East West University

Project Report

Project Tittle: Lung Cancer Detection Using Classification Algorithms

Course Tittle: Machine Learning

Course Code: CSE 475

FALL 2022

Section: 01

Date of Submission: 11th January 2022

Submitted by:

Mahmud Jamil

ID: 2017-2-60-147

Farhat Jebin

ID: 2017-2-60-059

Md. Fahimul Islam

ID:2017-1-62-022

Submitted to:

Dr. Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering

Description:

Lung cancer is the uncontrolled growth of abnormal cells in one or both lungs. These abnormal cells do not carry out the functions of normal lung cells and do not develop into healthy lung tissue. This one of the leading cause of cancer deaths worldwide. A copious number of researches are going on worldwide regarding the detection of lung cancer easily by analyzing symptoms primarily before going into many complex medical tasks. Implementation of various machine learning algorithm in a medical criteria has been always a great help. In this paper, we have tried to show how much accurately various predictive models can predict presence of lung cancer from a given set of symptoms. Also, we have discussed a comparison between some classifiers which has been used for our predictive modeling.

Introduction:

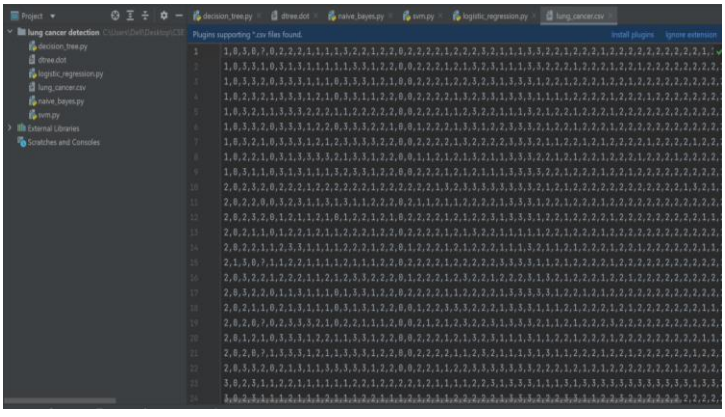
Medical Science is a very important area where data mining or machine learning techniques can be applied. In this present world, there are hundreds of thousands of diseases that people are suffering from. Any disease has a specific set of symptoms. Some specific combination of these symptoms causes that specific disease. So, from a number of medical checkup history of a selected disease, if a dataset can be created with the extraction of the combination of symptoms then a classifier algorithm can be trained with that dataset. Later the presence of that disease in a human body can be predicted easily only by taking the symptoms as input. In this case we have worked with lung cancer detection.

Problem Statement:

From a machine learning point of view, we need to work on a huge data for the prediction of lung cancer presence. we can emerge the machine learning technique for detection of many diseases. Mistakes are likely to be made in most of the cases where traditional analysis of data predicts if someone has lung cancer or not, without any complex medical exertion.

The main goal of applying machine learning in this criteria is to computerize this type of analysis and improve the accuracy of the prediction. Also, development of predictive models can pave the way to use this aspect in other areas of medical research.

About Dataset:



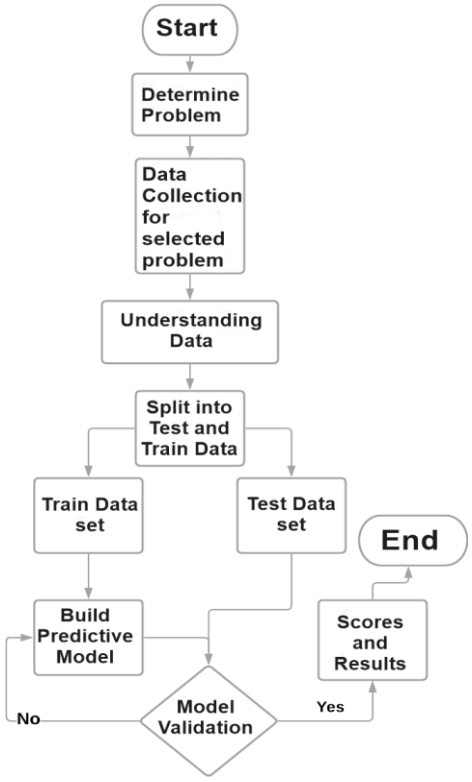
This file which includes 32 rows and 57 attributes. Each row is medical records of detection of lung cancer. The first attribute denotes the result. This is the class attribute. Other 56 attributes can be denoted as many symptoms. This data was used by Hong and Young to illustrate the power of the optimal discriminant plane in ill- posed settings. In the first column data described 3 types of pathological lung cancers which are denoted with respectively 1, 2 and 3. The Authors give no information on the individual variables nor on where the data was originally used. There were some missing or irrelevant values in original dataset. All of them are replaced with a (?).

Now,we have used four classifier algorithms to build up predictive models for lung cancer detection according to the dataset. The used classifiers are:

- Decision Tree
- Naïve Bayes
- Support Vector Machine
- Logistic Regression

We have calculated the accuracy of each model by five time cross validation. Also F1 scores have been calculated for each model. Each models has been run on the dataset for Ten times in order to find difference between them and distinguish which one is best .

Working Process Flowchart:



Tools:

We used Pycharm ide, Pyhton 3 and this Python packages for out project.

- We have used ‘pandas’ to read data from dataset.
- ‘Numpy’ has been used to handle the data and work with missing values.
- ‘Scikit Learn’ has been used for initiating the classifiers and train and test them with the data.
- From OS, system has been imported to draw the decision tree.
- ‘Scikit Learn’ has been also used to determine accuracy, f1 scores of the classifiers after testing them with the data.

Testing:

```
data = read_csv("lung_cancer.csv")
X=data.iloc[:,1:].values
Y=data.iloc[:,0]

imp = SimpleImputer(missing_values='?', strategy='most_frequent')
X=imp.fit_transform(X)
X=pd.DataFrame(X)
```

Spliting Data for Training and Testing.

We needed to split the data for training and testing. For a good training of a classifier, huge amount of record is necessary. However, there are only 32 instances in the provided dataset. We have taken 80% of the instances in order to train the classifiers and 20% of the instances to test the classifiers. With using this splitting prediction by the models.we have been able to achieve good Data has been divided in to two sets, X & Y. Class Label means the Y dataset which is the first attribute in the provided dataset. Rest of the attribute has been taken into X. Then X and Y both are split into X_train, X_test and Y_train, Y_test.

```
X_train, X_test, Y_train, Y_test =train_test_split(X,Y, test_size=0.2)
```

Result Analysis in all Algorithms:

1.Decision Tree Output:

We have applied the algorithm on the provided dataset for ten times and recorded the cross validation score, F1 Score and accuracy. It shows different cross validation score but always greater than 0.6. However, the accuracy is always 100% For each calculation the confusion matrix is calculated as a diagonal matrix. Diagonal Confusion Matrix means the prediction accuracy is 100%. Here is a screen shot of the output after applying decision tree classifier algorithm on the dataset for 10 times:

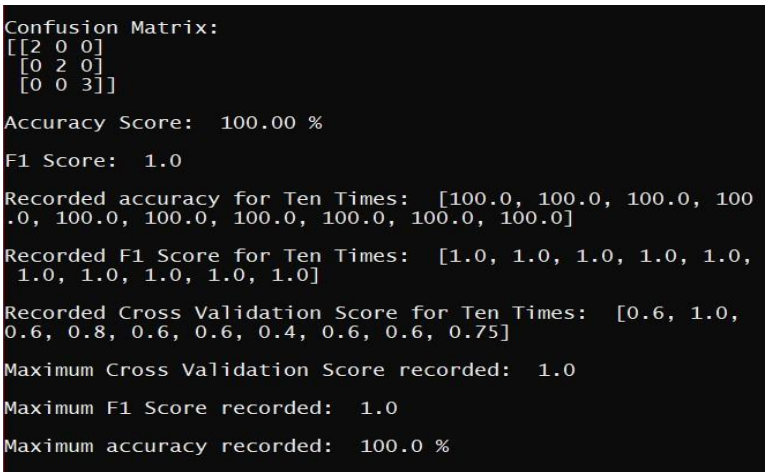


Figure :Decision Tree Output

Maximum Cross Validation Score: 1.0

Maximum F1 Score: 1.0

Maximum Accuracy: 100%

Confusion Matrix: Always Diagonal

2. Naïve Bayes Algorithm :

Applying Naïve Bayes Algorithm We have used Gaussian distribution for applying Naïve Bayes classifier algorithm on the provided dataset. Like decision tree algorithm we also ran this classifier for ten times and recorded the cross validation score, F1 Score and accuracy. It shows various accuracies, F1 Scores. But the cross validation score is always greater than 0.6. Also, the confusion matrix is not always diagonal. Here is a screen shot of the output after applying Naïve Bayes classifier algorithm on the dataset for 10 times:

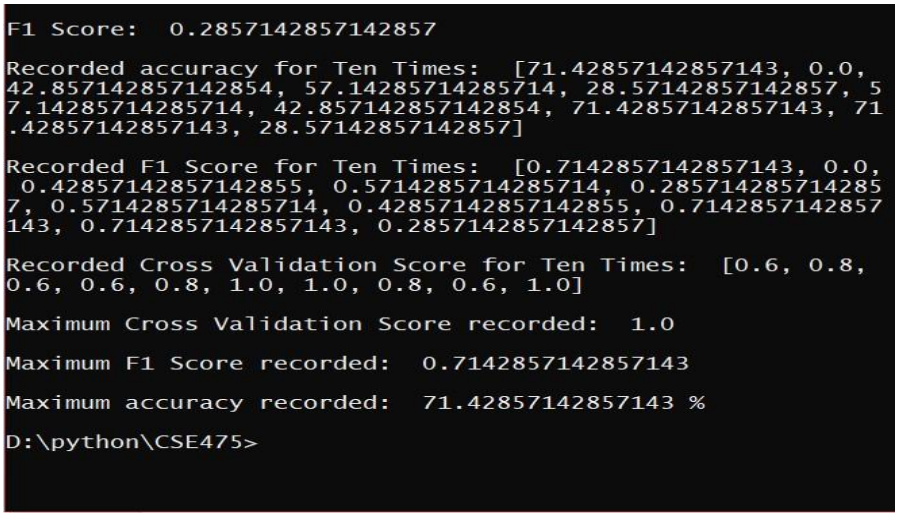
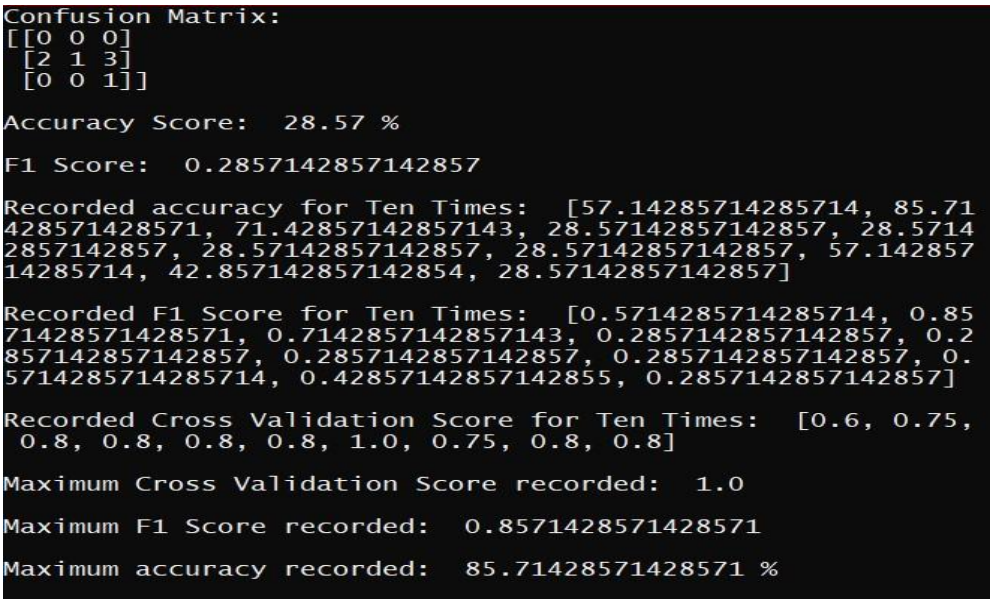


Figure: Output of Naïve Bayes Algorithm

3. Logistic Regression:

We also wanted to show how well a linear model can perform. So we have applied logistic regression algorithm the provided dataset. Also for this algorithm, we applied this on our dataset for 10 times. Like Naïve Bayes algorithm, the confusion matrix is not always diagonal for Logistic Regression. F1 score and accuracy vary a lot. Cross Validation Score is always greater than 0.6. Screenshot of the output after applying Logistic Regression Algorithm.



Logistic Regression Output

Maximum Cross Validation Score: 1.0
Maximum F1 Score: 0.8571
Maximum Accuracy: 85.71%

Confusion Matrix: Not Always Diagonal

Maximum Cross Validation Score: 1.0
Maximum F1 Score: 0.8571
Maximum Accuracy: 85.71%

Confusion Matrix: Not Always Diagonal

4. Applying Support Vector Classifier Algorithm :

Using Support Vector as classifier is also one of the simplest supervised learning algorithm. Applying this algorithm we achieved a result which is similar to the result given by naïve bayes algorithm.

This algorithm was also ran for 10 time on the dataset. The confusion matrix is sometimes diagonal, F1 score and accuracy vary a lot and the cross validation score is always greater than 0.6 but maximum is 0.8.

Here is a screenshot of the output after applying support vector classifier algorithm:

```
Accuracy Score:  42.86 %

F1 Score:  0.42857142857142855

Recorded accuracy for Ten Times:  [42.857142857142854, 71.4
2857142857143, 42.857142857142854, 42.857142857142854, 57.1
4285714285714, 57.14285714285714, 42.857142857142854, 57.14
285714285714, 71.42857142857143, 42.857142857142854]

Recorded F1 Score for Ten Times:  [0.42857142857142855, 0.7
142857142857143, 0.42857142857142855, 0.42857142857142855,
0.5714285714285714, 0.5714285714285714, 0.42857142857142855
, 0.5714285714285714, 0.7142857142857143, 0.428571428571428
55]

Recorded Cross Validation Score for Ten Times:  [0.75, 0.6,
 0.6, 0.6, 0.6, 0.6, 0.8, 0.8, 0.4, 0.6]

Maximum Cross Validation Score recorded:  0.8

Maximum F1 Score recorded:  0.7142857142857143
```

Support Vector Classifier Output

Maximum Cross Validation Score: 0.8
Maximum F1 Score: 0.7142
Maximum Accuracy: 71.42%
Confusion Matrix: Not Always Diagonal

DISCUSSION:

From the results section if we collect and show the average results in a table:

Classifier Algorithm	Confusion Matrix	Max F1 Score	Max Accuracy	Max Cross validation Score
Decision tree	Always Diagonal	1.0	100%	1.0
Naïve bayes	Not always diagonal	0.714	71.42%	1.0
Logistic Regression	Not always diagonal	0.857	85.71%	1.0
Support Vector Classifier	Not always diagonal	0.714	71.43%	0.8

We have trained and tested many classifier algorithm with same dataset and tried to predict the presence of lung cancer taking 56 different symptoms in consideration. The reason behind using more than one algorithm is to show the difference between them and find out which is best for creating a predictive model for lung cancer detection. Among all the classifiers, applying a CART algorithm, Decision Tree, has given the most accurate result with 100% . No other classifiers is doing that for this dataset. It must be noted that there are 56 attributes in the dataset and it must be distinguished that which attributes shows more importance and are likely to be selected for calculation. The accuracy 84.3% provides us the assurance that the program we have built is reliable. Also, if the program is trained with more music data, for instance, 10 or 15 thousand of data, the program's prediction will be more accurate.