AMJES | American Journal of Emerging Scholars

Prediction of heart attack risk using logistic regression and artificial neural networks

Alice Lin

ABSTRACT

Cardiovascular diseases are highly prevalent around the world, making heart attacks one of the leading causes of death. Being able to accurately predict heart attacks before they happen could help decrease the fatality rate of heart attacks, as well as decrease the number of heart attacks that occur. Because of the limitations of human time and focus, machine learning has been utilized to attempt to find new ways of prediction. This study uses two machine learning techniques, logistic regression and an artificial neural network, in order to build a model to predict the probability someone has of experiencing a heart attack. The dataset contained 13 different features, each a different piece of medical information about a certain patient. These features were then used to predict the probability of the patient getting a heart attack. In addition to building a model for prediction, this study also attempts to evaluate the different features of the dataset and compare the importance of each feature to the overall prediction. The correlation between each feature and the overall prediction was examined. Moreover, each feature was also compared with the rest to find the difference in influence the features had on the overall prediction of the model. The results of this study show that the logistic regression model and artificial neural network model share a similar accuracy, with both models reaching an overall accuracy of 84%.

1. INTRODUCTION

Cardiovascular diseases are the cause of about 17.9 million deaths per year, with more than 4/5 of these deaths being caused by heart attacks or strokes. Because of this, medical practitioners have highlighted the importance of predicting heart attacks, so that early treatment is possible. Early treatment can help to reduce the chance of mortality in heart attacks, as well as maybe even prevent heart attacks before they happen. However, recent studies have revealed that 4 out of 5 of the widely used clinical heart attack predictors are not very accurate at predicting, and considerably overestimate the risk of heart attacks. These high overestimates can cause people who don’t need treatment to unnecessarily be prescribed costly treatments, and can also cause people to mistrust these predictions. Personal doctors may be a more reliable source of prediction, but these doctors would usually be distracted and may not pick up on the chance of a heart attack because they are not solely focused on predicting this. Furthermore, people may want to be able to assess their risk without having to go into a medical clinic or travel to see a doctor, as this may be too time consuming.

Machine learning has been used to help make predictions and medical diagnoses in the past. In a paper written by Harshit Jindal et al., three algorithms were explored: logistic regression, k-nearest neighbors (KNN), and a random forest classifier. After testing, it was decided that the KNN network was the most efficient, with an accuracy of 88.52%.

Furthermore, an exploration of the different machine learning techniques and feature selection methods was executed by Hidayet Takci. His paper explored and compared 12 different machine learning classifiers and 4 different feature selection methods. The final conclusion was that the best machine learning technique was the support vector algorithm with the linear kennel, and the best feature selection method was the reliefF method. The accuracy for the combination of these two was found to be 84.81%.

Another paper that explored the prediction of heart attacks using different techniques written by Lubna Riyaz et al. compared a variety of different machine learning methods, including support vector machine (SVM), decision tree (DT), Naive Bayes (NB), KNN, and artificial neural network (ANN). Out of all of the different methods, the one with the highest average prediction accuracy turned out to be ANN, with an accuracy of 86.91%. The lowest prediction accuracy came from the C4.5 decision tree technique with an accuracy of 74.0%.

This paper uses two machine learning techniques: an artificial neural network and logistic regression to attempt to build a model to predict whether someone is likely to have a heart attack. Furthermore, it also looks to distinguish which of the features provided has the largest influence on whether a heart attack is likely.

2. METHODS

2.1: DATASET

The dataset used in this project, “Heart Attack Analysis and Prediction Dataset”, contained 13 different features which were used to predict one binary output, with 0 representing a low chance of having a heart attack and 1 representing a high chance of having a heart attack. The dataset contained a total of 303 different patients. 165 of the patients were classified as having a high risk of heart attack, and 138 of the patients were classified as not having a high risk of a heart attack. The 13 different features were different medical statistics for a patient, which were as follows:

Age: The age of the patient in years. Upon analyzing the dataset, it was found that the range of ages was all adults, with a minimum of 29 and a maximum of 77. The average age of all the patients was 54.366 years. The standard deviation was 9.055 years.

Sex: The gender of the patient. A 0 was used to denote females, and a 1 was used to show the patient was male. There were a total of 96 females in the dataset, and 207 males.

Cp: The type of chest pain the patient experienced. The pain was categorized into one of three types: typical angina (1), atypical angina (2), non-anginal pain (3), and asymptomatic (0). A total of 143 patients were asymptomatic, 50 exhibited typical angina, 87 exhibited atypical angina, and 23 exhibited non-anginal pain.

Trtbps: The resting blood pressure of the patient (mg/dl). Upon analyzing the dataset, it was found that the range went from 94 mg/dl to 200 mg/dl. The average resting blood pressure for all of the patients was 131.624 mg/dl. The standard deviation was 17.521 mg/dl.

Chol: The cholesterol level of the patient (mg/dl) obtained using a BMI sensor. The average cholesterol level for all of the patients was 246.264 mg/dl. The standard deviation was 51.827 mg/dl.

Fbs: Whether the fasting blood sugar level of the patient was >120 mg/dl. A 1 was used to represent if this was true, and a 0 was used if it was false. 45 patients were found to have a fasting blood sugar level >120 mg/dl, and 258 patients were found to have a fasting blood sugar level <120 mg/dl.

Restecg: The resting electrocardiographic results, categorized into three groups: normal (0), having ST-T wave abnormalities (1), and showing probable or definite left ventricular hypertrophy according to Estes’ criteria (2). 147 of the patients were found to have normal electrocardiographic results, 152 patients were found to have ST-T wave abnormalities, and 4 patients were found to have probable or definite left ventricular hypertrophy according to Estes’ criteria.

Thalachh: The maximum heart rate of the patient that was achieved. Upon analyzing the dataset, it was found that the range was from 71 to 202. The average maximum heart rate for all of the patients was 149.647 bpm. The standard deviation was 22.891 bpm.

Exng: Whether the patient experienced exercise induced angina. A 1 represented that the patient did experience it, and a 0 represented that they did not. 204 of the patients did not experience exercise induced angina, and 99 did experience exercise induced angina.

Oldpeak: The previous peak of the patient as seen from an ECG plot. The average value of the previous peak on the ECG for all of the patients was 1.040. The standard deviation was 1.161.

Slp: The slope as taken from the ECG plot. The feature was categorized into three categories: downsloping (0), flat (1), and upsloping (2). 21 of the patients were found to have a downsloping slope, 140 patients were found to have a flat slope, and 142 were found to have an upsloping slope.

Caa: The number of major vessels of the patient colored by fluoroscopy. The average number of vessels colored for all of the patients was 0.729. The standard deviation was 1.0.

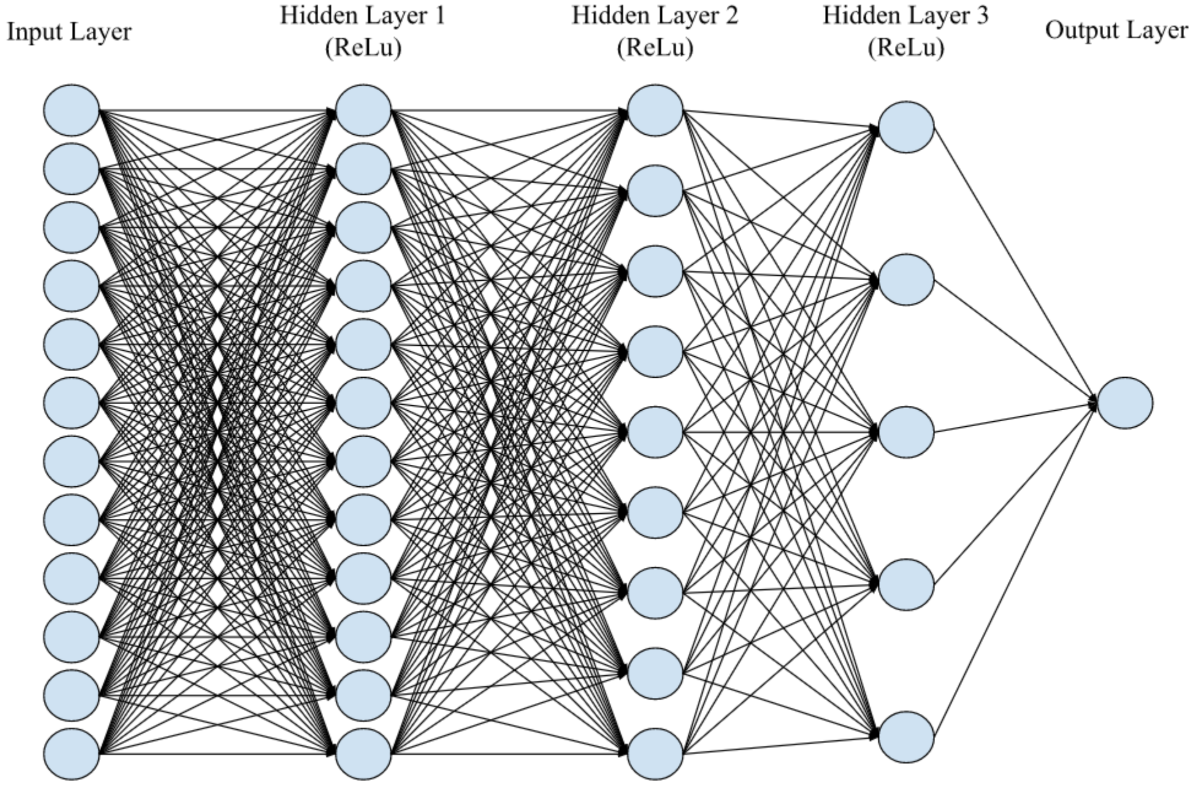
Thall: The observance of a blood disorder called thalassemia. The data was sorted into three categories: fixed defect (1), normal blood flow (2), and reversible defect (3). 18 patients were found to have a fixed defect, 166 to have a normal blood flow, and 117 to have a reversible defect.

2.2: MODEL BUILDING/TRAINING

The programming language used was Python. First the data was split into training and testing, with 25% of the data being designated as the testing set. The features were then scaled down to make the models more efficient. The scaling operation used was the Standard Scaler operation from the sklearn.preprocessing package.

The first machine learning technique used for prediction was a simple logistic regression model, to create a baseline accuracy to compare other techniques to. To see if a more accurate result could be obtained with a more complicated model, an artificial neural network was built and then trained with the features. Each layer of the neural network was densely connected, and after some experimentation the ending model had three hidden layers, with 13, 9, and 5 nodes respectively (Figure 1). The activation function of each hidden layer was the ReLu function. In addition, hyperparameter tuning was also implemented with the regularization parameter lambda. A starting value of 0.1 was used for lambda, and the value was decreased by 0.01 until it was seen that the accuracy was no longer increasing with the decrease of lambda. After a couple of trial runs, the final value of 0.04 was selected for lambda. Following all of the hidden layers, the output layer consisted of one node, and the sigmoid activation function to produce the final value of 0 or 1. The model was then compiled using the root mean square propagation (RMSprop) and the error calculated was the binary cross entropy error. The final model of the neural network can be visualized as seen below:

Figure 1. Visualization of the artificial neural network trained using the dataset to predict whether a heart attack was likely to occur.



The model was then run with 1000 epochs, and the loss function (Figure 2) was graphed for each step to ensure the model was learning correctly. The loss function used an optimizer of RMSprop and the binary cross entropy loss function.

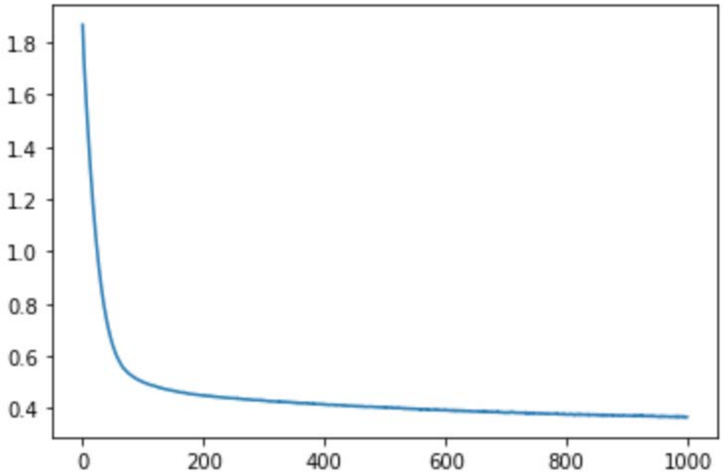


Figure 2. Graph of the loss function with the loss on the y-axis and the epoch number on the x-axis. The decreasing trend shows that the model was learning and the flattening at the end reveals that enough epochs were used to ensure that the model had finished learning.

2.3: FEATURE IMPORTANCE

In order to determine which feature held the most influence over the final output, feature importance was investigated. In order to do this, each feature was examined independently. A new set of data was created and analyzed for each feature. For the feature being examined, the values were replaced by values of equal increment ranging from the minimum to the maximum value of the feature. The rest of the features were kept constant at the median value of the respective feature. Each of these new datasets were then independently inserted into the model for predictions.

3. RESULTS

3.1: MODEL RESULTS

Classification reports were printed for both models used (Table 1, Table 2). In these reports, precision refers to the ratio of correctly predicted positives to the total positive predictions. Recall refers to the ratio of correctly predicted positives to the total actual positives. The f1 score refers to a weighted harmonic mean of the precision and recall. Support refers to the total number of each class in the dataset. The equations for each of the interpreters can be seen below:

$$\text{precision} = \text{true positives} / \text{predicted positives}$$
$$\text{recall} = \text{true positives} / \text{actual positives}$$
$$\text{f1 score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

Table 1. Classification report for the logistic regression model.

	precision	recall	f1-score	support
0	0.89	0.73	0.80	33
1	0.82	0.93	0.87	43
accuracy			0.84	76

Table 2. Classification report for the artificial neural network.

	precision	recall	f1-score	support
0	0.86	0.76	0.81	33
1	0.83	0.91	0.87	43
accuracy			0.84	76

3.2: FEATURE IMPORTANCE RESULTS

To interpret the results of the feature importance test, first the predictions for each of the features was graphed, in order to see the correlation between each individual feature and the overall prediction, or likelihood of a heart attack (Figure 3). Furthermore, a graph was also generated to compare the magnitudes of influence between all of the different features (Figure 4).

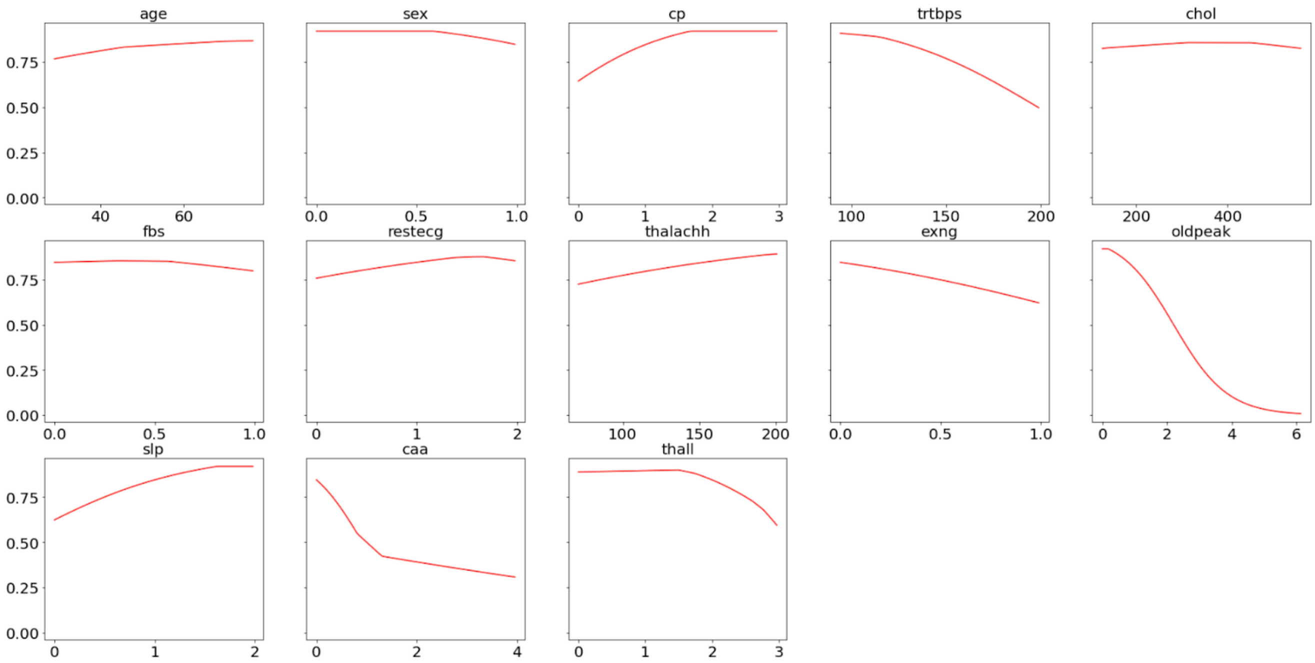


Figure 3. A table of graphs of each of the features and the results of the feature importance test. The y-axis for each of the graphs are the predictions, and the x-axis represents the values of the feature. Each of the graphs has been scaled to the same y-axis. The larger the change in prediction is, the more influential the feature is on the final output.

Afterwards, the "value" of the importance of each feature was calculated by subtracting the maximum predicted value by the minimum predicted value for each feature (Figure 4).

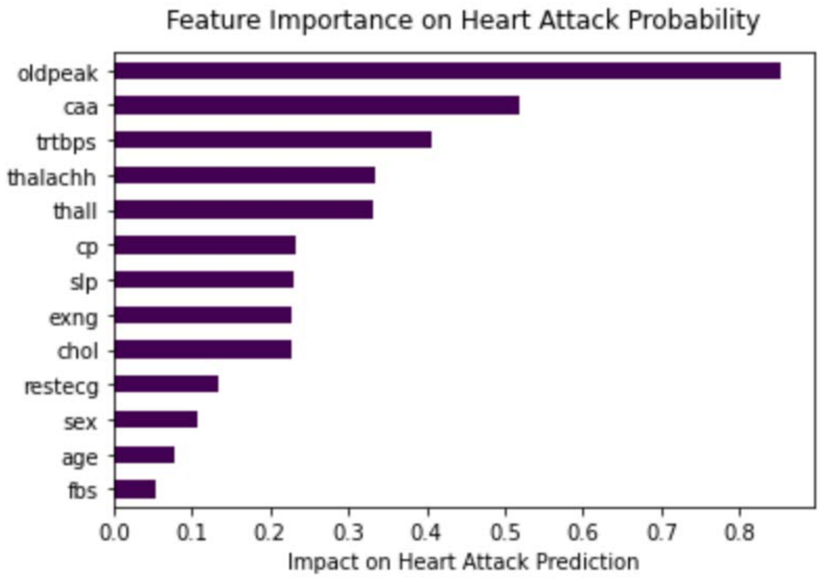


Figure 4. Graph of the importance of each feature on predicting the final output. The top, with the largest difference, represents the feature with the largest impact, and the bottom shows the feature with the lowest impact.

4. DISCUSSION

As seen in the results shown above, the logistic regression model and the artificial neural network reached a similar accuracy of 84%. This shows that a more complicated model does not necessarily bring better accuracy for predicting. The final accuracy is comparatively similar to previous accuracies obtained in other studies, such as the support vector algorithm with linear kennel which obtained an accuracy of 84.81%, and another artificial neural network which obtained an accuracy of 86.91%.

However, one study was able to achieve more accurate results, with an accuracy of 88.52% using a KNN network. Upon comparing the datasets, it was noted that the same features existed in both of the datasets. However, there were two main differences between that study and this one. The dataset used in that study contained more information in its features despite the same feature names, which could result in higher prediction accuracy. For example, the feature restecg (the resting electrocardiographic results) had only three categories in this study, while it had five categories in the other study. The other main difference between the two studies was the model used for prediction. For future studies, an area of exploration could be how to categorize the dataset to maximize accuracy.

As seen from the feature importance analysis, the old peak (previous peak of the patient as seen from an ECG plot) was the most influential feature and the fasting blood sugar level was the least important. This result is slightly surprising because higher cholesterol and blood sugar levels are often considered potential causes for heart attacks. It is possible that higher importance would be assigned if the feature was not categorical, but was numerical instead. Having age and sex near the bottom of the list was expected, because it is often noted that sex has little effect on one’s chance of a heart attack, and the range of ages wasn’t quite large enough to show the effects old age has on the likelihood of heart attacks.

This study has some limitations due to the limited sample size from the dataset. Therefore in future studies a larger dataset with a wider range of values could be analyzed to find a more generalized result. Furthermore, a study could analyze whether omitting certain features or adding other features could improve the accuracy of the model since it could improve signal-to-noise ratio.

This paper marks a step towards the improvement of heart attack prediction in hopes that in the future people will be able to be warned accurately when they are at risk of a heart attack without having to notice it themselves and seek out a professional.

5. REFERENCES

Cleveland Clinic. (n.d.). *Heart attack (myocardial infarction)*. Cleveland Clinic. Retrieved February 10, 2023, from https://my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction

Harshit Jindal *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* 1022 012072

John Hopkins Medicine. (2015, February 16). *Most Clinical 'Calculators' Over-Estimate Heart Attack Risk*. John Hopkins Medicine. Retrieved February 7, 2023, from https://www.hopkinsmedicine.org/news/media/releases/most_clinical_calculators_over_estimate_heart_attack_risk

Khan, T. (n.d.). *Cardiovascular diseases*. World Health Organization. Retrieved February 6, 2023, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Rahman, R. (2021). *Heart attack analysis & prediction dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Riyaz, L., Butt, M. A., Zaman, M., & Ayob, O. (2021, August 29). *Heart disease prediction using machine learning techniques: A quantitative review*. SpringerLink. Retrieved February 9, 2023, from https://link.springer.com/chapter/10.1007/978-981-16-3071-2_8

TAKCI, HİDAYET (2018) "Improvement of heart attack prediction by the feature selection methods," *Turkish Journal of Electrical Engineering and Computer Sciences*: Vol. 26: No. 1, Article 1. https://doi.org/10.3906/elk-1611-235