CSCI 4340 MACHINE LEARNING

Assoc. Prof. Dr. Amelia Ritahani binti Ismail

Assignment 2

Members

Names	Matric No.
Amnah salah majzob abdel maged	1824962
Emon Rifat Hasan	1832901
HIBO SULEIMAN AMEN	1825120

Due date: 12 JUNE 2022

Video Presentation link : https://youtu.be/_cBDNZ81dv4

Abstract

The heart is the most important or vital organ in our bodies. The heart is responsible for maintaining and conjugating blood in our bodies. There are numerous incidences of heart disease around the world. People are dying as a result of heart disease. Various symptoms are listed, such as chest pain, fasting heartbeat, and so on. The health care industries found a large amount of data. This assignment gives the idea of predicting heart disease using machine learning algorithms. Here, we will use various machine learning algorithms such as Support Function Machine (SVM),K-Nearest Neighbor (KNN), and Hybrid algorithms . The algorithms are used on the basis of features and for predicting heart disease. This paper uses different machine learning algorithms for comparing the accuracy among them.

Introduction

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate the chance of heart disease in humans. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients everyday in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

The dataset used is found in Kaggle and it is Predicting heart disease using machine learning .This paper gives the idea of predicting heart disease using machine learning algorithms. Here, we will use various machine learning algorithms such as Support Function Machine (SVM),K-Nearest Neighbor (KNN), and Hybrid algorithms .

Source: Predicting heart disease using machine learning 🩺 | Kaggle.

The Objective of the Heart Disease Prediction

- The goal of heart disease prediction is to determine if a patient should be diagnosed with heart disease or not, so:
  Positive result = 1, the patient will be diagnosed with heart disease.
  Negative result = 0, the patient will not be diagnosed with heart disease.
- We have to find which classification model has the greatest accuracy and identify correlations in our data. Finally, we also have to determine which features are the most influential in our heart disease diagnosis.

Features

This database contains 76 attributes, but all published experiments refer to using a 14 features to determine our predictor:

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

Experimental Setup

Pre-processing

In this experiment, Analyzing and observing the dataset helps to understand the features and characteristics of it, and allows to determine the labeled data and if there is any missing or duplicate value exist.

Furthermore, to ensure the format used in the dataset is suitable for the model all redundant values that don't assist in the diagnosis should be removed from the dataset, luckily, it wasn't found in our dataset. We have found that 165 persons with heart disease and 138 persons without heart disease.

subset of 14 of them.We used the following.
Experimental Setup

From the dataset we understand that:

- Chest Pain (cp): if cp equal to 1, 2, 3 then they are more likely to have heart disease unlike cp equal to 0.
- resting electrocardiographic results (restecg): value 1 are more likely to have heart disease.
- the slope of the peak exercise ST segment (slope): slope value 2 (Downsloping: unhealthy heart) more likely to have heart disease than slope value 0 (Upsloping: better heart) or 1 (Flatsloping: healthy heart).
- number of major vessels (0-3) colored by flourosopy (ca): the more blood movement the better heart. Ca=0 more likely to have heart disease.
- thalium stress result (thal): thal value

equal to 2 are more likely to have heart disease.

- serum cholestoral in mg/dl (chol): more than 200 is cause for concern.
- maximum heart rate achieved (thalach): thalach with more than 140 are more likely to have heart disease.

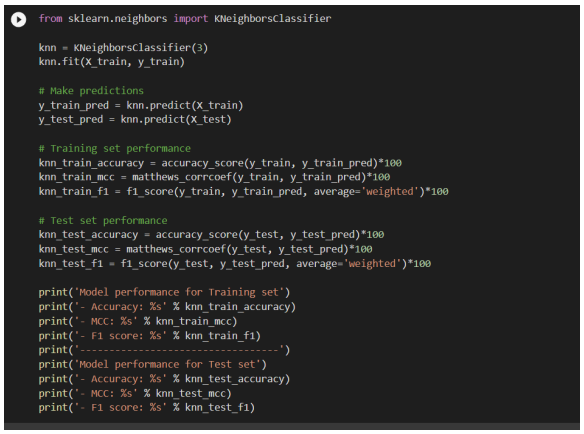
Processing:

After preparing and understanding the data, it needs to be splitted into training set (80%) and testing set (20%) .

Then we used 2 supervised learning method:-

1- k-nearest neighbors(KNN):

This method is non-parametric because it doesn't make any assumptions about underlying data, and it is called lazy learner because it stores the dataset until the time of classification then it performs an action on the dataset instead of learning from the training set.



In this code the dataset will be classified for KNN then it is going to make 2 predictions, train prediction and test prediction. A performance model will be applied for both predictions separately, it will calculate accuracy, F1, and MCC.

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\%$$

$$F1 = \frac{Spec \times Pre}{Spec + Pre} * 2$$

$$Rec = \frac{TP}{(TP + FP)} * 100\ \%$$

```
Model performance for Training set
- Accuracy: 76.03305785123968
- MCC: 55.81019406256292
- F1 score: 75.92379327913498
---------------------------------
Model performance for Test set
- Accuracy: 55.73770491803278
- MCC: 18.47549173297337
- F1 score: 55.36968885915022
```

As we can see the accuracy for the training set is 76%, MCC almost 56%, and F1 approximately scored 76%. Moreover, the test set accuracy dropped to 56%, MCC 18%, and F1 scored 55%. Finally, print the confusion matrix for train set and train prediction.

```
[[  1   0   1   0]
 [  0   6   6   2]
 [  2   4 109  18]
 [  0   2  23  68]]
```

2-Support vector machines (SVM):

A supervised learning method used for classification, regression and outliers detection problems in Machine Learning. it aims to create the best line (decision boundary) that can divide n-dimensional space into classes to make it easier to put the new data point in the correct category in the future.

```python
from sklearn.svm import SVC

svm_rbf = SVC(gamma=2, C=1)
svm_rbf.fit(X_train, y_train)

# Make predictions
y_train_pred = svm_rbf.predict(X_train)
y_test_pred = svm_rbf.predict(X_test)

# Training set performance
svm_rbf_train_accuracy = accuracy_score(y_train, y_train_pred)*100
svm_rbf_train_mcc = matthews_corrcoef(y_train, y_train_pred)*100
svm_rbf_train_f1 = f1_score(y_train, y_train_pred, average='weighted')*100

# Test set performance
svm_rbf_test_accuracy = accuracy_score(y_test, y_test_pred)*100
svm_rbf_test_mcc = matthews_corrcoef(y_test, y_test_pred)*100
svm_rbf_test_f1 = f1_score(y_test, y_test_pred, average='weighted')*100
print('Model performance for Training set')
print('- Accuracy: %s' % svm_rbf_train_accuracy)
print('- MCC: %s' % svm_rbf_train_mcc)
print('- F1 score: %s' % svm_rbf_train_f1)
print('---------------------------------')
print('Model performance for Test set')
print('- Accuracy: %s' % svm_rbf_test_accuracy)
print('- MCC: %s' % svm_rbf_test_mcc)
print('- F1 score: %s' % svm_rbf_test_f1)
```

Same as KNN the dataset is going to be splitted into 2 predictions, train prediction and test prediction. Then a performance model is going to be applied for both predictions, it will calculate accuracy, F1 Score , and MCC.

```
Model performance for Training set
- Accuracy: 100.0
- MCC: 100.0
- F1 score: 100.0
---------------------------------
Model performance for Test set
- Accuracy: 54.09836065573771
- MCC: 0.0
- F1 score: 37.98395535402861
```

We can notice how successful this method was in the training set accuracy, MCC, and F1 are all 100% unlike the test set, there was

a huge drop in accuracy 54%, MCC 0%, and F1 scored almost 38%. Then finally, print the confusion matrix.

```
[[  2   0   0   0]
 [  0  14   0   0]
 [  0   0 133   0]
 [  0   0   0  93]]
```

In the next step we combined both KNN and SVM algorithms to improve the expected result by creating a Hybrid Model. This model is used whenever we want to combine any 2 algorithms in machine learning. The steps are similar to SVM and KNN. First, set the predictions(train and test). Second, apply a performance model on both sets.

```
Model performance for Training set
- Accuracy: 69.00826446280992
- MCC: 44.04388577074535
- F1 score: 63.353220206367055

Model performance for Test set
- Accuracy: 57.377049180327866
- MCC: 14.690718258463603
- F1 score: 50.10958183571491
```

We can see that the 2 algorithms were successfully combined and the performance model shows that in the training method is a bit higher than those normal models. Unlike the result of the test set, which got higher accuracy than KNN and SVM with 57%, MCC with 15%, and F1 scored 50%.

Finally, the confusion matrix for Hybrid model is

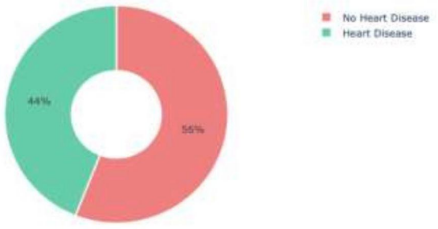
```
[[  0   0   2   0]
 [  0   0  11   3]
 [  0   0 133   0]
 [  0   0  59  34]]
```

CONCLUSION

A cardiovascular disease detection model has been developed using three ML classification modeling techniques. This assignment predicts people with cardiovascular disease by extracting the patient's medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain,

sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them being diagnosed with a previous heart disease. The algorithms used in building the given model are Hybrid Algorithm, SVM and KNN . The accuracy of our model is 87.19% for KNN algorithm . Use of more training data ensures higher chances of the model to accurately predict whether the given person has a heart disease or not.

Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying Hybrid algorithm and KNN to get an accuracy of an average of 94.2 % on our model which is better than the previous models having an accuracy of 87.19%. Also, it is concluded that accuracy of SVM is highest among the three algorithms that have accuracy of 100%. Figure below shows 44% of people that are listed in the dataset are suffering from Heart Disease.



REFERENCE

oni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overviewof heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

Harshit.J(2021).Heart disease prediction using machine learning algorithms.Retrieved from, https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf.

Aman,G.(2022 ,11 February).Heart disease prediction using machine learning algorithms.Retrieved from,https://www.analyticsvidhya.com/blog/2022/02/heart-disease-prediction-using-machine-learning/.