiovascular diseases continue to be a leading cause of death worldwide [1]
- Aim is to predict the risk of myocardial infarction in patients for early intervention
- Growing need for innovative and data-driven approaches in healthcare
- Using a heart attack risk dataset, can machine learning classification algorithms accurately predict heart attack risk in individuals based on physiological indicators?

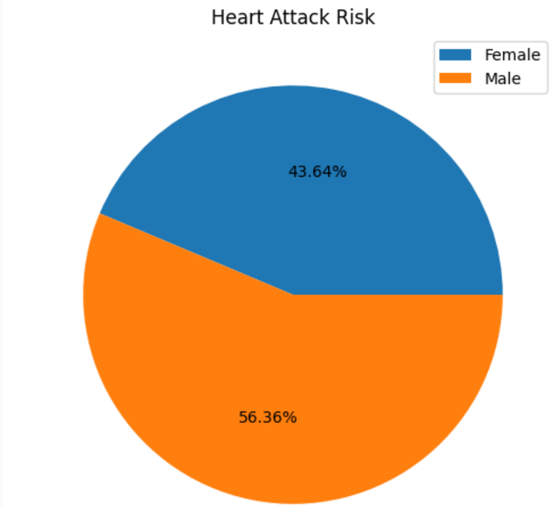
[1] Mattingly, Q. (no date) Cardiovascular diseases, World Health Organization. Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (Accessed: 01 December 2023).

Dataset Overview

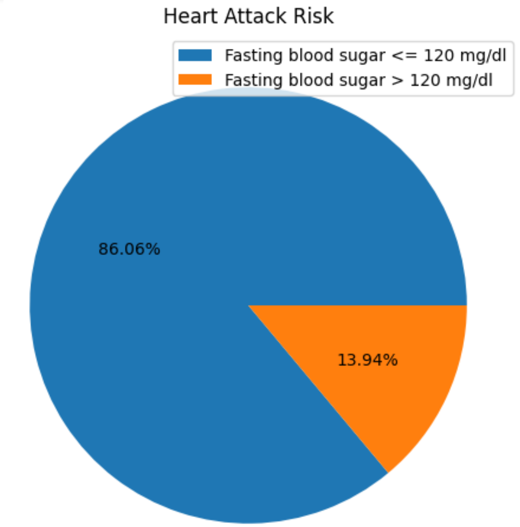
- Columns are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar > 120, resting electrocardiographic results, maximum heart rate, exercise-induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, number of major vessels, thallium stress test, target
- Physiological indicators include resting blood pressure, cholesterol, fasting blood sugar > 120, maximum heart rate, exercise-induced angina, ST depression induced by exercise
- Categorical variables are variables with discrete values
- Continuous variables are variables with continuous values

EXPLORATORY DATA ANALYSIS (EDA)

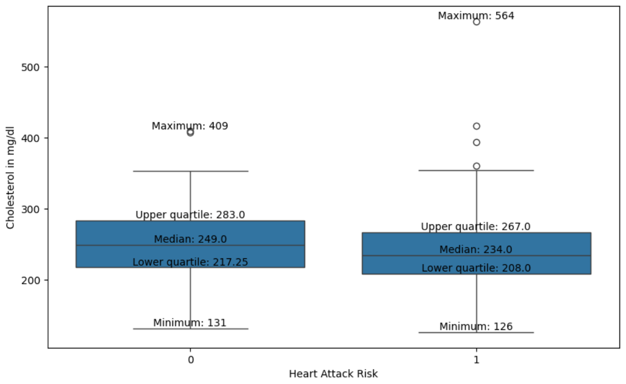
Pie charts, Box plots, pair plots and heatmap are used to gain further insight into the dataset



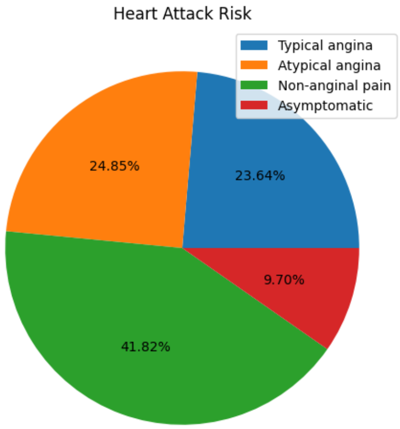
Males were at a higher risk of having a heart attack, but not significant.



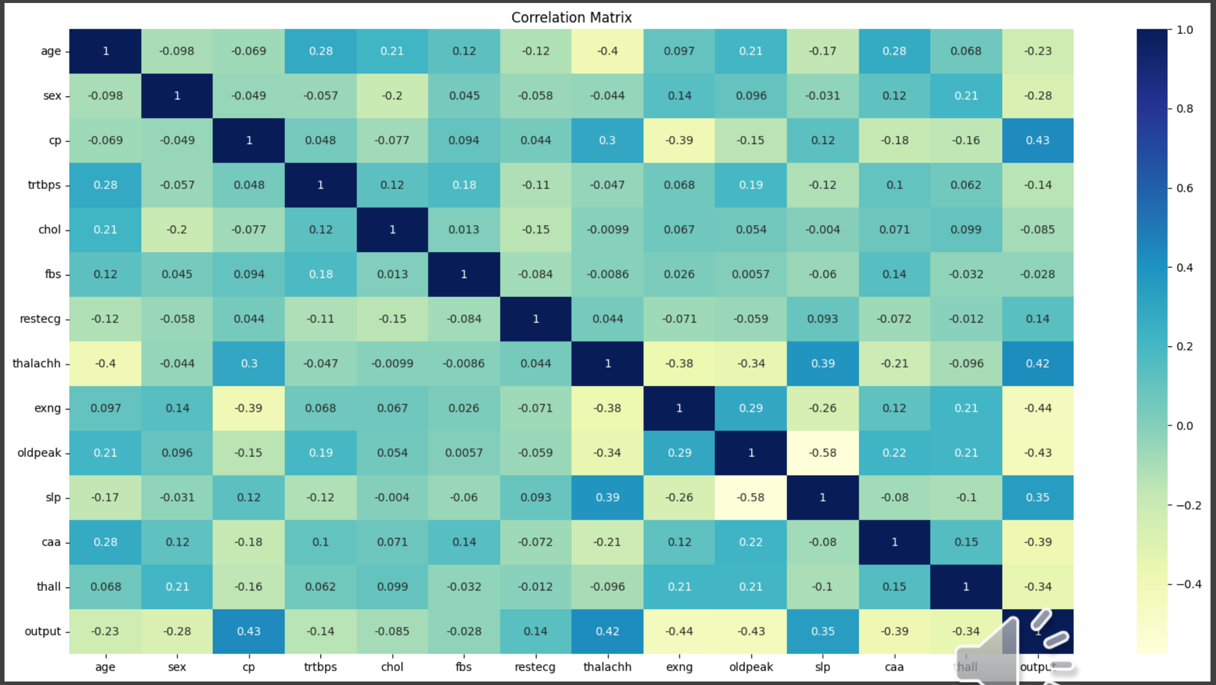
Low blood sugar adds to heart attack risk



Suggests that mildly low cholesterol is a high heart attack risk indicator



Patients with non-anginal chest pain are at a higher risk of a heart attack



Correlation Matrix

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
age	1	-0.098	-0.069	0.28	0.21	0.12	-0.12	-0.4	0.097	0.21	-0.17	0.28	0.068	-0.23
sex	-0.098	1	-0.049	-0.057	-0.2	0.045	-0.058	-0.044	0.14	0.096	-0.031	0.12	0.21	-0.28
cp	-0.069	-0.049	1	0.048	-0.077	0.094	0.044	0.3	-0.39	-0.15	0.12	-0.18	-0.16	0.43
trtbps	0.28	-0.057	0.048	1	0.12	0.18	-0.11	-0.047	0.068	0.19	-0.12	0.1	0.062	-0.14
chol	0.21	-0.2	-0.077	0.12	1	0.013	-0.15	-0.0099	0.067	0.054	-0.004	0.071	0.099	-0.085
fbs	0.12	0.045	0.094	0.18	0.013	1	-0.084	-0.0086	0.026	0.0057	-0.06	0.14	-0.032	-0.028
restecg	-0.12	-0.058	0.044	-0.11	-0.15	-0.084	1	0.044	-0.071	-0.059	0.093	-0.072	-0.012	0.14
thalachh	-0.4	-0.044	0.3	-0.047	-0.0099	-0.0086	0.044	1	-0.38	-0.34	0.39	-0.21	-0.096	0.42
exng	0.097	0.14	-0.39	0.068	0.067	0.026	-0.071	-0.38	1	0.29	-0.26	0.12	0.21	-0.44
oldpeak	0.21	0.096	-0.15	0.19	0.054	0.0057	-0.059	-0.34	0.29	1	-0.58	0.22	0.21	-0.43
slp	-0.17	-0.031	0.12	-0.12	-0.004	-0.06	0.093	0.39	-0.26	-0.58	1	-0.08	-0.1	0.35
caa	0.28	0.12	-0.18	0.1	0.071	0.14	-0.072	-0.21	0.12	0.22	-0.08	1	0.15	-0.39
thall	0.068	0.21	-0.16	0.062	0.099	-0.032	-0.012	-0.096	0.21	0.21	-0.1	0.15	1	-0.34
output	-0.23	-0.28	0.43	-0.14	-0.085	-0.028	0.14	0.42	-0.44	-0.43	0.35	-0.39	-0.34	1

cp, thalachh, slp are significantly correlated with output

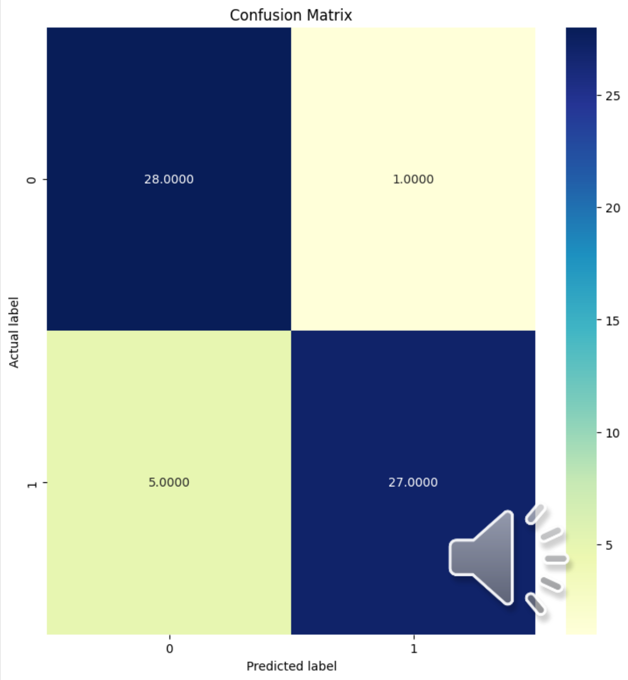
Machine Learning Overview

- Classification is a type of supervised learning that uses label data to categorise data into predefined labels
- Given the binary nature of the target (0 = less chance of a heart attack, 1 = more chance of heart attack) classification is most suitable
- Used Logistic Regression, Random Forest, Support Vector Machines (SVM), and K-Nearest Neighbours on the dataset
- All variables in the dataset are used
- Feature engineering techniques are used such as scaling and encoding
- Holdout Method using sklearn.model_selection.train_test_split with a training set of 80% with a random state of 42.
- Hyperparameter tuning is performed on all models using sklearn.model_selection.GridSearchCV is optimised by cross-validated grid search over a parameter grid

Logistic Regression [2]

- Classification algorithm in machine learning that uses one or more independent variables to determine a dichotomous (only two) possible outcomes
- Learns linear relationship from the given dataset and introduces non-linearity in the form of the sigmoid function
- C and penalty parameter a tuned (C = 10.0, penalty = 'l2'). The default solver lbfgs is used.
- Accuracy score of 90.2%

Advantages	Disadvantages
Easy to implement, interpret, very efficient to train	Only constructs linear boundary
Regularisation (L1 and L2) used to prevent overfitting	Not suited to complex relationships
Fast at classifying unknown records	Only work if the predicted variable is binary

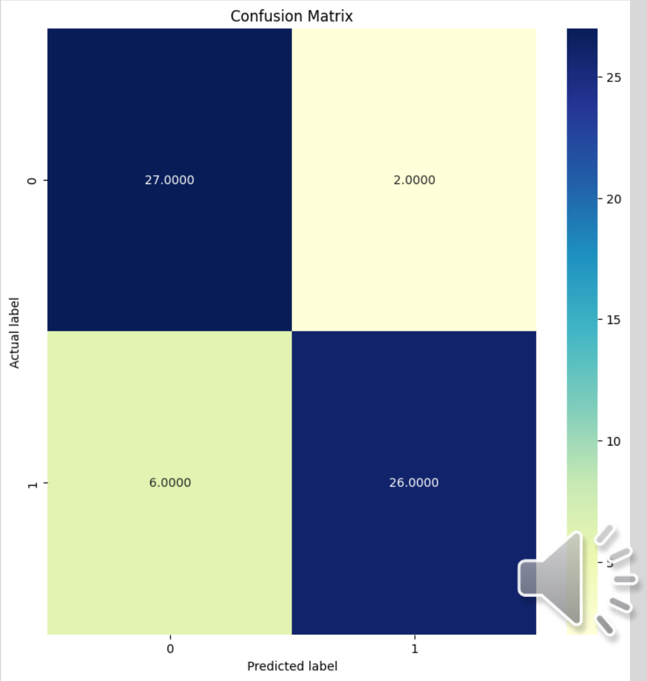


[2] Rout, A.R. (2023) Advantages and disadvantages of logistic regression, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/ (Accessed: 01 December 2023).

Support Vector Machines [3]

- Classifier that represents training data as points in space separated into categories by a gap as wide as possible
- New points are added to the space, which categorises which space they belong to
- C, gamma, and kernel are tuned (C = '10.0', gamma = 'scale', kernel = 'linear')
- Accuracy score of 86.9%

Advantages	Disadvantages
Productive in high-dimensional spaces	Not acceptable for large datasets
Regularisation can be used to prevent overfitting	Kernel matrix can be very large if dataset is large
Can model non-linear decision boundaries	Sensitive to parameters

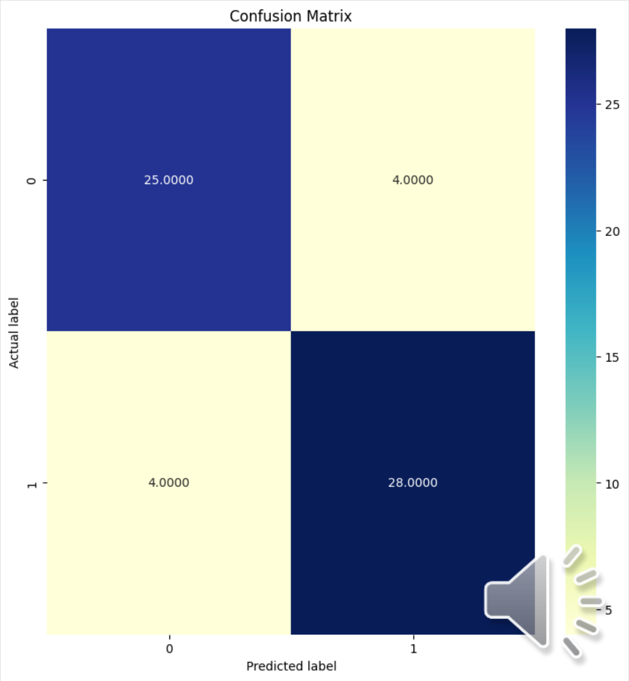


[3] Support Vector Machine (SVM) algorithm (2023) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/support-vector-machine-algorithm/ (Accessed: 06 December 2023).

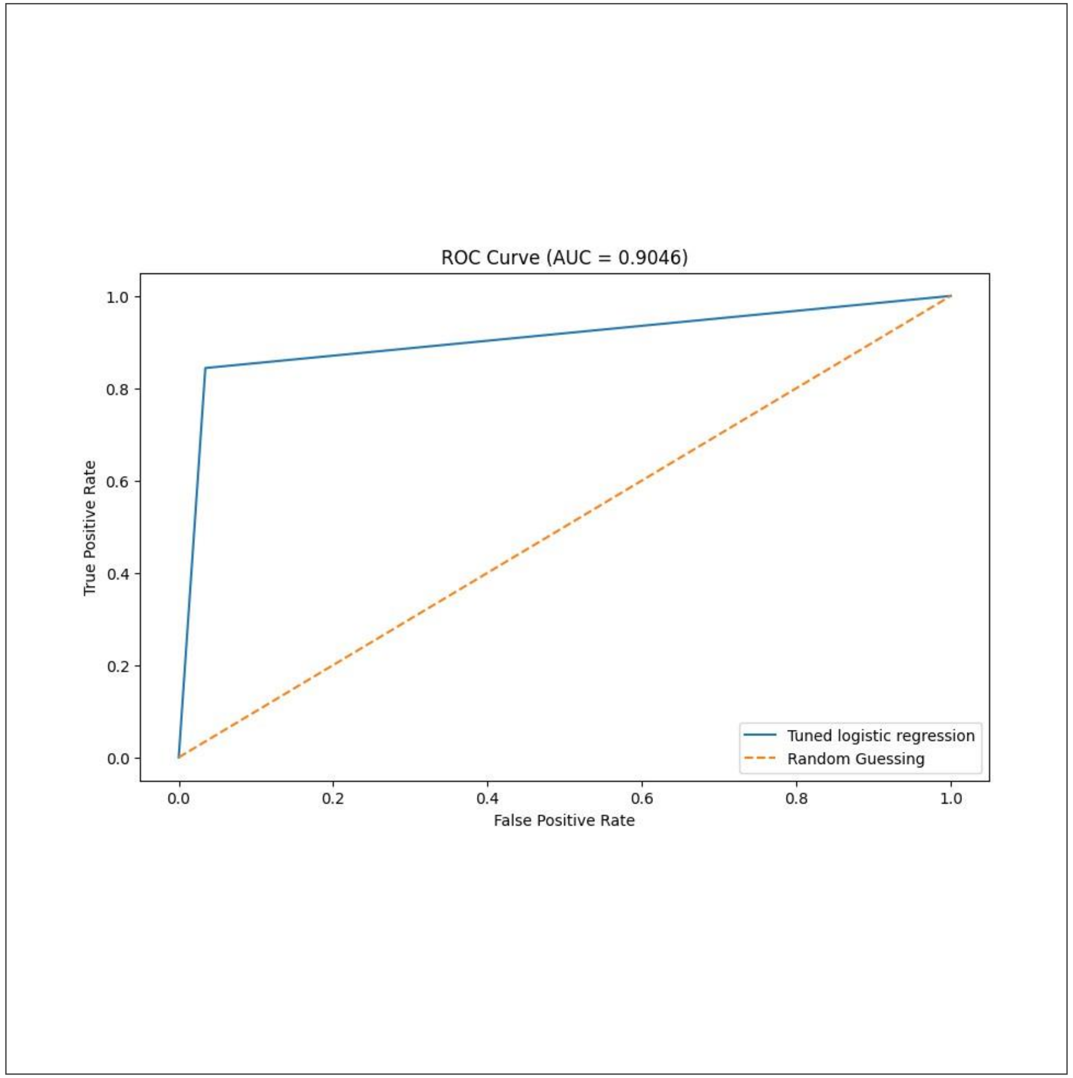
K-Nearest Neighbours [4]

- Lazy learning algorithm that stores all instances corresponding to training data in n-dimensional space. K is the number of neighbours it checks
- n_neighbors, weights, and p are all tuned (n_neighbors = 5, p = 1, weights = 'uniform')
- Accuracy score of 86.9%

Advantages	Disadvantages
Easy to implement and robust to noisy training	Computationally expensive
Adapts easily to new data points	Curse of dimensionality
Limited hyperparameters	High chance of overfitting



[4] K-Nearest Neighbor(KNN) algorithm (2023) GeeksforGeeks. Available at: https://www.geeksforgeeks.org/k-nearest-neighbours/ (Accessed: 06 December 2023).



Results

- Logistic regression performs the best with an accuracy of 90.2% but a cross-validated accuracy score of 84.6%
- An area under curve (AUC) = 0.5 is no better than random guessing
- An AUC > 0.9 is considered excellent discrimination
- AUC refers to the area under the receiver operating characteristic (ROC) curve.
- ROC curve is a graphical representation of the trade-off between true positive and false positive

Conclusions

Classification algorithms can be used to accurately predict the risk of heart attack in individuals based on physiological factors

Specifically logistic regression with an accuracy score of 90.2%

Vital in early identification of individuals at risk of heart attack

Applicable in real-world applications ultimately saving lives

Limitations and Considerations

- Dataset may be inaccurate
- Limited records in the dataset to be useful in healthcare applications
- Data is skewed based on sex
- Benefit from non-psychological indicators that are proven to be risk factors for heart attack [5]
- Benefit from more records in the dataset
- Benefit from geographical indicators (requires data security) [6]
- Benefit from including ethnic background (requires data security) [7]

[5] Risk factors (no date) British Heart Foundation. Available at: https://www.bhf.org.uk/informationsupport/risk-factors (Accessed: 01 December 2023).

[6] Growth360Partners (2023) The role of geography and race in cardiovascular disease, The Role of Geography and Race in Cardiovascular Disease. Available at: https://www.modernheartandvascular.com/the-role-of-geography-and-race-in-cardiovascular-disease/#:~:text=Research%20reveals%20that%20where%20you,a%20stroke%20or%20heart%20disease.) (Accessed: 01 December 2023).

[7] Heart disease risk: How race and ethnicity play a role (no date) How Race and Ethnicity Impact Heart Disease. Available at: https://my.clevelandclinic.org/health/articles/23051-ethnicity-and-heart-disease (Accessed: 01 December 2023).