University of Science and Technology at Zewail City

Communication and Information Engineering

CIE 417
Machine Learning

Heart Disease Classification

Team Information	*ID*
Asmaa Mohamed Ibrahim	201701056
El-Sayed Mohammed Mostafa	201700038
Fatma Moanes NourEl-Din	201700346
Muhammad Magdy Alasmar	201700316

Course Instructor
Dr. Mohamed ElShenawy

Contents

1 Problem Definition and Motivation

In the course of time, more attention is engaged to utilizing machine learning models in applications related to healthcare. In fact, Artificial Intelligence seems to be an essential tool in the medical industry as it offers solutions to problems such as insufficient numbers of doctors, absence of proper healthcare systems in several countries, and wrong or late diagnosis of patients. Cardiology, which is the field of medicine concerned about the study and treatment of heart diseases, is no exception to this. Heart diseases -also called cardiovascular diseases- are diseases that involve heart or blood vessels. They are the top cause of deaths in high-income countries, and one of the major three causes of deaths in low-income countries [1]. Late diagnosis of heart diseases is very serious, and it happens a lot. It is very common that one of the most occurring heart diseases -high blood pressure- is called "the silent killer". Easy diagnosis of heart diseases is a priority that might be achieved with machine learning algorithms. In our project, dataset from UCI Machine Learning Repository about heart disease is used [2].

The given dataset to used to find the best model (and set of information/features) to accurately predict the presence or absence of heart disease in an individual given some symptoms and laboratory findings. The main objective is to have a machine learning model that is able to accurately classify whether the patient has heart disease or not. Our accuracy estimation will be the test error which is needed to be as low as possible. To achieve the lowest possible test error, different models will be applied and compared to choose the best model.

2 Literature Review

The chosen heart attack dataset is one of the most popular data sets in research papers and machine learning platforms such as Kaggle. Thus, a lot of notebooks and papers have been interested in predicting the heart attack possibility using this dataset. In a research paper by D. Shah, S. Patel and S. K. Bharti, various classification algorithms were applied on the same heart attack dataset [3]. To illustrate, the applied algorithms were KNN classifier, naïve bayes, decision tree and random forest, etc. Moreover, according to the applied pre-processing and models' evaluation, the highest

accuracy score obtained on the test data was 90.789% using KNN classifier with k = 7. Within this evaluation, there was no specific interest or focus on the recall score or any other metric scores.

Another study on the same dataset was performed by V. Chaurasia and S. Pal [4]. Within their study, a decision tree model with bagging was found to get the highest test accuracy 85.03% and the highest recall value too compared to the other models.

Another research paper presented by K. Deepika and Dr. S. Seema accomplished one of the highest test accuracies obtained on the same dataset in their research [5]. The patients' data were trained by Naïve Bayes, SVM, Decision Tree and Artificial Neural Networks (ANN). Moreover, by achieving the highest f1-score, in addition to getting the optimum recall value which is 1, the SVM classifier announced itself as a very good classifier in this dataset with f1-score 0.930. However, no further information about the preprocessing or model specifications, such as whether soft margin or hard margin SVM classifier was used, were presented in Deepika's and Seema's research paper.

In his research paper, A.K. Dwivedi applied different classification models with cross validation on the UCI dataset [6]. Furthermore, among the trained classifiers, logistic regression stood out as the highest accuracy obtained with 85% with sensitivity and specificity of 89% and 81% respectively. Here, sensitivity and recall are the same metric (True Positive / (True Positive + False Negative)). In fact, the only model enhancement done in this research was 10 fold cross validation. However, higher specificity was obtained using SVM classifier with 0.89, the logistic regression model achieved the highest overall accuracy of 85

In a machine learning notebook on Kaggle, a logistic regression model was applied to the dataset with the help of using p-value estimation to exclude the insignificant features ('age' and 'pfs')[7]. Accordingly, the resultant accuracy was 85% using the confusion matrix. Also, another logistic regression model was applied in notebook [8] , the AUC of the ROC curve was 0.91 which is relatively high; however, there was no preference for recall over precision or vice versa. Besides logistic regression, a KNN model was trained and tested on the same dataset. As a result, the f1-score was 0.82 which is not satisfactory even after performing hyper-parameter tuning for K to choose K with a value 45. Although a deep learning model, a decision tree model were also applied and trained on this dataset, the best recall result was obtained

with SVM model with linear kernel; the recall reached 0.94 with relatively lower precision and f1-score 0.86. In an additional Kaggle notebook, a good recall and f1-score results were obtained using different machine learning models; however, there were almost no pre-processing procedures performed [9]. In this case, a lower credibility is given for such model development and models evaluation.

3 Approach and Methodology

Our approach follows the standard methodology in any machine learning problem. First, the dataset exploration was performed to investigate any drawbacks in the data, such as missing values, in addition to investigating some of the relationships among the data features. Specifically, the pre-processing steps were applied to make the data records clean and ready for the fitting algorithms. After pre-processing, a feature selection method was applied to obtain better representation of the data besides achieving the target of dimensionality reduction.Then, multiple classification models were applied to the data records one at a time. All model development processes were done in parallel. Each model has its own parameters/hyper-parameters that have to be tuned and set. For models evaluation, the f1-score metric is suitable for classification problems hence it is used as the evaluation and the comparing metric. Due to the nature of the dataset, and the application in which these models can be integrated, the recall metric is much more important than the other metrics; it's more important to catch all the positive cases from the dataset with more tolerance of assigning negative cases to the positive ones. In other words, it's more acceptable to predict a positive case wrongly than ignoring a true positive case as it's a heart attack dataset and this later false classification may lead to death.

3.1 Data Exploration Pre-processing

Through the dataset description, the meaning and interpretation of each attribute is stated and matched with the corresponding values of each attribute; however, some of the features values were not within the right range of values. The attribute named 'trestbps' (Resting blood pressure) has one value although it should have 2 values (Systolic BP and Diastolic BP). For Serum Cholesterol normal levels range up to 200 and it can "rarely" increase

up to 1000 [10], however; in this dataset 'chol' attribute ranges from 126 to 564. Apparently, the most obvious drawback in the dataset is the names given to each column, hence a renaming step is applied to give the columns much more expressive names. There are no missing values but one duplicate data record is discovered and removed.

To investigate whether the dataset is balanced or not, the number of positive and negative samples is plotted as in figure 1. The figure reveals that the dataset is balanced with respect to the number of positive and negative samples.

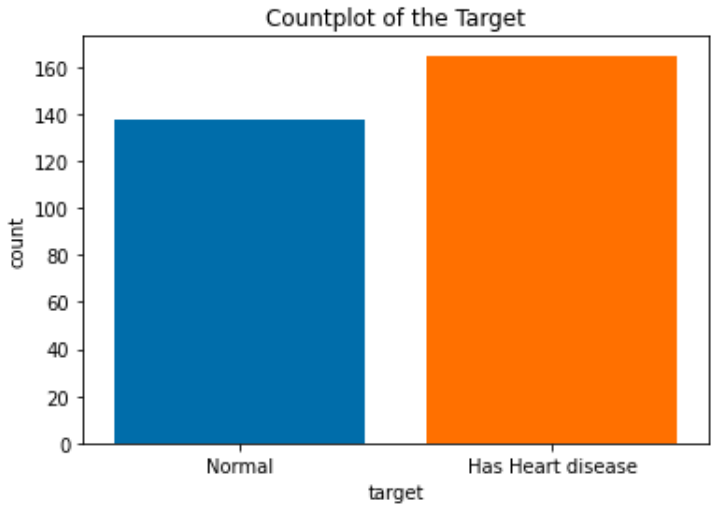

Figure 1: Countplot of the target showing whether the dataset is balanced or not

Another important aspect in investigating the dataset is the bias and fairness investigation. From figure 2 the number of males is higher than the females by an unfair amount. Hence, the data has a bias towards giving more credible results for males than females.

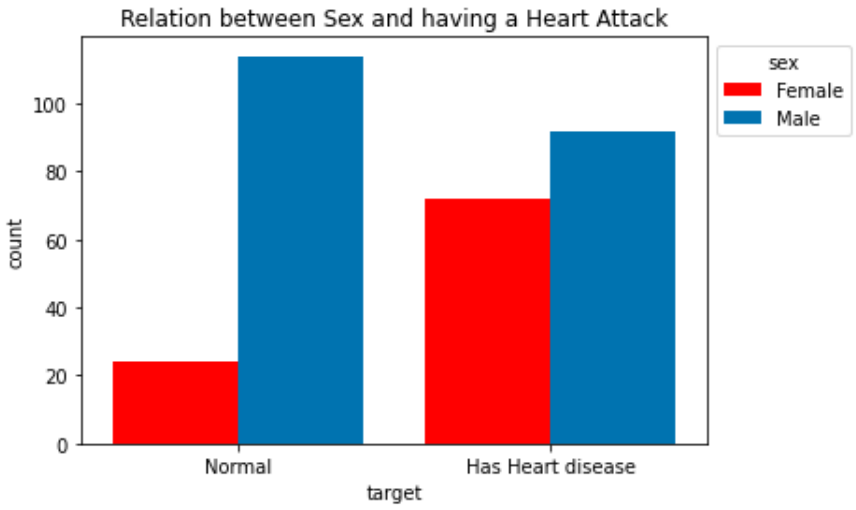

Figure 2: Countplot showing the relation between the sex and having heart attack

To move on to models development, the dataset is splitted into training set and test set where the chosen test set ratio of the dataset is 10% due to the few number of records existing in the dataset.

3.1.1 Feature Selection

Furthermore, feature selection is performed in order to get fewer numbers of attributes that act like data features instead of using the whole 13 attributes of the dataset. The strong motivation behind using feature selection is to achieve the two important trade-offs: dimensionality reduction and better performance. Different methods of feature selection have been studied in machine learning literature such as principal component analysis (PCA). The idea in PCA is to extract a new representation for the features depending on the variance measure; the newly created axes "features" are linear combinations of the original features however each new feature axis chosen by an optimization problem in order to explain as maximum variance as

possible. To obtain 100% of the variance in the dataset, the whole created principle components (13 as the number of features) have to be used as the new features which contradicts with the dimensionality reduction target thus a subset of these components are chose to explain only 85% of the variance in the dataset. From the resultant curve of the accumulative explained variance versus the number of principal components, the number of principal components chosen is only five. The first five principal components explain more than 95% of the variance.

3.2 Model Selection

As there is no rule of thumb to pick a machine learning classification algorithm that would be the most suitable with the dataset under study, several algorithms are picked to be applied on this heart attack dataset. The development and parameters tuning of each model is done exclusively for each model. Each model/algorithm has its own parameters which are tuned to achieve the maximum possible accurate prediction behaviour. The models proposed are logistic regression, K-nearest neighbours (KNN), support vector machines (SVM) and random forest.

3.3 Model Development

3.3.1 Logistic Regression

To investigate the advantage of using the principle components, two different logistic regression models are applied, one with the original data features and the other with the principle components. In the model with the original data features, the logistic regression model is trained and tested with its sigmoid function; however, ROC curve is used to decide on the best threshold to obtain as high recall as possible while keeping acceptable value of precision. To do so and choose the best threshold, k-fold cross validation is applied. After cross validation, the training set is used to also decide on the best threshold to be taken. Finally, the two models are trained separately on the training set, the best probabilities thresholds are set and the models are tested on the test set to evaluate each model. In both models, the number of folds (k) for cross validation is set to 5 as it is a suitable number according to the few number of samples available.

3.3.2 K-Nearest Neighbors

First of all, KNN with K = 3 was applied on the dataset obtained from PCA with 5 principal components to know the resulting scores of each metric and improve it. Afterwards, hyperparameter tuning was applied to tune the hyperparameters: number of nearest neighbors (odd numbers from 3 to 15), weight function used in prediction (distance or uniform), power parameter for the Minkowski metric (1: Manhattan distance or 2: Euclidean distance), and the algorithm used to compute the nearest neighbors (kd tree or ball tree). Typically, the tuning was performed using grid search algorithm with 5- fold cross validation and the performance scoring metric set to "f1".

3.3.3 Random Forest

For the random forest model, an already bagging is applied to perform ensemble learning on multiple decision trees. A grid search is performed for hyper-parameters tuning to choose the optimal tree depth, number of features and number of the performed estimators. As the main sklearn ready-made function for grid search applies cross-validation also, the number of folds is chosen set to one to practically turn off the cross validation as the data samples in many trees would be too small due to the relatively small number of samples available. For the data features, only the nine principle components are used as the input features to the random forest classifier with the grid search as they are greatly predicted to have better performance than the original features.

3.3.4 Support Vector Machines

SVM is predicted to achieve good results, because it fits linear and non linear data. Due to the small size of the dataset and the multiple hyperparameters that need to be tuned, cross validation and grid search were used. Where cross validation was used to determine a range at which each parameter provides the best results, then grid search was used to find the best combination of all parameters.

3.4 Model Evaluation

As mentioned before, the evaluation criteria includes both the recall and precision values, which in turn interjects the f1-score metric. The recall value is much more important due to the nature of its representation in the used heart attack dataset. Due to the few test samples available, the evaluation is dependent on model performance on both training set and test dataset. There is no absolute preference for the performance on test dataset.

4 Discussion and Analysis

4.0.1 Logistic Regression

The logistic regression model with the original features has given a precision value of 0.81 and recall value of 0.93 on the training set. However, by using the validation set and tune the threshold, instead of the default probability threshold, the threshold is set to be 0.531 to get a high true positive rate (TPR) and a low false positive rate (FPR; the true positive rate (TPR) is the exact same metric as recall. The threshold value is very near to 0.5, the default threshold, and this makes sense as the recall value obtained by the training set, which is obtained using the default threshold, is already high. Unfortunately, the FPR value is high which is not good for such an application but its lower in interest compared to TPR value. The final test for this model is done by the test set. The precision and recall of the test set classification results are both 0.941 which is high. Although the model seems to be acceptable, the great test result may be due to the few number of samples and a better performance is predicted from the second logistic regression model.

For the second logistic regression model with the principal components, the exact same procedure is followed. The threshold is tuned to be 0.37. By using the test set, the precision value reaches 0.85 while the recall is 1 which is way better than the previous model. Hence, using PCA with logistic regression has really paid off a much better classifier.

4.0.2 K-Nearest Neighbors

In the previous model, the best results were obtained from the dataset after performing PCA. Therefore, KNN was applied directly on the PCA dataset. Accordingly, f1 score = 0.755, precision = 0.755 and recall = 0.755 on the train data. As for the test data , the f1 score was = 0.7879, precision = 0.8125 and recall = 0.7647 .

Afterwards, grid search with cross validation gave f1 score = 0.722, precision = 0.675, recall = 0.776 . For the test data: f1 score = 0.91426 , precision = 0.888 and recall = 0.9412 . Moreover, the best hyperparameters found were: algorithm = 'auto', number of neighbors = 25, p =1 and weights = 'uniform' .As can be seen, grid search with cross validation obtained better performance metrics, as expected. However, it is important to mention that the test error for this dataset cannot be completely relied on , as the test data is very small and not representative. Therefore, it is important to consider the training error.

4.0.3 Random Forest

By fitting the random forest and applying the grid search hyper-parameters tuning on the selected principle components, the optimum maximum tree depth is set to 5, the maximum number of features is 5 and the number of performed estimators is 60. These parameters are then used for the accepted random forest classifier. By applying to model to the test set, the output precision is 0.83 and the recall is 0.88. The resultant accuracy is less than the one obtained by the logistic regression model.

4.0.4 Support Vector Machines

By applying cross validation and grid search the best results obtained were when the kernel was linear with C =1. The precision and recall obtained from train data were 0.89, 0.78 and thus f1 = 0.81.These results are not very promising, and it was not expected from the SVM. Especially, since the logistic regression obtained better values, which means there is a separation between the two classes, and it is even linear. After using the fitted model, the test results were 0.82 for precision, recall and f1.

Finally, from the results obtained from the four different models, the best model, with respect to the performance metric f1-score and recall, is the logistic regression model after using the five principal components.

5 Ethical Considerations

Through the process of modeling, the ethical dimension was not neglected. Investigating the dataset for any causes of bias or violations of fairness revealed that the data is almost fair. However, as mentioned before and illustrated in a graph, the males (207) were more represented than females (96) in the dataset, which raised suspicion that the data would be biased toward males. By the investigation of the results in train and test data for logistic regression and SVM the results were not skewed. Where for the train data the males were 183(82 +ve/101 -ve), and the females were 88(65 +ve/23 -ve), the false positives of the females were 8, and false negatives are 2, and for males 34 were false positive and 7 are false negative. The values are proportional and do not raise any ethical issue.

For the test results, only three samples were false positive and all of them were males. This does not raise a worry, because it is still within proportion and the worry was that data would be biased toward misdiagnosing negative female patients because they are poorly represented. Also, by repeating this for the SVM, 6 samples were faulty predictions 2 of them females, and 4 males.

After all, it would be certainly better if females:males ratio was 1 and of course if we had more samples to look deeper into the possibility of bias. Besides that, all features are relevant to the study are fair.

6 Professional Responsibility and Accountability

Regarding the professional responsibilities, the process of making the model was very aware of how critical the results are. This motivated us to take special care of the recall score because a false negative error is very critical when it comes to diagnosing a disease. Though, we definitely did not neglect the precision to get an overall good performance of the model. In addition

to that, though these data are biological, and they do not change with time which means that the model should work even after the passage of time, but the small dataset suggests better performance if more data is collected, which means that collecting more data and refitting the model after that, would be very helpful to increase the performance of the model, and very recommended before depending fully on the model. Despite the results being acceptable, heart disease is very serious and should be treated with caution. For now, the model can be used for very good guidance in the process of diagnosis.

7 Conclusion

Different essential pre-processing procedures are applied to the heart attack dataset before fitting different classification models. Logistic regression model, K nearest neighbour model, random forest and support vector machine model are all fitted and tested on the data. Cross validation and grid search procedures are also applied in different situations. The logistic regression model with five principal components turned out to be the best model depending on the test performance metrics. The ethical and professional considerations and responsibility are studied for such models.

References

[1] The top 10 causes of death, WHO, 2018. From: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] UCI, Heart Disease Dataset. From: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 (2020). DOI: 10.1007/s42979-020-00365-y

[4] Chaurasia, Vikas & Pal, Saurabh. (2013). Data Mining Approach to Detect Heart Diseases. International Journal of Advanced Computer Science and Information Technology (IJACSIT). 2. 56-66.

[5] K. Deepika and S. Seema, "Predictive analytics to prevent and control chronic diseases," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, 2016, pp. 381-386, doi: 10.1109/ICATCCT.2016.7912028.

[6] Dwivedi, A.K. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput & Applic 29, 685–693 (2018). DOI: 10.1007/s00521-016-2604-1

[7] Predicting heart-attack using Logistic regression", Kaggle.com, 2021. [Online]. Available: https://www.kaggle.com/ninad19298/predicting-heart-attack-using-logistic-regresssion.

[8] "logistic regression 91.5% & DNN 85% & PCA", Kaggle.com, 2021. [Online]. Available: https://www.kaggle.com/genjihasky/logistic-regression-91-5-dnn-85-pca.

[9] Heart Attack Prediction Using Different ML Models", Kaggle.com, 2021. [Online]. Available: https://www.kaggle.com/nareshbhat/heart-attack-prediction-using-different-ml-models.

[10] 10 surprising facts about cholesterol - CNN.com", Edition.cnn.com, 2021. [Online]. Available: http://edition.cnn.com/2009/HEALTH/11/24/moh.healthmag.cholesterol.surprises/index.html. [Accessed: 14- Jan- 2021].