Machine Learning for Heart Attacks

Jayden Yap Jean Hng

School of Computing

Singapore Polytechnic

Singapore

JAYDENYAP.21@ichat.sp.edu.sg

Abstract

A heart attack occurs when there is an occlusion of a coronary artery, leading to tissue damage. A heart attack is fatal when there is a subsequent disruption of the normal electrical signal of the heart, leading to fibrillation. Scientists are working on a test to determine who is at risk of heart attack or stroke. Current methods to detect heart disease include looking at stress tests, CT scans, and family history. The new blood test is still being studied to determine if it is accurate enough to roll out on a wide scale. Let's apply Machine Learning Classification to try to aid in this area.

I. Introduction

We will make use of Supervised Machine Learning to improve Heart Attack predictions. Therefore, we need to build classifier models.

II. Definitions

A. Machine Learning

Machine Learning or automatic learning is a scientific field, and more specifically a subcategory of artificial intelligence. It consists of letting algorithms discover "patterns", namely recurring patterns, in data sets. This data can be numbers, words,images or statistics. Types of Machine Learning include: Supervised Learning, Unsupervised Learning and Reinforcement Learning.

B. Supervised Learning

Supervised learning is an approach to creating artificial intelligence (AI), where a computer algorithm is trained on input data that has been labeled for a particular output. The model is trained until it can detect the underlying patterns and relationships between the input data and the output labels, enabling it to yield accurate labeling results when presented with never-before-seen data.

III. Related Works

There have been several previous studies conducted on the issue of heart attack predictions with machine learning. One such investigation was conducted this year in 2022 by several authors at the National Library of Medicine. (Javeed, A,2022). They scrutinized various ML approaches for the development of automated diagnostic systems for heart disease detection based on different kinds of modalities.

Another one was by Adil Hussain Seh in 2019. (Seh, Adil Hussain. (2019) In this paper commonly used data mining and machine learning techniques and their complexities are summarized.

IV. Methodology

A. Abbreviations and Acronyms

ML may be used to abbreviate Machine Learning

Similarly, abbreviations for some model names such as KNN for K Nearest Neighbors, SBM for Support Vector Machines may be used. These are widely used in the industry.

B. Units

We will be using SI units in our investigation unless otherwise stated such as with our data's unit measurements such as mg/dl for cholesterol values

C. Equations

One's total cholesterol score is calculated using the equation: HDL level + LDL level + 20% of your triglyceride level

$$
\begin{aligned}
precision &= \frac{TP}{TP + FP} \\
recall &= \frac{TP}{TP + FN} \\
F1 &= \frac{2 \times precision \times recall}{precision + recall} \\
accuracy &= \frac{TP + TN}{TP + FN + TN + FP} \\
specificity &= \frac{TN}{TN + FP}
\end{aligned}
$$

Figure 1: Classification metric equations

IV. Data Source

The dataset I will use for this investigation is from Kaggle, a prestigious website commonly used by Data Scientists to obtain free datasets and learn more about Machine Learning. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=download The dataset is provided by Rashik Rahman.

V. Discussion

My investigation continued with fitting models like the Random Forest classifier and K Nearest Neighbors classifier onto the data, this was after splitting my data.

From Figure 2, you can see that there was a similar amount of error between false positives and false negatives. Our dataset is relatively balanced between number of positives and negatives.

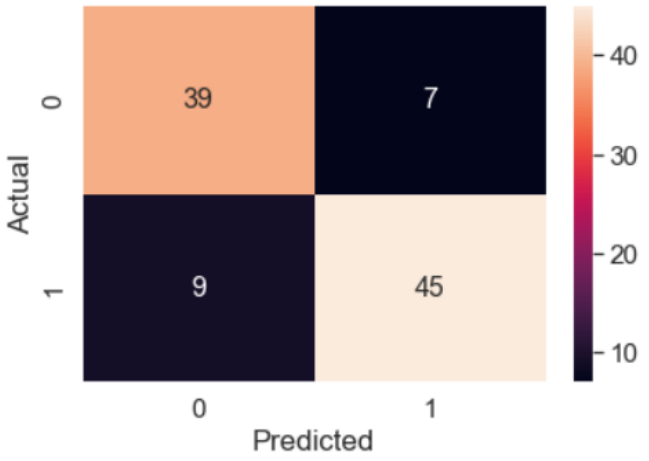

Figure 2: Confusion Matrix: Random Forest

From Figure 3 on the K Nearest Neighbors model, we can confirm that the performance was significant worse than Random Forest. Making more false negatives and false positives.

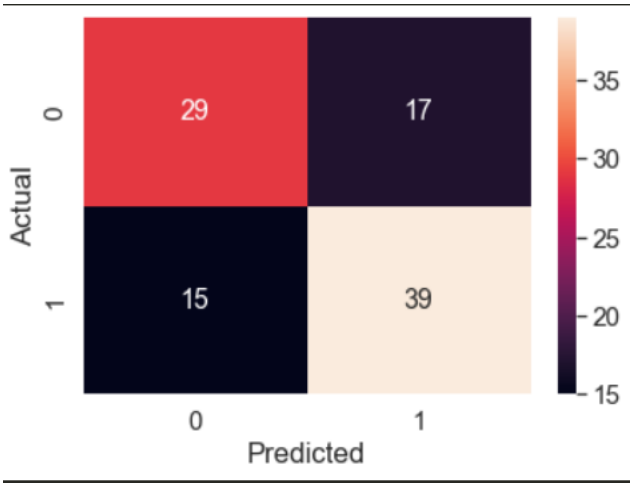

Figure 3: Confusion Matrix: K Nearest Neighbors

Heart failure has many risk factors like sex, family history, which are classified into uncontrolled reasons, however this dataset we are using contains the following factors: Age, Sex,Exercise induced angina, Number of major vessels, Chest Pain Type, Resting blood pressure, Cholestoral via BMI sensor, fasting blood sugar amount, and resting electrocardiographic results. These are the most common factors doctors look at today.Factors like high cholesterol and high blood pressure are classified into controllable risk factors. The model is possibly able to identify these 2 groups of risk factors to better determine the predictions. Figure 4 depicts the four chambers of the heart that are responsible for blood pumping.

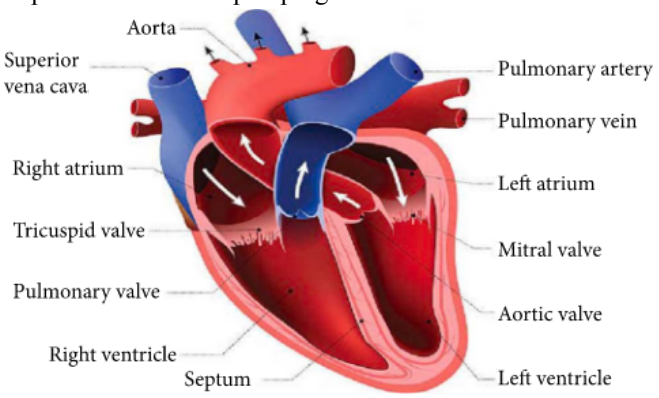

Figure 4: Anatomy of Human Heart

VI. Conclusion

We can conclude from this investigation that Random Forest is superior to the K Nearest Neighbors model for this dataset with random state 1 of both the train test splitter and the models themselves. Heart attack rate can be predicted with ~80% accuracy using the random forest model, so if we can use this in conjunction with current methods used by doctors for predictions, then we can make the world a better place. More discussion can take place for whether these can be applied in a practical fashion. But we can definitely see there is some potential in this sector.

References

[1] Javeed, A., Khan, S. U., Ali, L., Ali, S., Imrana, Y., & Rahman, A. (2022). Machine Learning-Based Automated Diagnostic Systems Developed for Heart Failure Prediction Using Different Types of Data Modalities: A Systematic Review and Future Directions. Computational and mathematical methods in medicine, 2022, 9288452. https://doi.org/10.1155/2022/9288452

[2] Seh, Adil Hussain. (2019). A Review on Heart Disease Prediction Using Machine Learning Techniques. https://www.researchgate.net/publication/337063133_A_Review_on_Heart_Disease_Prediction_Using_Machine_Learning_Techniques

[3] Rashik Rahman. (2021) Heart Attack Analysis & Prediction Dataset https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset?resource=download