Course - IS 733 – Data Mining

(Spring 2020)

Final Project Report

On

Heart disease prediction

Instructor

Dr. James Foulds

Assistant Professor - Information Systems

University of Maryland, Baltimore County

Submitted by,

Sakshi Parikh

Pranitha Lolla

Anila Chintalapati

Anjana Kethineni

Arjun Kumar

Dharmil Shah

Submitted on - May 12th, 2020

Abstract:

Cardiovascular disease is the vital cause of death in human beings. Men tend to be more vulnerable to cardiovascular diseases when compared to women. There are various risk factors and health features that contribute to the cause of cardiovascular diseases. According to the American Heart Association every 37 seconds, 2,353 people die from cardiovascular disease in the United States. Hence, there is a need to identify the causes and risk factors to predict the disease before its final stage. This prediction can help in recognizing the disease at an early stage and can help in providing medication before the diseases become fatal. Hence, the motive of this project is to predict the presence or absence of cardiovascular diseases among human beings. A dataset with 70,000 instances is considered for the prediction process. This data set is refined and then a series of methodologies are implemented in the process in order to derive the accuracy of the prediction. Finally, a strategy of predicting heart disease is derived and the most accurate method for prediction is implemented. In this project, the overview, and the phases like data pre-processing, data analysis, process flow, methodologies, accuracy results, challenges are keenly explained. The following figure depicts the causes of death in both men and women separately.

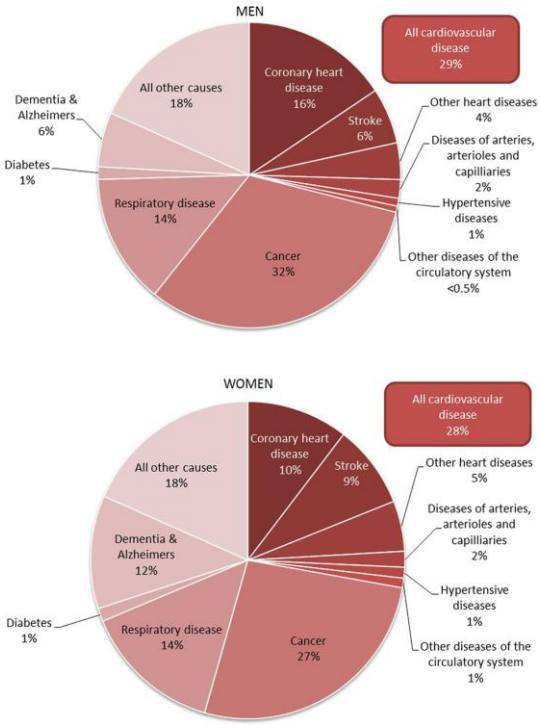


Figure 1

Deaths by cause and sex, UK. This figure compiles data from the four countries of the UK. In Northern Ireland, the data for lung cancer only includes International Classification of Diseases-10 code C34. Adapted from England and Wales, Office for National Statistics (2014) Deaths registered by cause, sex and age. http://www.statistics.gov.uk (accessed January 2014); Scotland, National Records of Scotland (2014) Deaths, by sex, age and cause. http://www.gro-scotland.gov.uk (accessed January 2014); Northern Ireland, Statistics and Research Agency (2014) Registrar General Annual Report. NISRA: Belfast.

Introduction:

Cardiovascular diseases can prove to be fatal when neglected. It becomes impossible to cure when diagnosed in the final stages. To be precise according to these sources Cleveland Clinic Heart and Vascular Institute, American College of Cardiology, American Heart Association. There are four stages of cardiovascular disease. They are Stage A, stage B, stage C, stage D. In stage A, an individual is more likely to get a heart stroke for various reasons. One of the most common factors depends on the health habits of individuals. The factors like hereditary diabetes, alcohol intake, high blood pressure supplement to the cause of heart stroke. Stage B occurs when there are no symptoms of a heart attack, but the patient is diagnosed with systolic, diastolic dysfunction. Stage C is the failure of the systolic heart. This stage may not be fatal but if this stage goes unnoticed or untreated, the patient may lose their life. In stage D the patient shows advanced symptoms. This may result in heart replantation or various surgeries.
For this project, the dataset constituting numerous entries and instances was considered. This dataset consists of 70,000 records of patient's data split into 11 accommodating features/attributes. There are three types of input features in this dataset. Firstly, Objective inputs are the ones focusing on the facts of respective patients. For example, facts like age, height, weight, the gender of a patient are considered as the objective inputs. The examination is the second input feature. This is the result of medical examinations of the patients. Systolic blood pressures, diastolic blood pressures, cholesterol, and glucose are considered as examination inputs in this data set. The last feature is subjective input. It is the information given by each cardiovascular patient. These inputs are taken as binary values. Components such as smoking, alcohol intake, physical activities of the patient are considered as the subjective inputs. Finally, the presence of the cardiovascular disease is represented as 1 and the absence of the disease is represented as 0.

Data pre-processing:

It is evident that data pre-processing is an important step in the data mining process. In most instances, the data gathering methods have erroneous instances and attributes with missing values, beyond the range factors and unreasonable data entries and formats. The original dataset considered for this project has numerous attributes. So, the features which contribute the most in cardiovascular disease prediction are only examined. Data cleaning is the first step taken to clean the data. In this project, data cleaning included standardizing the input values and removing spaces between them. Missing data values in the dataset were handled by eliminating the null values. Hence, 70,000 instances are considered after the data pre-processing phase. The resulting data is then split into a training set and test data on a random basis. Finally, the data is encoded. Encoding is a process that secures the data, saves memory, and encoding also results in uninterrupted speed when the data is manipulated and saved. All the nominal attributes like cholesterol, glucose, alcohol intake, smoking is taken as binary values. And the presence or absence of cardiovascular disease is also considered as a binary value. Presence =1 and absence=0. The current data set has the age of patients in the count of days. This age of patients is converted into years for precise calculation.

The following figure is the block diagram of the heart disease prediction.

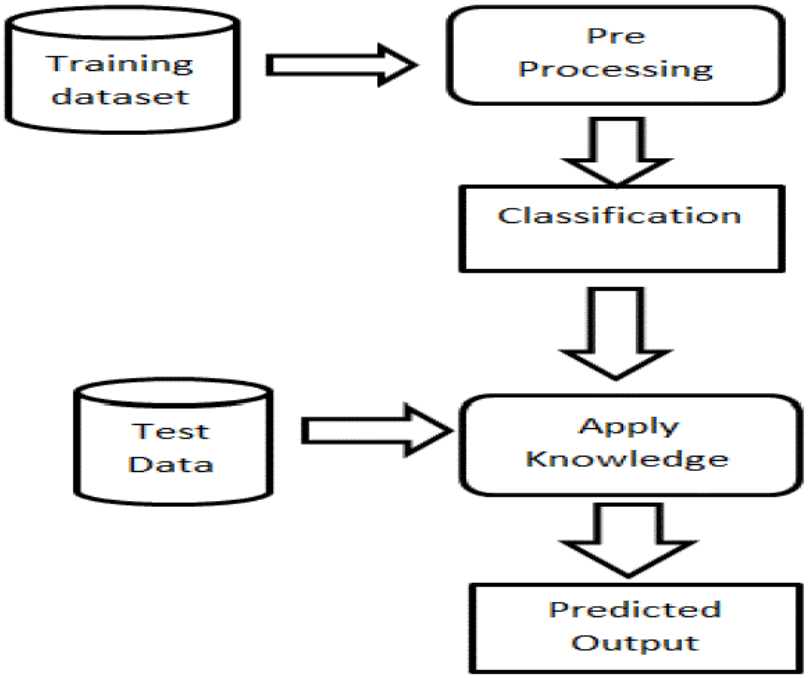


Figure 2

Methods:

To find exact relation between the attributes and the target attribute which is if the person is having heart disease or not, we will be performing the 3 following algorithms:

a. Logistic regression classifier

b. Random forest classifier

c. K-Nearest Neighbors (KNN), and

d. Voting classifier to predict the maximum accuracy.

After implementing the first 3 methods independently, we implemented a voting classifier to get the best possible output from all the three methods that we are using.

Methodology:

a. Logistic Regression Classifier:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. It is a predictive analysis algorithm and based on the concept of probability. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". This leads to the most straightforward interpretation. In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labelled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labelling; the function that converts log-odds to probability is the logistic function.

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation

In [13]:
```
# Logistic Regression

lr_scores=[]

lr = LogisticRegression(penalty='l1',solver= 'liblinear')
lr.fit(x_train,y_train)
lr_score = accuracy_score(y_test,lr.predict(x_test))
print(accuracy_score(y_test,lr.predict(x_test)))
```

```
0.7225714285714285
```

This is a binary logistic regression and we got the accuracy of 72%

b. Random Forest Classifier:

Random forest is a supervised learning algorithm. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Random forest adds additional randomness to the model, while growing the trees. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds.

In [28]:
```
# Random Forest Classifier

rfc = RandomForestClassifier(n_estimators=500, criterion = 'entropy')
rfc.fit(x_train,y_train)
rf_score = rfc.score(x_test,y_test)
print(rfc.score(x_test,y_test))
```

```
0.7130857142857143
```

This is the prediction of our algorithm using random forest regression. We got the accuracy score of 71% which is lower than the logistic regression classifier.

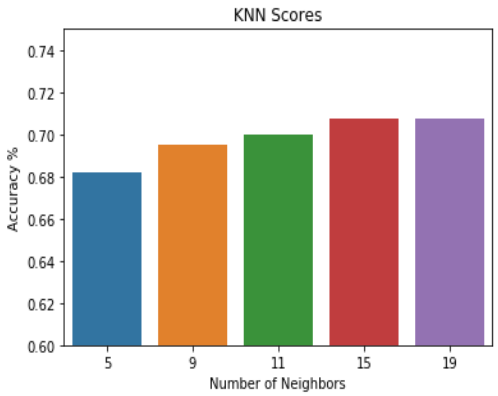
c. K-Nearest Neighbors (KNN):

K-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature vector. A feature vector is our mathematical representation of data, and since the desired characteristics of our data may not be inherently numerical, pre-processing and feature-engineering is required to create these vectors. Given data with N unique features, the feature vector would be a vector of length N, where entry I of the vector represents that data point's value for feature I. Each feature vector can thus be thought of as a point in R^N. Unlike most other methods of classification, KNN falls under lazy learning, which means that there is no explicit training phase before classification. Instead, any attempts to generalize or abstract the data is made upon classification.

In [8]:
```
# K Nearest neighbor Classifier

kn_scores = []
knn = [5,9,11,15,19]

for k in knn:
    knc = KNeighborsClassifier(n_neighbors=k)
    knc.fit(x_train,y_train)
    kn_scores.append(knc.score(x_test,y_test))
```

In [12]:
```
# Accuracy of different KNN Classifiers

sns.barplot(knn, kn_scores)
plt.ylim(0.6,0.75)
plt.title('KNN Scores')
plt.xlabel('Number of Neighbors')
plt.ylabel('Accuracy %')
```

Out[12]: Text(0, 0.5, 'Accuracy %')



In this supervised KNN approach we got an accuracy of 70%, which is clearly less than the other two classifiers.

d. Voting Classifier:

Voting classifier is an ensemble method that takes the majority outcomes or plurality outcomes of different machine learning algorithms. Majority as the name suggests means the outcome that appears the most number of times while plurality means that even if a particular machine learning algorithm hasn't received the majority but has more votes than the other algorithm it is considered as the majority and we chose that algorithm.

Voting classifier is better to use as it takes the poll of all the machine learning algorithms that you have implemented and provides the answer according to the majority of the outputs. Which is most of the time correct.

Soft Voting can only be done when all your classifiers can calculate probabilities for the outcomes. Soft voting arrives at the best result by averaging out the probabilities calculated by individual algorithms.

In [16]:
```
# Voting Classifier

knn = KNeighborsClassifier(n_neighbors=15)
mlp = MLPClassifier(hidden_layer_sizes=512,activation='relu')
lr = LogisticRegression(penalty='l1',solver= 'liblinear')
gnb = GaussianNB()
rfc = RandomForestClassifier(n_estimators=500, criterion = 'entropy')

vc = VotingClassifier(estimators=[('knn',knn),('lr',lr),('rfc',rfc)], voting='soft')
```

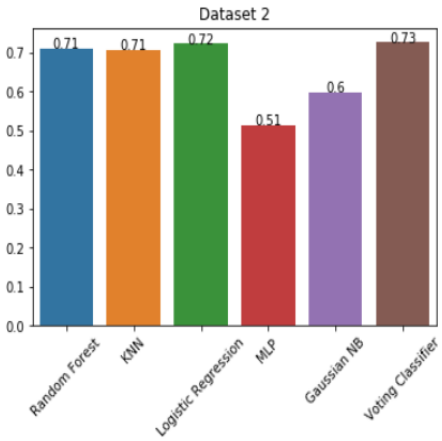
In [17]:
```
vc.fit(x_train,y_train)
```

Out[17]:
```
VotingClassifier(estimators=[('knn', KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=15, p=2,
           weights='uniform')), ('lr', LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scalin...obs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False))],
         flatten_transform=None, n_jobs=None, voting='soft', weights=None)
```

In [18]:
```
# Voting Score

vc_score = vc.score(x_test, y_test)
print (vc.score(x_test, y_test))
```

```
0.7262857142857143
```

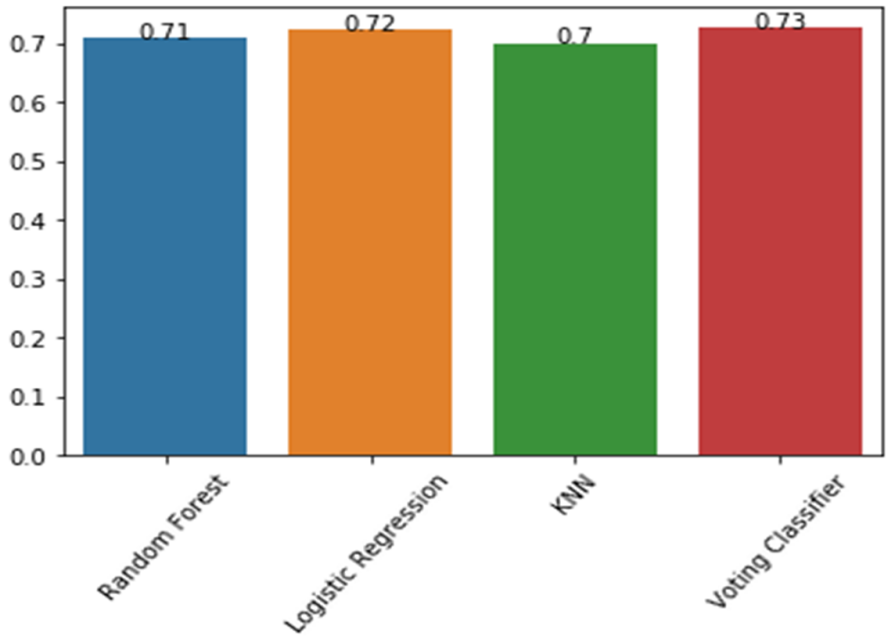
In [19]:
```
label = ['Random Forest','KNN','Logistic Regression','MLP','Gaussian NB','Voting Classifier']
scores = [rf_score,kn_scores[3],lr_score,mlp_score,gnb_score,vc_score]
sns.barplot(x=label,y=scores)
plt.xticks(rotation=45)
for i in range(len(label)):
    plt.text(i,scores[i],round(scores[i],2),horizontalalignment='center')
plt.title('Dataset 1')
plt.savefig('results.png')
```



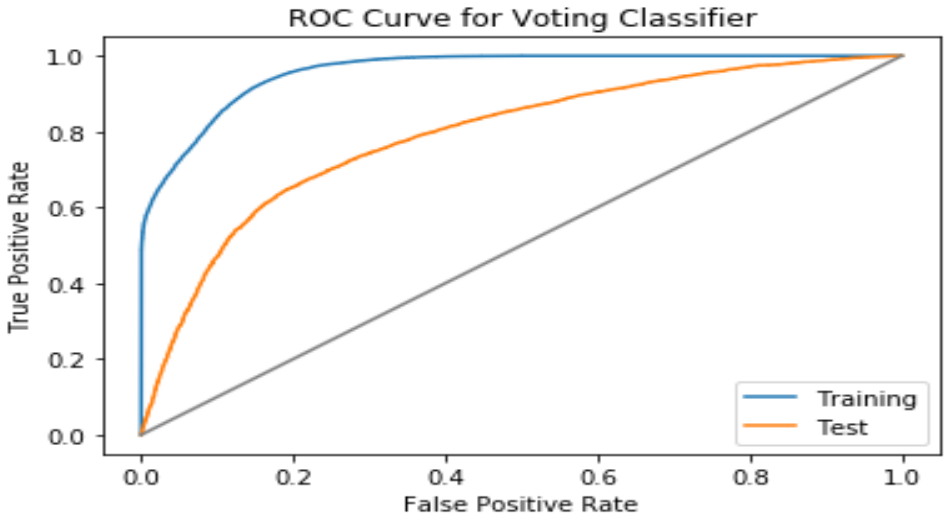
Here the voting classifier has the highest accuracy i.e., 73%.

Results:

We have used two machine learning algorithms which are KNN and Logistic Regression and two ensemble methods Voting Classifier and Random Forest. In this dataset, Age and Weight of the individual are the most important features for prediction of the class (presence of heart condition in the individual). In, Voting Classifier we have observed the highest accuracy. In this classifier we have used the final class label which was voted the in majority by highest predictions and the result depends on the majority test cases. So, the accuracy captured was 73% in our model. This was done when we took a poll of all machine learning classification algorithms and in our case, it turned out to be that most of the majority test cases were correct. This graph shows us all the classifiers with their accuracies with classifiers on X-axis and accuracy on Y-axis.



The figure below depicts the ROC curve for voting classifier which is Receiver Operating Curves. This curve represents False Positive Rates on X-axis vs True Positive Rates on Y-axis. We have used ROC curves to visualize our result on the voting classifier because it has determined the highest accuracy. The true positive rate for training data was higher than the true positive rate for testing data. It had a gradual increase and flattened with an intersection. ROC is most useful in the initial stages for assessment.



Conclusion and Future work:

Models	Accuracy
Random Forest	71 %
KNN	70 %
Logistic Regression	72 %
Voting Classifier	73%

We have used the four models above to predict heart diseases and from the evaluation result we can conclude that Voting Classifier provides better accuracy of 73% which is almost 2% higher than other models such as Logistic Regression, KNN and Random Forest Classifier. It is not easy to manually detect the chances of diagnosing a heart disease based on certain risk factor. So, we have used these machine learning algorithms to be used to predict the results from existing data sets.

This is a more generalized way to predict heart diseases but in future this can be used to analyse different data sets. We can also work on improving the performance of the diagnosis of health by handling various class labels during the process of prediction. We can improve the models and their accuracy by implementing either deeper neural network algorithms or training more data. We have also used grid search for best parameter values, the best method was random forest classifier, and its accuracy was still 72%. From the visualization of the rules we found that systolic and diastolic blood pressures are the important attributes in classifying. Also, in most cases the determination of class is independent of attributes such as smoking, drinking or exercises or not.

References:

1. 'Improving the accuracy of prediction of heart diseases'- https://www.sciencedirect.com/science/article/pii/S235291481830217X .

2. 'The future of Heart Attack Prediction' - https://mendedhearts.org/story/the-future-of-heart-attack-prediction/ .

3. (Figure 1) The epidemiology of cardiovascular disease in the UK 2014- https://heart.bmj.com/content/101/15/1182

4. (Figure 2) Block diagram of the proposed system- http://www.rroij.com/articles-images/IJIRCCE-1311-g001.html