**Using Classification Algorithms to Predict if a Patient is Susceptible to a Heart Attack**

This report will contain various algorithms which will determine whether a person has a heart disease based on certain, laid out variables. The accuracy and precision of these algorithms will be determined and then compared as well as graphed to evaluate which algorithm works best for this dataset. Performance of the algorithms will be measured using the cross-fold validation method, in the case of this dataset, there will be a comparison between patients, determined using classification algorithms, to which the cross-fold validation method will compare each machine learning algorithm to see which one performs the best. This will be done by training and testing each algorithm, comparing methods to observe performance (Fushiki, 2011). Due to the dataset not being large, a 5-fold cross validation will be carried out. Another method which will be carried out using a precision metric and a confusion matrix, this will help output the number of correct positive predictions made which is done by the precision metric, whereas the confusion matrix provides a summary of prediction results. Combining all these methods will decide which algorithm has the highest accuracy and precision in deciding if a patient is susceptible to a heart attack or not.

**The Data**

The data within the .csv file is all numerical with various variables which influence whether the patient is inclined to get a heart attack which is represented with 1, higher chance of heart attack, and 0, for lower chance of heart attack. The target variable is represented with a binary. Since there are no outliers or missing values as shown in figure 1, the data can be used for our algorithms.

Table

Description automatically generated

Figure 1 Displaying no Missing Values

Since the K-nearest algorithm will be used it is important to scale the data, certain magnitudes of variables may be much greater than that of another variable which will affect the distance between the two points, thus impacting the performance of the KNN model by it being biased towards the variable with the greater magnitude (Pultik Sharma, 2019). To solve this problem, we can bring down the variables to the same scale as shown in figure 2. Since all variables are integers, we do not have to transform variables from objects.

Text

Description automatically generated with medium confidence

Figure 2 Showing Code to Scale the Data

**Presenting the Models**

The following algorithm models have been used to test the dataset: Random Forest, Logistic Regression, K-nearest Neighbour, and Decision Tree. In each of these models, the accuracy and cross-fold validation has been calculated along with the accuracy of these folds and the standard deviation, as well as that the confusion matrix had been carried out. An analysis of these solutions will be proposed to see which algorithm has the best and highest performance and which has the most successful predictions on whether the patient will have a heart attack or not. The accuracy is simply the number of predictions that the model got right compared to the total number of predictions, the higher the accuracy the more precise the model is in predicting the chances of a heart attack. The model test accuracy is the total of true positives and negatives compared to the entirety of false and true positives. This gives an overall accuracy of how well our model does. The confusion matrix provides a table, which highlights the true positives, true negatives, false positives, and false negatives. These results show the type of errors that occur within the dataset in the specified algorithm. A showcase of each result for each algorithm will be displayed as a table and then later compared to each other.

The first algorithm that has been tested on the provided dataset is the random forest algorithm. The way this algorithm works is by taking k number of records from the dataset to which individual decision trees are created per sample, the output is based on majority voting of each sample. The following table displays all accuracy results from training the data, cross validation, and confusion matrix:

Table 1 Showing Results Obtained From Random Forest Algorithm

|  |  |  |
| --- | --- | --- |
| Train/Test Accuracy | Model Test Accuracy | Cross-fold Validation Accuracy |
| 0.98760 | 0.8196 | 0.80 (standard deviation of 0.08) |

The second algorithm that was used was the logistic regression algorithm, this algorithm evaluates the relationship between variables via estimating probabilities using its own logistic function, the probabilities are then transformed into binary values via the sigmoid function (Wright, 1995). Interesting results have been found in terms of the accuracy of this model, using the MinMax scaler provided better performance than using the standard scaler, as shown:

Table

Description automatically generated

Figure 3 Showing Results from using Standard Scaler

A screenshot of a computer

Description automatically generated with medium confidence

Thus, this algorithm has shown the following results:

Figure 4 Showing Results using MinMax Scaler

Table 2 Showing Results from Logistic Regression Algorithm

|  |  |  |
| --- | --- | --- |
| Train/Test Accuracy | Model Test Accuracy | Cross-fold Validation Accuracy |
| 0.89 | 0.88 | 0.83 (standard deviation of 0.04) |

The third algorithm used was K-nearest Neighbour which works by finding the distance between the data to be queried on and the examples in the data, the number of neighbours is specified, then votes based on the frequency of the labels around it, thus classifying the new data point based on similarity. At first, just by assigning the algorithm certain ‘k’ neighbours, the accuracy score turned out to be very low compared to other algorithms, however it is possible to find the best k neighbour to improve this accuracy by using iteration, looping through 1 and 20 neighbours, and finding the best one, in this case k = 11 was found to provide the maximum score as seen in figure 5

Chart, line chart

Description automatically generated

Therefore, giving the following results:

Table 3 Showing Results Obtained from K-nearest Algorithm

|  |  |  |
| --- | --- | --- |
| Train/Test Accuracy | Model Test Accuracy | Cross-fold Validation Accuracy |
| 0.75 | 0.65 | 0.80 (with standard deviation of 0.08) |

Finally, the fourth algorithm used was the decision tree algorithm which provides a similar concept to the random forest algorithm. Each internal node is compared with the record’s attribute and follows the branch until a leaf node is reached with a predicted class value. Table 4 shows the following results:

|  |  |  |
| --- | --- | --- |
| Train/Test Accuracy | Model Test Accuracy | Cross-fold Validation Accuracy |
| 0.83 | 0.83 | 0.77 (with standard deviation of 0.05) |

**Conclusion and Comparisons**

Here we are comparing the performance of all models used. Starting with the accuracy of each model using a bar graph:

As shown, random forest has the best accuracy score. This means that it is the algorithm that most correctly predicts the chances of a person having a heart attack. This is most ideal as in the real world, predictions made on patients must be as accurate as possible to stop a potential fatality from happening.

The algorithm with the best cross-fold validation was logistic regression followed by random forest. Thus, the logistic regression algorithm has the best ability and most effective when predicting new data that was not used in estimating it.

Overall, the best algorithm for this dataset is the random forest, as it provides on average the best performance due to its high accuracy score and high cross fold validation score

FUSHIKI, T. 2011. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing,* 21**,** 137-146.

WRIGHT, R. E. 1995. Logistic regression.

Pultik, S. 2019. Why is Scaling Required in KNN and K-Means? Date Accessed (02/02/2022). Available at <https://medium.com/analytics-vidhya/why-is-scaling-required-in-knn-and-k-means-8129e4d88ed7>

Dataset Reference: <https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>