INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 3 | Year: 2020

COMPARISON OF CLASSIFICATION ALGORITHMS USING CANCER DATA

T. Chalapathi Rao [1]

Sr. Assistant Professor,
Department of Computer Science and Engineering,
Aditya Institute of Technology and Management (A), Tekkali
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh, India

P. Preethi [2], **A. Madhuri** [3], **P. Sowmya** [4], **S. Neelima** [5]

[2,3,4,5] IV B.Tech Students,
Department of Computer Science and Engineering,
Aditya Institute of Technology and Management (A), Tekkali
Jawaharlal Nehru Technological University Kakinada, Andhra Pradesh

Abstract: *Cancer is one of the leading causes of death worldwide. Early detection and prevention of Cancer plays a very important role in reducing deaths caused by cancer. Cancer is the un controlled growth of abnormal cells in the parts of the body. We will analyze the cancer prediction using classification algorithm such as Naive Bayes, KNN, Random Forest and Decision Tree algorithm. Initially, we will collect different datasets of specific cancer. We then collect cancer and non-cancer patient's data of this locality. And we will prepare data sets using the collected data. The data is preprocessed and analyzed using different classification algorithm.*

The main aim of this project is to provide the earlier warning to the users and compare the performance analysis of the classification algorithms.

Key words: *Machine Learning, Lung Cancer Disease, Breast Cancer Disease, Liver Cancer Disease, Collected data, Random Forest, Decision Tree, KNN, Naïve Bayes.*

1. INTRODUCTION

Data mining is a crucial step in discovery of knowledge from large data sets. Data mining is major role in extracting the hidden information in the medical database. The main leading cause of death in today's world is cancer. Cancer is one of the most common diseases in the world that results in majority of death. Cancer is the one of the leading cause of cancer deaths in both women and men Cancer is a type of diseases that causes the cells of the body to change its characteristics and cause abnormal growth of cells.

In this we will collect different cancer datasets. Data mining consists of different classification algorithms like Naïve Bayes , Decision tree, Random forest, KNN etc. In this we implement different data mining techniques. We remove the noise data, preprocess the data and will apply the algorithms. For every dataset we will apply all the four algorithms and determines which algorithm has best accuracy and which the suitable algorithm for the dataset is.

The main aim is to compare the algorithms and decides which algorithm has best accuracy.

2. RELATED WORKS

George et al... applied naive Bayes and decision tree algorithms for the prediction of lung cancer data. Using data mining techniques, data can be classified to bring forward interesting patterns or serve as a predictor for future trends. Numerous algorithmic techniques have been developed to extract information and discern patterns that may be useful for decision support.

Krishnaiah et al.. proposed to a model for nearly detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient. Using data mining lung cancer symptoms such as age, sex, wheezing, shortness of breath, Pain in shoulder, chest, arm, it can predict the likelihood of patients getting a lung cancer disease.

Charles et al.. Suggests that none of the data mining and statistical learning algorithms applied to breast cancer dataset outperformed the others in such way that it could be declared the optimal algorithm and none of the algorithm performed poorly as to be eliminated from future prediction model in breast cancer survivability tasks.

Sazzaduretal..Liver cancer is the leading cause of global death that impacts the massive quantity of humans around the world. This disease is caused by an assortment of elements that harm the liver. For example, obesity, an undiagnosed hepatitis infection, alcohol misuse. The goal of this work is to evaluate the performance of different Machine Learning algorithms in order to reduce the high cost of liver cancer diagnosis by prediction.

In this they used six algorithms Logistic Regression, K Nearest Neighbors, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance of different classification techniques was evaluated on different measurement techniques such as accuracy, precision, recall, f-1 score, and specificity. Present studies mainly focused onthe use of clinical data for liver disease prediction and explore different ways of representing such data through our analysis.

3. DATA COLLECTION:

3.1 **Breast Cancer Attributes:** Clump thickness, Unif-cell-size, Unif-cell-shape, Bare nuclei, Norm nuclei,Mitosis.

3.2 **Lung Cancer Attributes:** Air pollution, Alcohol, Dust Allergy, Occupational Hazards, Genetic risk, Smoking, Dry cough, Weight loss, Swallowing, Chronic Lung disease, Balanced diet, Coughing of blood, Fatigue, Snoring.

3.3 **Liver Cancer Attributes:** Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphatases, Alamine, Amino transferase, Total-protiens-albumin.

3.4 **Collected Dataset Attributes:** Age, Address, Living, Type of cancer, Suffer family, Treatment method, Type of therapy, Diagnosis method, Height, Weight, Blood pressure, Alcohol, Smoking, Sleep duration, Job, Pain, Gutkha, Milk, Blood group, Oil used, Eat fast food, Other disease, Walk, Mobile, Take drugs, Meditation, Target.

4. DATA PREPROCESSING:

In this study, we analyzed 699 breast cancer patient's, 1000 lung cancer patient's, 583 liver cancer patient's, 1008 cancer and non-cancer patient's data whereas 504 samples are cancer patient and 504 samples are non-cancer patients. The ratio of total breast cancer patients is presented inFig.1.

The ratio of total lung patients is presented in Fig.2. The ratio of total lung cancer patients is presented in Fig. 3. The ratio of total cancer and non-cancer patients is presented in Fig.4.

BREAST CANCER :

NO – 458
YES – 231
TOTAL 699

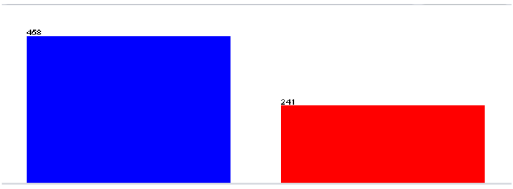

Fig 1: Count Plot shows the ratio of breast cancer patients

LUNG CANCER:

TOTAL :1000
LOW 303
MEDIUM: 332
HIGH 365

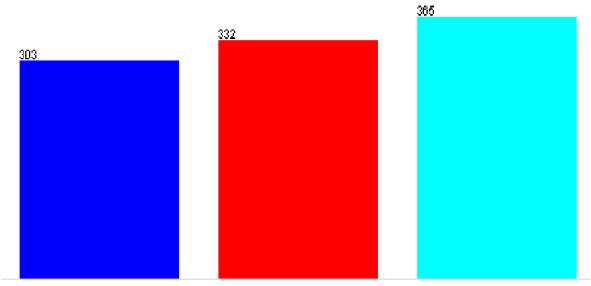

Fig 2 : Count Plot shows the ratio of lung Cancer patients.

LIVER CANCER:

TOTAL =583
YES=416
NO=167

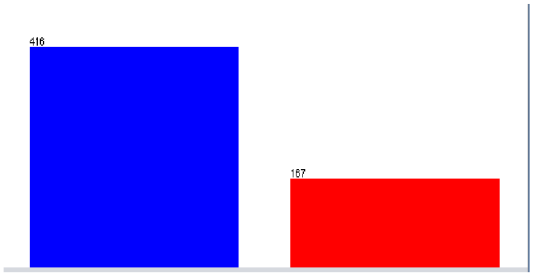

Fig 3: Count Plot shows the ratio of liver cancer patients

CANCER/NON CANCER:

OTAL=1008
YES=504
NO=504

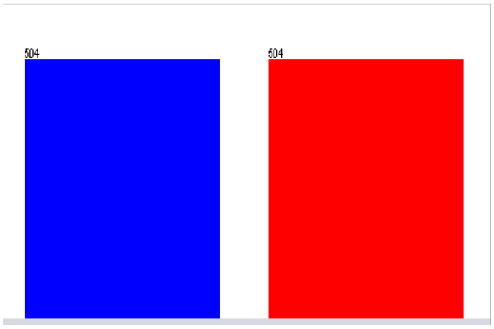

Fig 4: Count Plot shows the ratio of Cancer and non-cancer patient's

4.1. Tool and Language:

In this study, we used the spyder notebook as a tool and python 3.7 as a programming language.

5. CLASSIFICATIONTECHNIQUES:

Building accurate and efficient classifiers for large databases is one of the essential tasks of data mining and machine learning research. Building effective classification systems is one of the central tasks of data mining.

Many different types of classification techniques have been proposed in literature that includes Decision Trees, Naive- Bayesian methods, Random Forest and KNNetc.

5.1 NAÏVE BAYES

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a NaiveBayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

5.2 DECISION TREE ALGORITHM

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

5.3 KNN

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

5.4 RANDOMFOREST

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. There is a direct relationship between the number of trees in the forest and the results it can get:

the larger the number of trees, the more accurate the result. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

6. EXPERIMENTAL RESULT

In this study, the accuracy of four data mining techniques is compared .The paper deals with Naïve Bayes, Decision tree, KNN, Random forest algorithms. Experimental Process is discussed using 4 data sets and the results are compared. The performance analysis is done by considering only the accuracy of the model. The data analysis processed using data mining tool for data analysis, machine learning and statistical learning algorithms.

The training dataset which is used for finding the accuracy consist of 4 data sets with different attributes. The attributes used for training sets for cancer prediction. With the help of these attributes the prediction is done on Testing Process. The performances of these algorithms are evaluated and the results are analyzed. Initially, the data are pre-processed to make the data mining process more efficient.

The accuracy of the algorithms is compared. The pre-processed data is then classified. The algorithms can be applied directly to the data set. Classification is done to know the exactly how data is being classified.

The algorithm which considered to more accurate is considered as the best algorithm and chosen for predicting the test attributes. The below diagrams shows the accuracy of these algorithms from that comparison is done and result is obtained.

6.1 BREAST CANCER:

```
In [1]: runfile('C:/Users/Om/Desktop/data with code/Breast cancer code.py', wdir=
Users/Om/Desktop/data with code')
        id  clump_thickness  unif_cell_size  ...  norm_nucleoli  mitoses  classes
0  1000025                5               1  ...              1        1       no
1  1002945                5               4  ...              2        1       no
2  1015425                3               1  ...              1        1       no
3  1016277                6               8  ...              7        1       no
4  1017023                4               1  ...              1        1       no
```

```
['0' '0' '0' '0' '1' '0' '0' '1' '1' '0' '0' '0' '0' '0' '1' '0'
 '1' '0' '1' '1' '0' '1' '0' '0' '0' '1' '0' '0' '1' '1' '1' '0'
 '1' '0' '0' '1' '0' '0' '0' '0' '0' '0' '1' '1' '0' '0' '1' '1'
 '0' '0' '0' '0' '0' '0' '0' '0' '0' '1' '1' '0' '1' '0' '0' '0'
 '0' '0' '1' '1' '1' '0' '0' '0' '0' '1' '0' '1' '0' '0' '1' '1'
 '1' '1' '0' '1' '1' '0' '0' '0' '0' '1' '0' '0' '0' '1' '1' '1'
 '0' '0' '0' '1' '1' '1' '0' '0' '0' '1' '0' '1' '1' '0' '0' '0'
 '0' '0' '1' '1' '0' '0' '0' '1' '0' '0' '0' '0']
KNN Accuracy:  0.9642857142857143
['0' '0' '1' '0' '1' '0' '0' '1' '1' '0' '0' '0' '0' '0' '1' '0'
 '1' '0' '1' '1' '0' '1' '0' '0' '0' '1' '0' '0' '1' '1' '1' '0'
 '1' '0' '0' '1' '0' '0' '0' '0' '0' '0' '1' '1' '0' '0' '1' '1'
 '0' '0' '0' '0' '0' '0' '0' '0' '0' '1' '1' '0' '1' '0' '0' '0'
 '0' '0' '1' '1' '1' '0' '0' '0' '0' '1' '0' '1' '0' '0' '1' '1'
 '1' '1' '0' '1' '1' '0' '0' '0' '0' '1' '0' '0' '0' '1' '1' '1'
 '0' '0' '0' '1' '1' '1' '0' '0' '0' '1' '0' '1' '1' '0' '0' '0'
 '0' '0' '1' '1' '0' '0' '0' '1' '0' '0' '0' '0']
Navie Bayes Accuracy:  0.9571428571428572
['0' '0' '0' '0' '1' '0' '0' '1' '1' '0' '0' '0' '0' '0' '1' '0'
 '1' '0' '1' '1' '0' '1' '0' '0' '0' '1' '0' '0' '1' '1' '1' '0'
 '1' '0' '0' '1' '0' '0' '0' '0' '0' '0' '1' '1' '0' '0' '1' '1'
 '0' '0' '0' '0' '0' '0' '0' '0' '0' '1' '1' '0' '1' '0' '0' '0'
 '0' '0' '1' '1' '1' '0' '0' '0' '0' '1' '0' '1' '0' '0' '1' '1'
 '1' '1' '0' '1' '1' '0' '0' '0' '0' '1' '0' '0' '0' '1' '1' '1'
 '0' '0' '0' '1' '1' '1' '0' '0' '0' '1' '0' '1' '1' '0' '0' '0'
 '0' '0' '1' '1' '0' '0' '0' '1' '0' '0' '0' '0']
Decision Tree Accuracy:  0.9357142857142857
['0' '0' '0' '0' '1' '0' '0' '1' '1' '0' '0' '0' '0' '0' '1' '0'
 '1' '0' '1' '1' '0' '1' '0' '0' '0' '1' '0' '0' '1' '1' '1' '0'
 '1' '0' '0' '1' '0' '0' '0' '0' '0' '0' '1' '1' '0' '0' '1' '1'
 '0' '0' '0' '0' '0' '0' '0' '0' '0' '1' '1' '0' '1' '0' '0' '0'
 '0' '0' '1' '1' '1' '0' '0' '0' '0' '1' '0' '1' '0' '0' '1' '1'
 '1' '1' '0' '1' '1' '0' '0' '0' '0' '1' '0' '0' '0' '1' '1' '1'
 '0' '0' '0' '1' '1' '1' '0' '0' '0' '1' '0' '1' '1' '0' '0' '0'
 '0' '0' '1' '1' '0' '0' '0' '1' '0' '0' '0' '0']
Random Forest Accuracy:  0.9571428571428572
```

6.2 LUNG CANCER:

```
In [1]: runfile('C:/Users/Om/Desktop/data with code/lu
Users/Om/Desktop/data with code')
  Patient Id  Age  Gender  ...  Dry Cough  Snoring   I
0         P1   33       1  ...          3        4
1        P10   17       1  ...          7        2  Me
2       P100   35       1  ...          7        2
3      P1000   37       1  ...          7        5
4       P101   46       1  ...          2        3
```

```
    0    5   10   15   20   25
[1 0 1 0 1 0 1 1 1 0 1 0 1 1 1 1 0 0 0 0 1 1 0 0 1 0 1 1 1 0 1 0 1 1 1 1 1
 1 1 1 0 0 0 0 0 1 1 1 0 0 1 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 0 0 1 0 1 1 0
 1 0 1 1 1 0 0 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 0 0 0 1 0 0 0 0 0
 0 0 1 0 1 0 1 1 0 0 1 1 1 0 0 0 0 0 0 1 1 1 0 1 0 1 0 0 1 1 0 0 0 0 1 0 1
 0 1 1 1 0 0 1 0 1 0 1 0 0 1 1 0 0 1 1 1 0 0 1 0 1 0 0 1 0 1 1 1 1 1 1 1 0
 1 1 0 0 1 0 1 1 1 0 1 1 0 1 0]
KNN Accuracy:  0.7623762376237624
[1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 0 1 1 0 0 0
 1 0 1 1 0 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0
 1 1 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 1 0 0 0 1 0
 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 0 1
 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 0 0 1
 1 0 0 0 1 0 1 1 1 0 1 1 0 1 1 0]
Navie Bayes Accuracy:  1.0
[1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 0 1 1 0 0 0
 1 0 1 1 0 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0
 1 1 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 1 0 0 0 1 0
 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 0 1
 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 0 0 1
 1 0 0 0 1 0 1 1 1 0 1 1 0 1 1 0]
Decision Tree Accuracy:  1.0
[1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 0 1 1 0 0 0
 1 0 1 1 0 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1 1 0
 1 1 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 1 0 0 0 1 0
 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 0 1
 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 0 0 1
 1 0 0 0 1 0 1 1 1 0 1 1 0 1 1 0]
Random Forest Accuracy:  1.0
```

6.3 LIVER CANCER:

```
In [1]: runfile('C:/Users/Om/Desktop/data with code/liver cancer code.py', wdir='C:/
Users/Om/Desktop/data with code')
   Age  Gender  Total_Bilirubin  ...  Albumin  Albumin_and_Globulin_Ratio  cancer
0   65  Female              0.7  ...      3.3                        0.90     yes
1   62    Male             10.9  ...      3.2                        0.74     yes
2   62    Male              7.3  ...      3.3                        0.89     yes
3   58    Male              1.0  ...      3.4                        1.00     yes
4   72    Male              3.9  ...      2.4                        0.40     yes
```

```
['1' '1' '1' '1' '0' '1' '1' '1' '1' '1' '1' '1' '0' '0' '0' '1' '1' '1'
 '1' '1' '1' '1' '0' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '0'
 '0' '1' '0' '1' '1' '1' '0' '0' '1' '0' '0' '1' '0' '1' '0' '1' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '0' '1' '0' '1' '1' '1' '0' '1'
 '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '0' '1' '1' '1' '0' '1' '1' '0' '1' '1' '1'
 '1' '0' '1' '1' '1' '0' '1' '0']
KNN Accuracy:  0.7155172413793104
['1' '0' '1' '1' '1' '0' '1' '1' '1' '1' '1' '1' '0' '0' '1' '1' '1' '1'
 '0' '1' '1' '1' '1' '0' '0' '1' '1' '1' '1' '1' '1' '1' '0' '1' '1' '1'
 '1' '1' '0' '1' '1' '1' '1' '1' '0' '0' '1' '1' '0' '1' '1' '1' '1' '1'
 '1' '1' '1' '0' '0' '1' '1' '1' '1' '0' '1' '1' '0' '1' '1' '0' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '0' '1' '1' '0' '1' '1' '1' '1' '0' '1' '1'
 '0' '0' '1' '1' '0' '1' '1' '1' '0' '1' '1' '0' '1' '1' '1' '1' '1' '1'
 '1' '0' '1' '1' '0' '1' '1' '0']
Navie Bayes Accuracy:  1.0
['1' '0' '1' '1' '1' '0' '1' '1' '1' '1' '1' '1' '0' '0' '1' '1' '1' '1'
 '0' '1' '1' '1' '1' '1' '0' '0' '1' '1' '1' '1' '1' '1' '1' '0' '1' '1'
 '1' '1' '0' '1' '1' '1' '1' '1' '0' '0' '1' '1' '0' '1' '1' '1' '1' '1'
 '1' '1' '1' '0' '0' '1' '1' '1' '1' '0' '1' '1' '0' '1' '1' '0' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '0' '1' '1' '0' '1' '1' '1' '1' '0' '1' '1'
 '0' '0' '1' '1' '0' '1' '1' '1' '0' '1' '1' '0' '1' '1' '1' '1' '1' '1'
 '1' '0' '1' '1' '0' '1' '1' '0']
Decision Tree Accuracy:  1.0
['1' '0' '1' '1' '1' '0' '1' '1' '1' '1' '1' '1' '0' '0' '1' '1' '1' '1'
 '0' '1' '1' '1' '1' '1' '0' '0' '1' '1' '1' '1' '1' '1' '1' '0' '1' '1'
 '1' '1' '0' '1' '1' '1' '1' '1' '0' '0' '1' '1' '0' '1' '1' '1' '1' '1'
 '1' '1' '1' '0' '0' '1' '1' '1' '1' '0' '1' '1' '0' '1' '1' '0' '1' '1'
 '1' '1' '1' '1' '1' '1' '1' '0' '1' '1' '0' '1' '1' '1' '1' '0' '1' '1'
 '0' '0' '1' '1' '0' '1' '1' '1' '0' '1' '1' '0' '1' '1' '1' '1' '1' '1'
 '1' '0' '1' '1' '0' '1' '1' '0']
Random Forest Accuracy:  1.0
116
```

6.4 CANCER / NON CANCER:

```
In [1]: runfile('C:/Users/Om/Desktop/data with code/live data cancer and non
cancer.py', wdir='C:/Users/Om/Desktop/data with code')
   id address  age  ... mosquito repellents   think cancer/non cancer
0   1     vsp   50  ...                 yes  medium              yes
1   2     vsp   50  ...                 yes  medium              yes
2   3     vsp   40  ...                 yes    more              yes
3   4     egd   55  ...                 yes  medium              yes
4   5    sklm   28  ...                 yes  medium              yes
```

```
[1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 1 0 1 1 1 0 1 0 1 1 1 1 1
 1 1 1 0 0 0 0 0 1 1 1 0 0 1 0 1 0 1 1 0 0 0 0 0 1 0 0 1 0 1 0 0 1 0 1 1 0
 1 0 1 1 1 0 0 1 1 1 0 0 1 1 1 0 1 1 1 0 1 1 1 1 1 1 0 1 1 0 0 1 1 0 0 0 1 0
 0 0 1 0 1 0 1 1 0 0 1 1 0 0 0 0 0 1 1 1 0 1 0 1 0 0 1 1 1 0 0 0 0 1 0 1 1 1 1
 0 1 1 1 0 0 1 0 1 0 1 0 0 1 1 0 0 1 1 1 0 0 1 0 0 1 0 1 1 1 1 1 1 1 1 0 1 1
 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 1 0]
KNN Accuracy:  0.7623762376237624
[1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 0 1 1 0 0 0
 1 0 1 1 0 0 0 0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0 0 0 0 0 1 1 1 1 0 1 1 0
 1 1 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 1 0 0 0 0 1 0
 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 1 0 1
 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 0 0 1 1
 1 0 0 0 1 0 1 1 0 1 1 1 0 1 0]
Navie Bayes Accuracy:  1.0
[1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 0 1 1 0 0 0
 1 0 1 1 0 0 0 0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0 0 0 0 0 1 1 1 1 0 1 1 0
 1 1 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 1 0 0 0 0 1 0
 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 1 0 1
 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 0 0 1 1
 1 0 0 0 1 0 1 1 0 1 1 1 0 1 0]
Decision Tree Accuracy:  1.0
[1 0 1 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 1 0 0 1 1 0 1 0 1 1 0 0 0
 1 0 1 1 0 0 0 0 1 1 0 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 0 0 0 0 0 1 1 1 1 0 1 1 0
 1 1 1 0 1 1 0 0 1 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 0 1 1 0 1 0 1 0 1 0 0 0 0 1 0
 0 0 1 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 1 1 0 1
 0 1 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 1 1 1 1 0 0 1 1
 1 0 0 0 1 0 1 1 0 1 1 1 0 1 0]
Random Forest Accuracy:  1.0
```

	Lung Cancer	Breast Cancer	Liver Cancer	Live Data
Naïve Bayes	0.89	0.95812	0.55746	0.9
Decision Tree	1.0	0.93571	1.0	1.0
KNN	0.995	0.96428	0.71551	0.762
Random Forest	1.0	0.95714	1.0	1.0

7 CONCLUSION

The accuracy of classification techniques is evaluated based on the selected classifier algorithm. An important challenge in data mining and machine learning areas is to build precise and computationally efficient classifiers for Medical applications. With the help of data mining, people can know the possible outcome of cancer prediction.

This project describes the data mining algorithms such as Naïve Bayes, Random forest, Decision tree and KNN are used compared. The performance of Random Forest shows the high level compare with other classifiers. Therefore Random Forest classifier is suggested for prediction of any Cancer disease based classification to get better results with accuracy and performance.

References

[1]. Shah, C., & Jivani, A. G. (2013, July). Comparison of data mining classification algorithms for breast cancer prediction. In 2013 Fourth international conference on computing, communications and networking technologies (ICCCNT) (pp. 1-4).IEEE.

[2]. Mayilvaganan, M., & Thamaraikannan, N. (2019). PERFORMANCE COMPARISION OF DATA CLASSIFICATION ALGORITHM FOR ANALYSIS LUNG CANCER DATASETS.

[3]. Bin Othman, M. F., &Yau, T. M. S. (2007). Comparison of different classification techniques usingWEKAforbreastcancer.In3rdKualaLumpur International Conference on BiomedicalEngineering 2006 (pp. 520-523). Springer, Berlin,Heidelberg.

[4]. Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for lung cancer risk prediction and diagnosis. Procedia Computer Science, 83, 1064- 1069.

[5]. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2),113-127.

[6]. Sarvestani, A. S., Safavi, A. A., Parandeh, N.M., &Salehi,M.(2010,October).Predictinglungcancer survivability using data mining techniques. In 2010 2nd International Conference on Software Technology and Engineering (Vol. 2, pp. V2-227). IEEE.

[7]. Ramachandran,P., Girija,N.,&Bhuvaneswari,T. (2014). Early detection and prevention of cancer using data mining techniques. International Journal of Computer Applications,97(13).

[8]. Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. ArXiv preprint arXiv:1205.1923.

[9]. Nakte, J., & Himmatramka, V. (2016). Breast cancer prediction using Data mining techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 4(11), 55-60.