IMPROVED LUNG CANCER DETECTION USING MACHINE LEARNING TECHNIQUES

A dissertation

Submitted in fulfillment of the requirements for the award of the degree of

MASTER OF TECHNOLOGY

IN

INFORMATION SYSTEMS

Submitted by:

PUNEET

(2K18/ISY/08)

Under the Supervision of

MS. ANAMIKA CHAUHAN

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi, India-110042

JUNE, 2020

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi, India-110042

CANDIDATE'S DECLARATION

I, **PUNEET**, Roll No. **2K18/ISY/08** student of M.Tech Information Systems, hereby declare that the project Dissertation titled **"Improved Lung Cancer detection using Machine learning Techniques"** which is submitted by me to the Department of Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of **Master of Technology**, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi **PUNEET**

Date: 29-06-2020 **2K18/ISY/08**

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi, India-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled **"Improved Lung Cancer detection using Machine learning Techniques"** which is submitted by **PUNEET**, Roll No. **2K18/ISY/08** Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of **Master of Technology**, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi **MS. ANAMIKA CHAUHAN**
Date: **SUPERVISOR**

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi, India-110042

ACKNOWLEDGEMENT

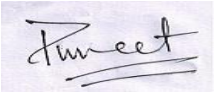
It is a great pleasure for me to express my respect and deep sense of gratitude to my Supervisor **Ms. Anamika Chauhan**, Assistant Professor, Department of Information Technology, Delhi Technological University, for his wisdom, vision, expertise, guidance, enthusiastic involvement and persistent encouragement during the planning and development of this research work. I also gratefully acknowledge his painstaking efforts in thoroughly going through and improving the manuscripts without which this work could not have been completed.

I am highly obliged to **Dr. Kapil Sharma**, Head of the Department, Department of Information Technology, for providing all the facilities, help and encouragement for carrying out the research work.

I am obliged to my parents **Krishan Kumar** and **Raj Dulari** for their moral support, love, encouragement and blessings to complete this task.

I wish to express my appreciation to my friends **Rajat Mahey**, **Nitin Bhagwat** and **Chandan Kumar** and grateful thanks to research fellows at department for their help and motivation throughout my research work. I also would like to express my deep and sincerely thanks to my friends and all other persons whose names do not appear here, for helping me either directly or indirectly in all even and odd times.

I would also like to extend my special thanks to staff members of IT Department, for their timely help and cooperation extended throughout the course of investigation. Finally, I am indebted and grateful to the Almighty for helping me in this endeavor.



Place: Delhi **PUNEET**
Date: 29-06-2020 **2K18/ISY/08**

ABSTRACT

Lung Cancer have become one of the most common cause of death among human beings. Many human beings die early because of lung cancer. The early detection of lung cancer is tough due to the structure of cancer cells and less awareness among human beings. Diagnosis of lung cancer is done using various tests like imaging tests, sputum cytology and biopsy, which are costly and time taking. Classifying lung cancer is not an easy task and needs experienced physicians and a lot of money. Cancer recurrence in recovered patients leads to a high cost, and not everyone can afford it. We have used lung cancer data from Wu, Jiangpeng et al. [11] as it is an unbalanced dataset. We used various evaluation parameters like Accuracy, Confusion matrix, AUC- ROC [6] and FNR, which gives us more insight. XGBoost provides the best accuracy of 92.16% for lung cancer. We have used various Machine Learning classification techniques under the library of scikit-learn like KNN, Logistic Regression, XGBoost, Gaussian Naive Bayes, Decision Tree and SVM. Different algorithms under the library of scikit learn have been used to select the features, and only those features that are important to our model are selected. Our models achieved more accuracy score and sensitivity than Wu, Jiangpeng et al. [11] for lung cancer. Parameter tuning has helped to improve the performance of the model. In various phases of machine learning pipeline, we have improved the result of Wu, Jiangpeng et al. [11]. Starting from preprocessing to the development of the model.

KEYWORDS

Cancer, Classification Techniques, Lung Cancer Classification, Dataset, Machine Learning Techniques, Machine Learning, Scikit-Learn Algorithms

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

ML = Machine Learning

PCA = Principal Component Analysis

ICA = Independent Component Analysis

NSCLC = Non-Small Cell Lung Cancer

GBM = Gradient Boosting Machines

SEER = Surveillance, Epidemiology, and End Results

EDA = Exploratory Data Analysis

1-D = One Dimensional

2-D = Two Dimensional

PDF = Probability Density Function

CDF = Cumulative Distribution Function

SVM = Support Vector Machine

KNN = K- Nearest Neighbors

ILCD = Improved Lung Cancer Detection

RBLC = Routine Blood Indices Model for Lung Cancer

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

TPR = True Positive Rate

FPR = False Positive Rate

TNR = True Negative Rate

FNR = False Negative Rate

MR = Miss Rate

Acc = Accuracy

Sen = Sensitivity

Spec = Specificity

AUC= Area under the Curve

ROC = Receiver Operating Characteristic

MCC = Matthews Correlation Coefficient

μ = Mean

σ = Variance

% = Percentage

P (n, r) = Permutation formula

CHAPTER 1

INTRODUCTION

There is an abundance of data in this world. When humans use the internet to communicate from one person to another, sitting in one corner of the world to the other- data is created. Data means pictures, videos, text, spreadsheets, music, word file, tweets etc. Lots of data is being generated in the world by human being, smart IoT devices and other electronic devices. In future, increase in IoT devices and the use of 5G network will lead to rapid growth of data.

Traditionally human have analyze data manually, and have tried to create system that adapts to change in data patterns. However, this approach is not sustainable because of the volume of data and the failure to comprehend and create rules for such data. Today we have shifted to automated systems that can learn from the data and adjust to the changes in the data environment.

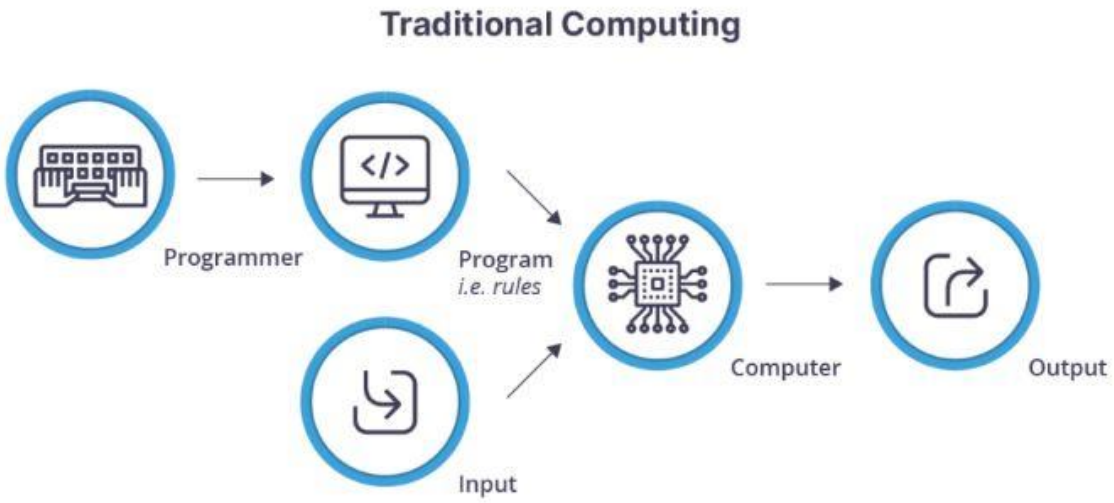


Figure 1 Traditional Computing diagram [34]

Machine learning is used to derive meanings or patterns from the data. Arthur C. Clarke used to say: "Any sufficiently advanced technology is indistinguishable from magic".

In Machine Learning we analyze data to make predictions or answer questions related to that data. In Machine Learning we first use training data for the creation of rules and learning patterns present in the data. Using this training data, models are built and these models are used to predict and answer questions for new data in the future. As more data is gathered, the model can be improved over time, and new predictive models can be deployed.

All of the Machine Learning process is dependent on data. Machine learning is the learning of patterns or identifying connections from the data to answer questions or make predictions.

Machine Learning is a tool, technology and algorithms that can be utilized to answer or can be utilized to predict the answer of a given question based on the data.

One of the biggest example of machine learning in practice is Google Search. Google Search incorporates many machine learning system from analyzing and understanding the query of user in form of text or voice to personalizing users' search result and ads.

Today machine learning applications are widely used in many areas like image detection, Cancer detection, weather prediction and fraud detection etc.

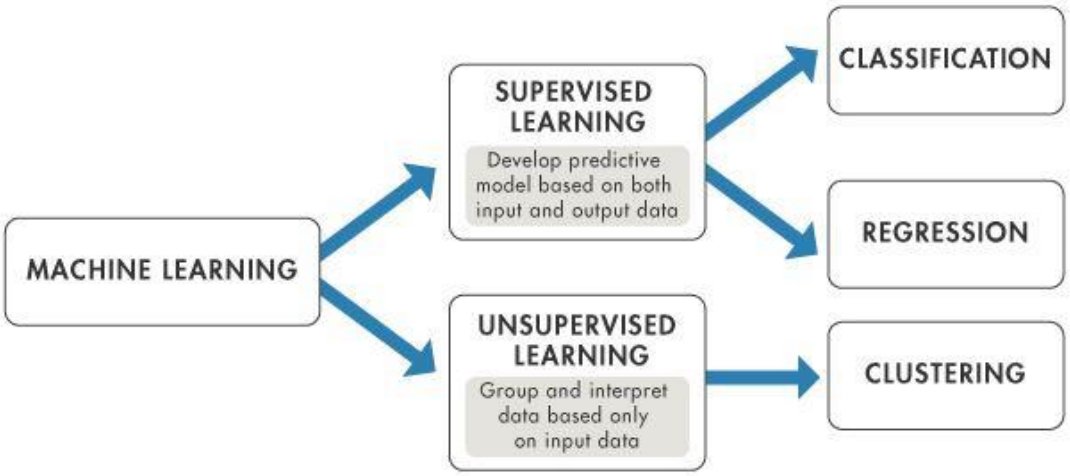


Figure 2 Machine Learning diagram [35]

There are usually three ways by which machines can learn:

1.1 Supervised Learning

Supervised Learning is one of the area of ML. In Supervised Learning, learning happens under the supervision of a teacher. In this we have a dataset which has class-label, i.e. we have a sample dataset from which algorithm can learn.

The dataset consists of train and test sets. Train set helps us to train the model, i.e. capture the patterns from the dataset known as fitting or training the model. Once the model had been trained, it is then applied to new data for validation.

1.2 Unsupervised Learning

Unsupervised Learning is one of the area of ML. In Unsupervised Learning, learning happens without the supervision of a teacher. In this, we have a dataset which doesn't have class-label. PCA, ICA, clustering etc. are some of the unsupervised learning approaches. Market segmentation, Astronomical data analysis, and Social network analysis are some of the examples in which unsupervised learning is used.

1.3 Reinforcement Learning

Reinforcement Learning is one of the area of ML. Reinforcement Learning is about taking effective action to maximize rewards in a given situation. It is used by software and machines to find the best course of action or direction to take in a specific situation. Reinforcement learning differs from supervised learning because in this we don't have the response key for the training data which is there in supervised learning. In reinforcement learning, the reinforcement agent decides what is needed to achieve the task and maximize the rewards. It is required to learn from its past, in the absence of a training dataset.

1.4 Background

Lung cancer is the uncontrolled growth of abnormal cells in one or both of the lungs [15]. These abnormal cells do not develop into healthy lung tissue and hence could not perform the normal lung cell functions. These abnormal cells can form tumors as they expand and interfere with the functioning of the lung, which is to provide the body with oxygenated blood. Lung Cancer is very common among humans and causes a large number of deaths every year. We can reduce the deaths caused by Lung Cancer by detecting it in the earlier stage. Using a blood test, we can detect lung cancer in the early stage, and chances may increase to save human life [1]. Various tests like imaging tests, sputum cytology and biopsy are used to detect lung cancer [2].

Cancer Research has made steady progress in the last few decades [7]. About 18.1 million [5] new patients are diagnosed with various subtypes of cancer in 2018 and are projected to grow to 23.6 million by 2030 [3, 4]. Prediction of cancer subtypes in early stages can help the doctors to give the patients proper medication and treatment. If cancer stages are predicted accurately early, then the patients will have chances of being cured. Underdeveloped and developing countries are the worst affected because of the lack of sufficient medical services leading to high mortality rates relative to developed countries.

Ignorance of cancer symptoms and failure to seek proper early-stage medical consultation or late seeking medical advice leads to a higher mortality rate.

The lack of physicians and medical equipment in rural areas also adds to the problem of early-stage diagnosis of cancer disease, leading to higher mortality rates. Machine learning approaches are commonly used for identification and classifying different diseases in medical fields. Machine learning methods are now also being used in medical field areas to predict cancer early, which can lead to better chances of patients' survival.

Doctors and scientists have applied various methods, such as screening, detection, and classification to diagnose different types of cancer even before they cause any symptoms. Scientists have also developed various new methods to predict the outcome of cancer treatment in the early stage [8]. A significant amount of cancer data were collected and published for medical research with the emergence of modern medical technologies. Predicting precisely outcome of a disease is one of the most difficult tasks for physicians. Due to this Machine Learning (ML) approaches is becoming a best tool in medical research and can also help in predicting potential outcomes of different cancer types.

CHAPTER 2

LITERATURE REVIEW

Wu, Jiangpeng et al. [11] has identified Lung Cancer based on Routine Blood Indices. Blood-based liquid biopsies are widely recognized as a method for cancer monitoring tool and diagnostic for cancer detection. Immensely high sensitivity is also required because of the very low levels of correctly identified protein biomarkers, RNA or DNA released into the blood. Doctors also order regular tests of blood indices because they are easy to manage and cost-efficient. Machine learning models like Random Forest has been adopted to create a model of identification between lung cancer and routine blood indices that would decide if they are potentially likely linked.

Leng, Shaoyi et al. [9] had used the integrity of cell-free DNA to identify patients with lung cancer. The cfDNA levels and cfDNA integrity were significantly higher in patients with NSCLC than in patients with tuberculosis. Also, cfDNA and its integrity may be used as markers to classify tuberculosis-related NSCLC.

Fradkin, Dmitriy et al. [10] used SVM and penalized logistic regression to build a model that can predict the survival of patients diagnosed with lung cancer and to analyze the importance of model parameter based features.

Lynch, Chip M. et al. [14] applies unsupervised machine-learning clustering and classification techniques to a group of lung cancer patients. The aim is to automatically classify lung cancer patients into groups on the basis of clinically measurable disease-specific variables to estimate the chances of their survival.

Dimitoglou, George et al. [13] used the Naïve Bayes classifier and C4.5 algorithm for the prediction of survivability of lung cancer patients.

Lynch, Chip M. et al. [12] used supervised Machine Learning techniques like Decision Trees, GBM, Custom ensemble, Linear regression and SVM which is applied on the SEER data to classify patients with lung cancer in terms of survival. Main attributes of the data when applying these methods include gender, age, stage, tumor grade, tumor size and number of primaries, with the objective of comparing predictive power between the different methods.

CHAPTER 3

SOURCE OF DATA

The source of the dataset for the cancer patients is taken from the research paper Wu, Jiangpeng et al. [11] collected at Second Hospital, China, Lanzhou University. There are 277 patients in total in the dataset, with 49 different forms of blood indices for each patient in the dataset.

	Platelet distribution width	Basophil ratio	Albumin	Age	Large platelet ratio	Lymphocyte ratio	ALB/GLB	Neutrophile granulocyte ratio	White blood cell	Creatine kinase isoenzymes
count	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000	277.000000
mean	13.407942	0.003383	39.752708	56.862816	27.350181	0.240296	1.224296	0.671097	6.663466	20.754152
std	3.329155	0.004651	5.211571	13.243646	11.084908	0.102024	0.341425	0.113945	3.006356	32.663335
min	0.000000	0.000000	24.100000	20.000000	0.000000	0.030000	0.590000	0.270000	2.180000	3.000000
25%	10.700000	0.000000	36.100000	50.000000	19.200000	0.170000	0.980000	0.600000	4.790000	11.000000
50%	14.000000	0.000000	40.100000	59.000000	26.000000	0.230000	1.180000	0.660000	5.820000	14.000000
75%	16.200000	0.009000	43.100000	66.000000	33.800000	0.310000	1.380000	0.740000	7.740000	21.000000
max	20.400000	0.020000	51.600000	81.000000	61.200000	0.600000	2.660000	0.930000	24.730000	383.000000

Figure 3 Dataset description

In Figure 3, description of the dataset is of features Platelet distribution width, Basophil ratio, Albumin, Age, Large platelet ratio, Lymphocyte ratio, ALB/GLB, Neutrophile granulocyte ratio, White blood cell and Creatine kinase isoenzymes. The description of the dataset includes count, mean, std, min, 25%, 50%, 75% and max.

	Label	Count
0	1	183
1	-1	94

Figure 4 Dataset count

Of the 277 patients, 183 were positive for lung cancer. The tissue biopsies were used to diagnose these patients. The other remaining 94 patients have been diagnosed as not having lung cancer. We can see from Figure 4 that our dataset is imbalanced.

The imbalanced dataset is those datasets where you have an unequal number of data points for each of your classes. For example- In our Cancer dataset, which are having two classes and the value count for two classes are:

Lung Cancer	183
Without Lung Cancer	94

The Imbalanced data set somehow look like above.

The **balanced dataset** is those datasets where you have an almost equal number of data points for each of your classes. For example - We had a flower dataset which is having three species and the value count for three species are:

Versicolor	50
Virginica	50
Setosa	50

The balanced data set somehow look like above.

Figure 4 shows Label as one of the columns of the dataset, '1' means having Lung Cancer which is 183 patients, and '-1' means not having Lung Cancer which is 94 patients. These patients include females and males, with age, ranges between 20 and 81 years. Such positive or '1' patients find themselves at a different level of cancer. Each patient's Smoking status was also reported into the dataset.

Out of the 94 patients, 51 patients are suffering from tuberculosis. These tuberculosis patients were especially included because the use of CT scans to differentiate lung cancer from tuberculosis has a high false-positive rate. The remaining patients went for regular checkups', and they were not detected with any disease related to lung cancer. The dataset [11] contains the column 'Dataset' which had entries as the training set and test set. The training set is consist of 226 patients out of which 153 were patients with Lung Cancer, 37 patients with tuberculosis and 36 were patients

with other diseases. 51 patients are from test data.

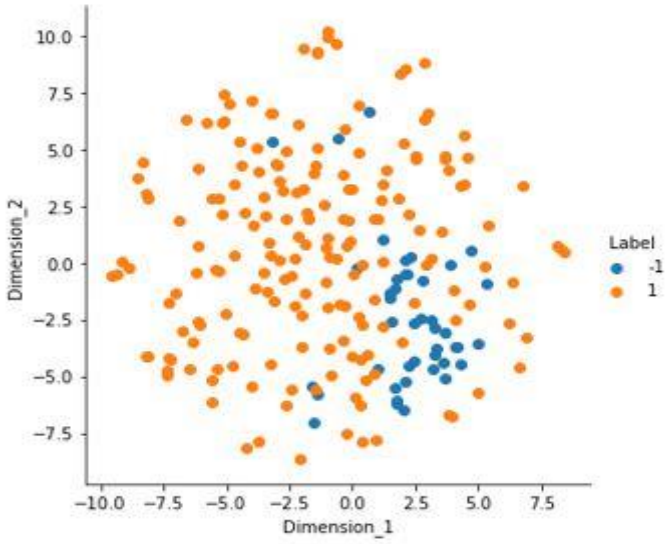


Figure 5 Data visualization in 2-D using TSNE

3.1 EXPLORATORY DATA ANALYSIS (EDA)

EDA understands what dataset uses simple statistical methods, graphs, visualizes data before we actually go and start machine learning.

Using statistical analysis and visualization tools, EDA is used for data analysis to understand the data and to use visualization tools.

There are 58 columns out of which 50 will be used as our features in our feature engineering. Feature engineering is essential as it helps to improve the performance of the model.

Out of the 50 features, I have applied ExtraTreeClassifier method as a feature ranking for the ranking of 50 features in decreasing order. I have used Univariate Exploratory Data Analysis and Bivariate Exploratory Data Analysis in feature engineering on our 50 features but I will be taking here top 5 ranked features as I couldn’t be able to show all here.

3.1.1 UNIVARIATE EXPLORATORY DATA ANALYSIS

In Univariate EDA, we only take one variable or feature into consideration at a time. Using SelectKBest method, it was applied to the top 5 features of the dataset [11] obtained. We graphed the plot 1-D Scatter plot, Histogram and PDF, CDF, Boxplot and Violin plot into Univariate Analysis.

1-D Scatter Plot

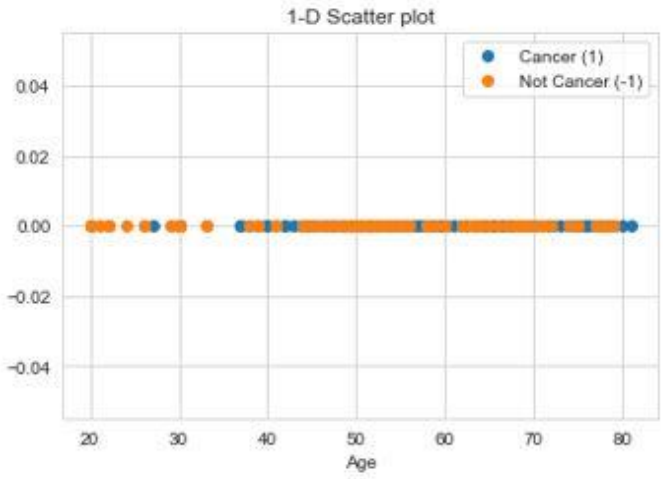


Figure 6 1-D Scatter plot of Age

OBSERVATIONS

- Very hard to make sense, as points are overlapping a lot.
- Hard to know how many points are there.

Histogram and Probability density function (PDF)

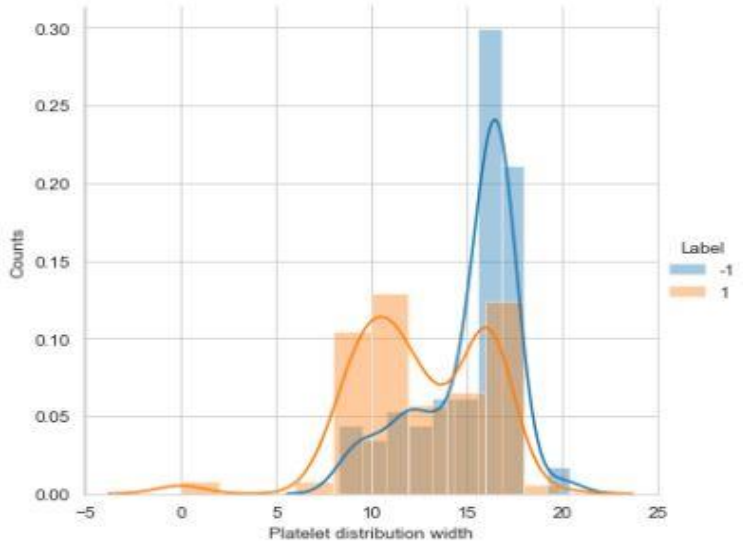


Figure 7 Histogram and PDF of Platelet distribution width

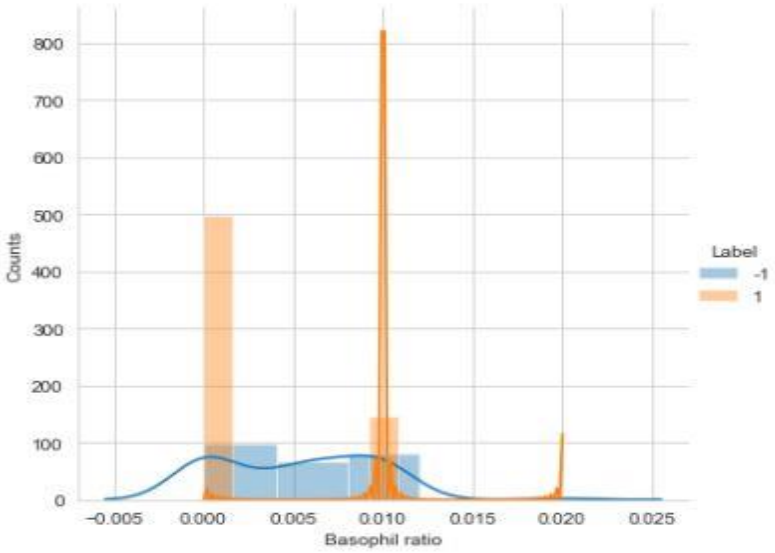


Figure 8 Histogram and PDF of Basophil ratio

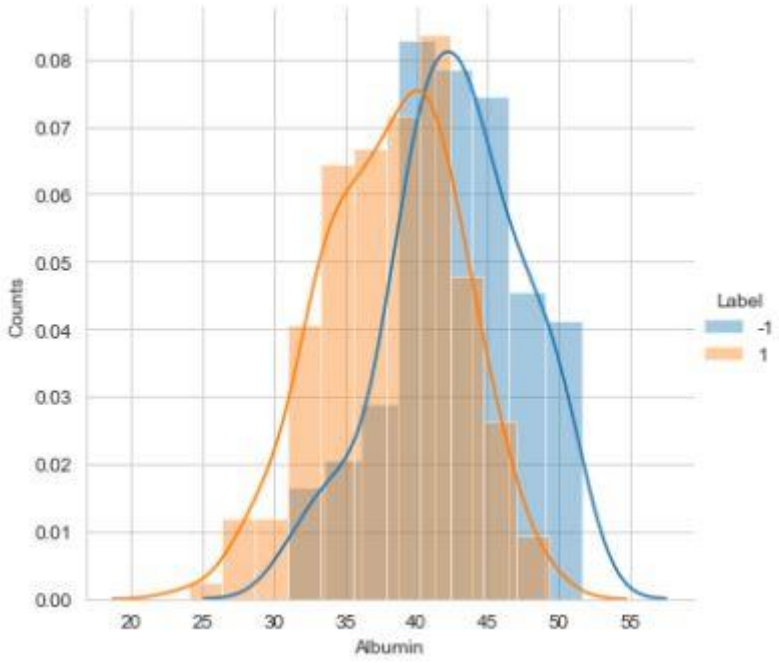


Figure 9 Histogram and PDF of Albumin

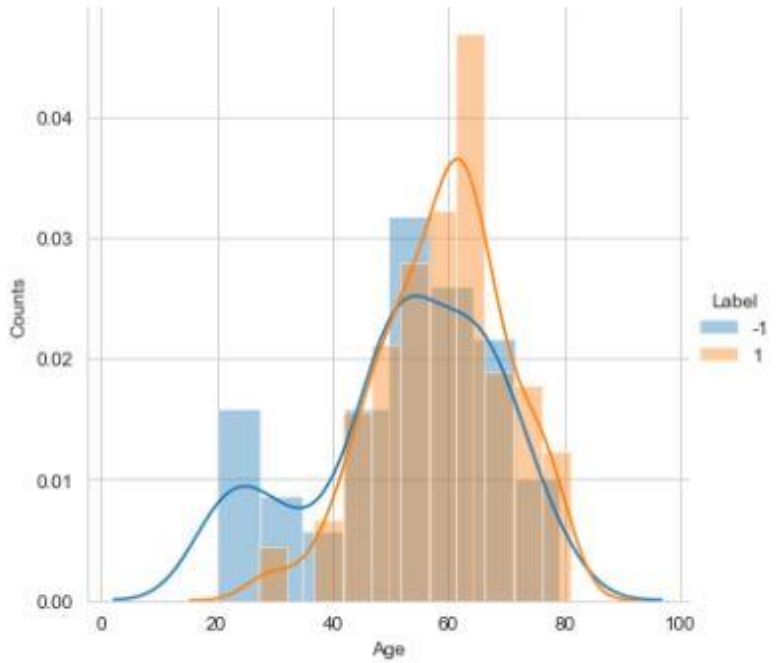


Figure 10 Histogram and PDF of Age

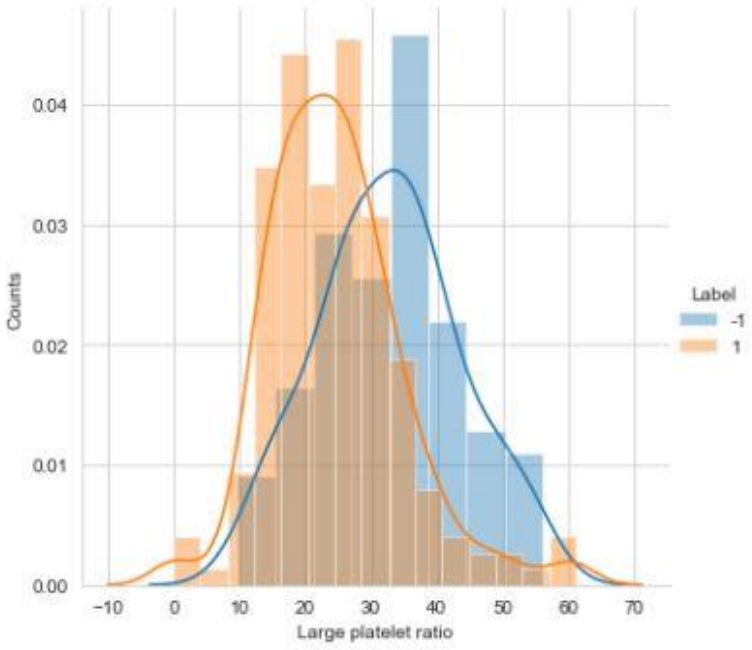


Figure 11 Histogram and PDF of Large platelet ratio

OBSERVATIONS

As can be seen from the above figures, the cancer and non can cancer patients have almost overlapping probability distribution. Therefore these features are not capable or sufficient enough to use as standalone features to classify patients.

Cumulative distribution function (CDF)

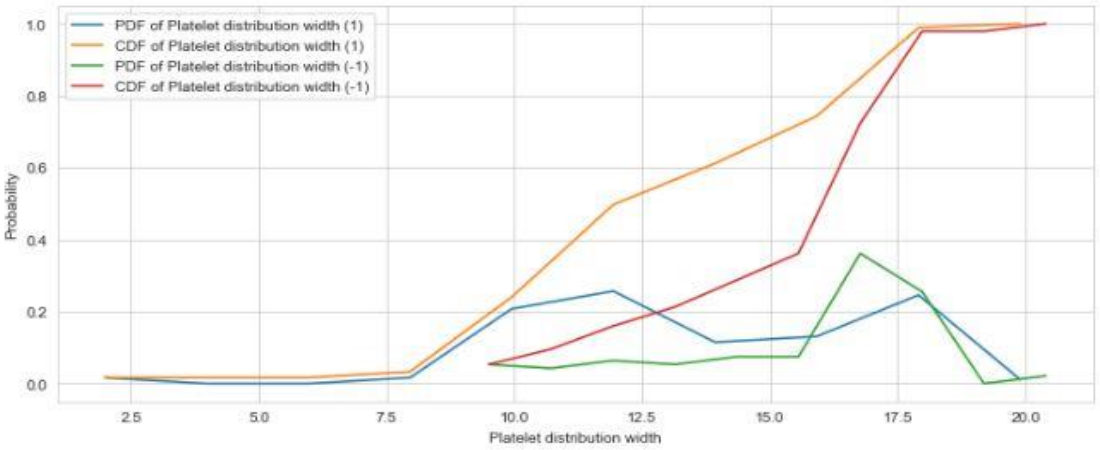


Figure 12 PDF and CDF of Platelet distribution width

OBSERVATIONS

- There are 70% of cancer patients' values that have Platelet distribution width ≤ 15.0 as evident by CDF.
- There are 100% of cancer patients' values that have Platelet distribution width < 20.0 as evident by CDF.
- There are 80% of non-cancer patients' values that have Platelet distribution width ≤ 17.2 as evident by CDF.
- 100% of non-cancer patients' have Platelet distribution width ≤ 21.0 as evident by CDF.

After reading the plot, we observe that the patients whose platelet distribution width is less than 12.0 are more likely to have lung cancer, as 55% approx. of the cancerous patients lie between that range whereas very few, i.e. 16% approx. of non- cancerous patients lie in that range.

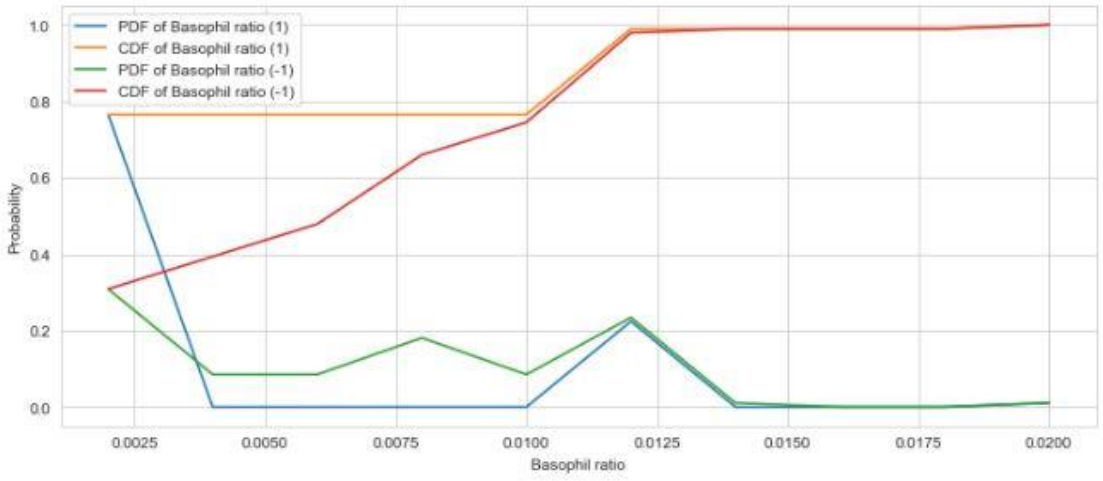


Figure 13 PDF and CDF of Basophil ratio

OBSERVATIONS

- There are 78% of cancer patients' values that have Basophil ratio ≤ 0.0100 as evident by CDF.
- There are 100% of cancer patients' values that have Basophil ratio ≤ 0.0200 as

evident by CDF.

- There are 75% of non-cancer patients' values that have Basophil ratio ≤ 0.0100 as evident by CDF.
- 100% of non-cancer patients' have Basophil ratio ≤ 0.0200 as evident by CDF.

After reading the plot, we observe that the patients whose basophil ratio is less than 0.0025 are more likely to have lung cancer, as 78% approx. of the cancerous patients lie between that range whereas, i.e. 33% approx. of non- cancerous patients lie in that range.

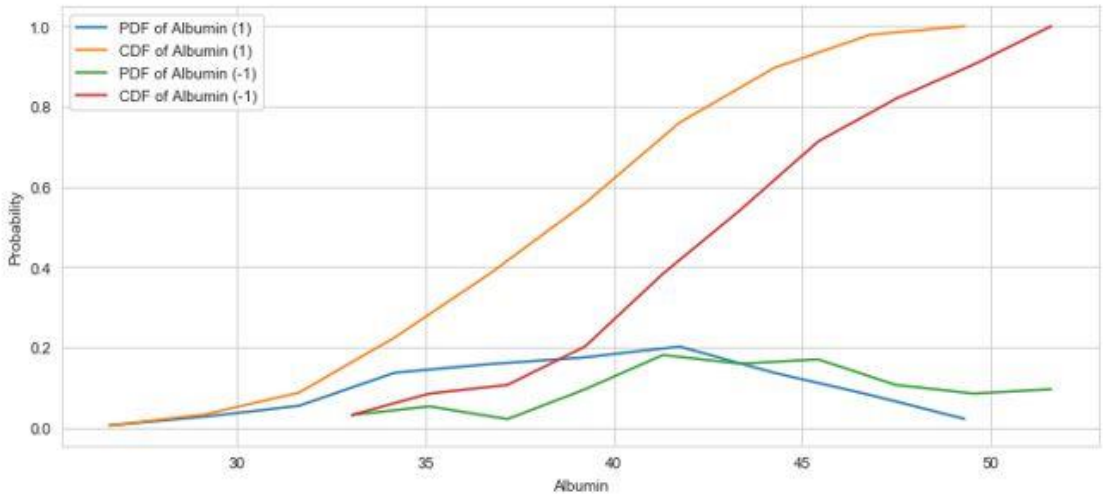


Figure 14 PDF and CDF of Albumin

OBSERVATIONS

- There are 62% of cancer patients' values that have Albumin ≤ 40 as evident by CDF.
- There are 100% of cancer patients' values that have Albumin ≤ 49 as evident by CDF.
- There are 70% of non-cancer patients' values that have Albumin ≤ 45 as evident by CDF.
- 100% of non-cancer patients' have Albumin ≤ 52 as evident by CDF.

After reading the plot, we observe that the patients whose albumin is less than 37 are more likely to have lung cancer, as 56% approx. of the cancerous patients lie between that range whereas very few, i.e. 18% approx. of non- cancerous patients lie in that range.

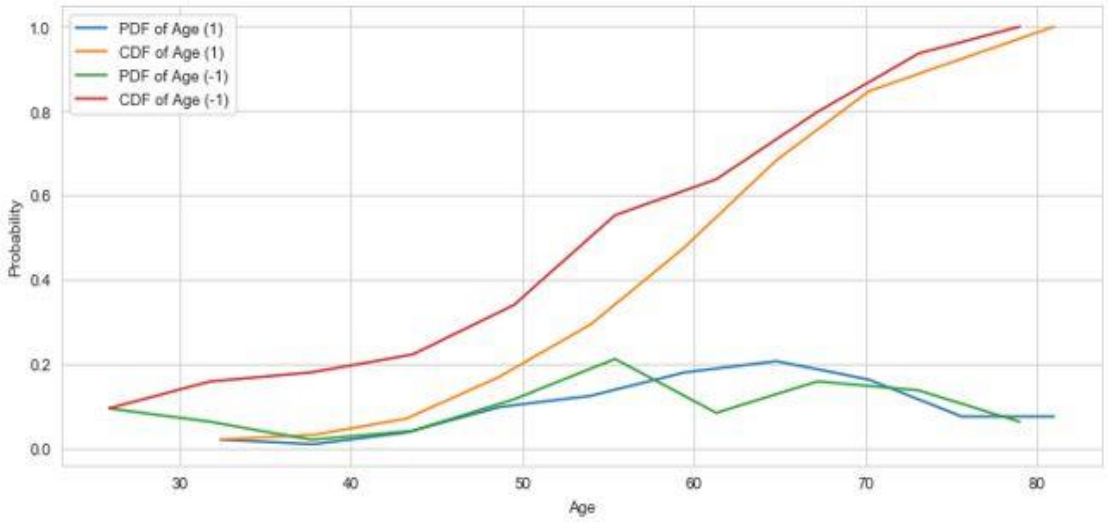


Figure 15 PDF and CDF of Age

OBSERVATIONS

- There are 50% of cancer patients' values that have Age ≤ 60 as evident by CDF.
- There are 100% of cancer patients' values that have Age ≤ 82 as evident by CDF.
- There are 62% of non-cancer patients' values that have Age ≤ 60 as evident by CDF.
- 100% of non-cancer patients' have Age ≤ 79 as evident by CDF.

After reading the plot, we observe that the patients whose age is less than 56 are more likely to have lung cancer, as 40% approx. of the cancerous patients lie between that range whereas, i.e. 58% approx. of non- cancerous patients lie in that range.

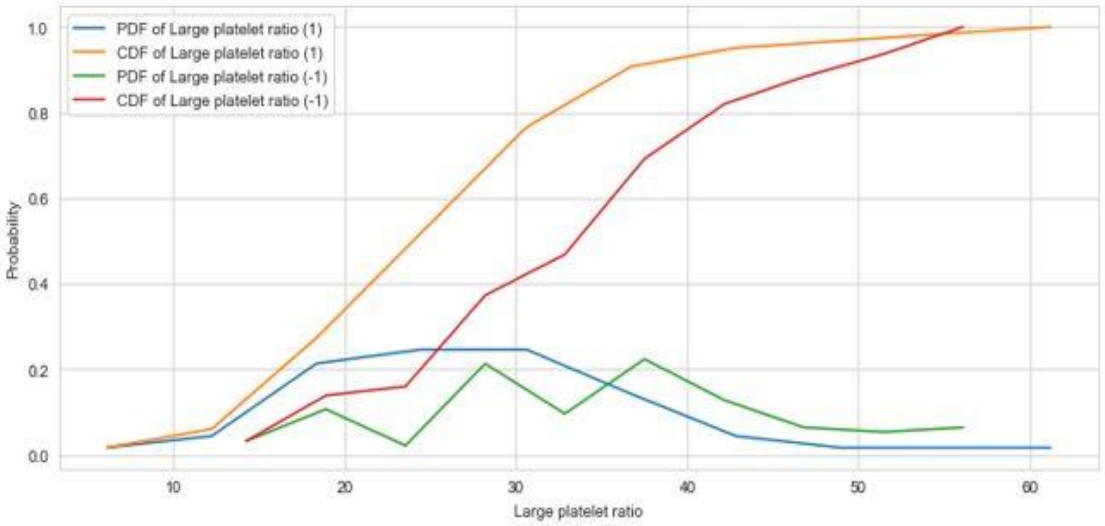


Figure 16 PDF and CDF of Large platelet ratio

OBSERVATIONS

- There are 77% of cancer patients' values that have large platelet ratio ≤ 30 as evident by CDF.
- There are 100% of cancer patients' values that have large platelet ratio ≤ 62 as evident by CDF.
- There are 60% of non-cancer patients' values that have large platelet ratio ≤ 36 as evident by CDF.
- 100% of non-cancer patients' have large platelet ratio ≤ 57 as evident by CDF.

After reading the plot, we observe that the patients whose large platelet ratio is less than 22 are more likely to have lung cancer, as 40% approx. of the cancerous patients lie between that range whereas very few, i.e. 17% approx. of non- cancerous patients lie in that range.

Boxplot

Boxplot can be visualized as a PDF on the side ways. It is another method of visualizing the 1-D scatter plot with the concept of median, percentile and quantiles.

A technique called Inter-Quartile range (IQR) is used in plotting the whiskers. While histograms are very good to understand the density or how many points exists in a range but it can't tell us what 25th, 50th (Median), 75th and 100th percentiles values. CDF graphs still can tell us the percentile values but Boxplot gives clear view for this.

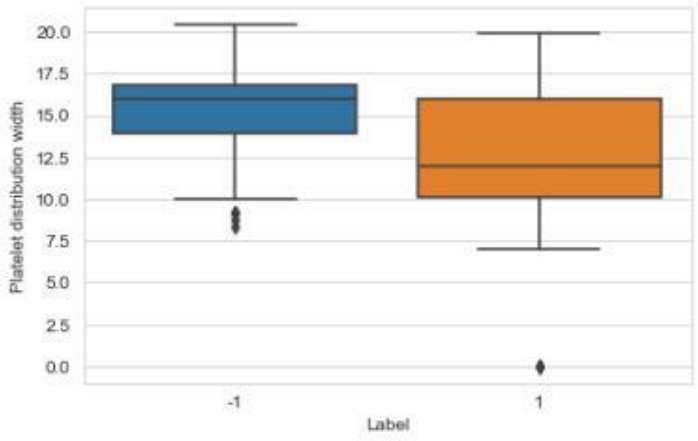


Figure 17 Boxplot of Platelet distribution width

OBSERVATIONS

- 25th percentile of the cancer patients' values are having Platelet distribution width ≤ 10.
- 50th percentile (Median) of the cancer patients' values are having Platelet distribution width ≤ 12.
- 25th percentile of the non-cancer patients' values are having Platelet distribution width ≤ 14.
- 50th percentile (Median) of the non-cancer patients' values are having Platelet distribution width ≤ 16.

After reading the plot, we can say that if we choose large Platelet distribution width ≤ 15.0 as our threshold for classification then we can see that for class-label '-1' more than 50% of the values have level greater than this threshold and for class label '1' more than 65% of the values less than this threshold.

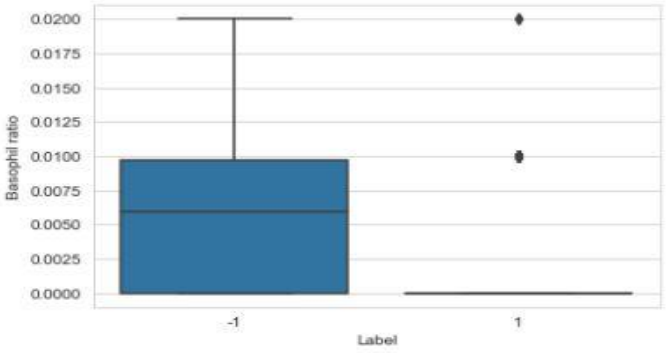


Figure 18 Boxplot of Basophil ratio

OBSERVATIONS

- 25^{th} percentile of the non-cancer patients' values are having Basophil ratio ≤ 0.
- 50^{th} percentile (Median) of the non-cancer patients' values are having Basophil ratio ≤ 0.0060.
- 75^{th} percentile of the non-cancer patients' values are having Basophil ratio ≤ 0.0095.
- 100^{th} percentile of non-cancer patients' have Basophil ratio ≤ 0.0200 as evident by CDF.

After reading the plot, we can say that if we choose large Basophil ratio ≤ 0 as our threshold for classification then we can see that for class-label '-1' more than 75% of the values have level greater than this threshold and for class label '1' it contains whiskers for this threshold.

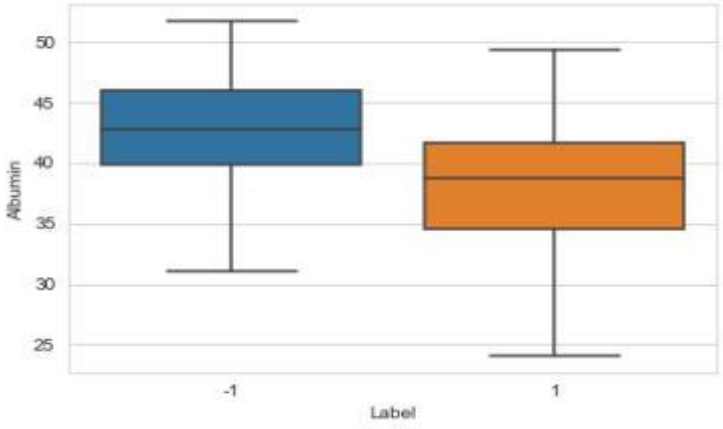


Figure 19 Boxplot of Albumin

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having Albumin ≤ 34.50.
- 50^{th} percentile (Median) of the cancer patients' values are having Albumin ≤ 39.
- 25^{th} percentile of the non-cancer patients' values are having Albumin ≤ 40.
- 50^{th} percentile (Median) of the non-cancer patients' values are Albumin ≤ 42.50.

After reading the plot, we can say that if we choose Albumin ≤ 40 as our threshold for classification then we can see that for class-label '-1' approximately 75% of the values have level greater than this threshold and for class label '1' more than 50% of the values less than this threshold.

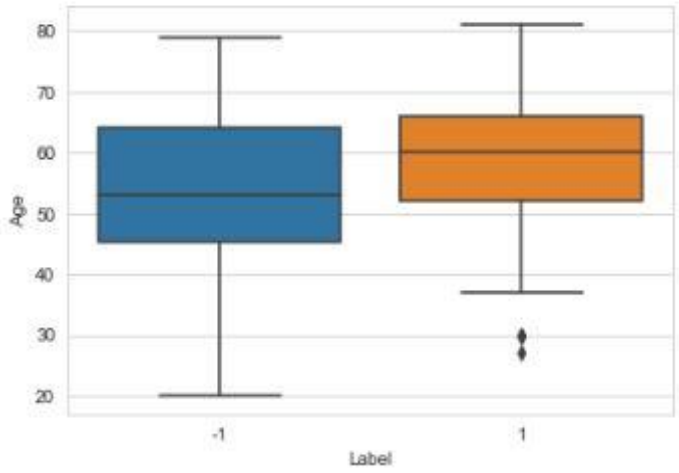


Figure 20 Boxplot of Age

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having Age ≤ 52.
- 50^{th} percentile (Median) of the cancer patients' values are having Age ≤ 60.
- 25^{th} percentile of the non-cancer patients' values are having Age ≤ 45.
- 50^{th} percentile (Median) of the non-cancer patients' values are Age ≤ 53.

After reading the plot, we can say that if we choose Age ≤ 50 as our threshold for classification then we can see that for class-label '-1' more than 50% of the values have level greater than this threshold and for class label '1' more than 25% of the values less than this threshold.

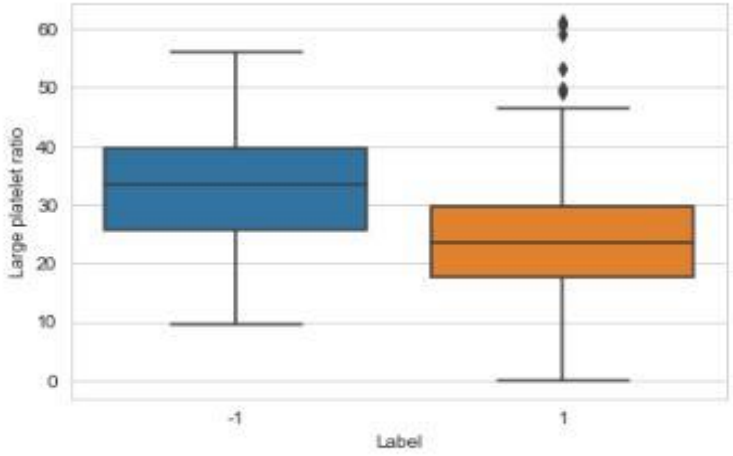


Figure 21 Boxplot of Large platelet ratio

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having large platelet ratio ≤ 18.
- 50^{th} percentile (Median) of the cancer patients' values are having large platelet ratio ≤ 24.
- 25^{th} percentile of the non-cancer patients' values are having large platelet ratio ≤ 26.
- 50^{th} percentile (Median) of the non-cancer patients' values are large platelet ratio ≤ 34.

After reading the plot, we can say that if we choose large platelet ratio ≤ 30 as our threshold for classification then we can see that for class-label '-1' more than 50% of the values have level greater than this threshold and for class label '1' approximately 60% of the values less than this threshold.

Violin Plot

Violin plot includes both PDF and boxplots information. It combines the benefits of the previous two plots (histograms, PDF and boxplots) and simplifies them. We can find the 25^{th}, 50^{th} (Median) and 75^{th} percentiles using Violin plots also evaluate the feasibility of if-else based machine learning models.

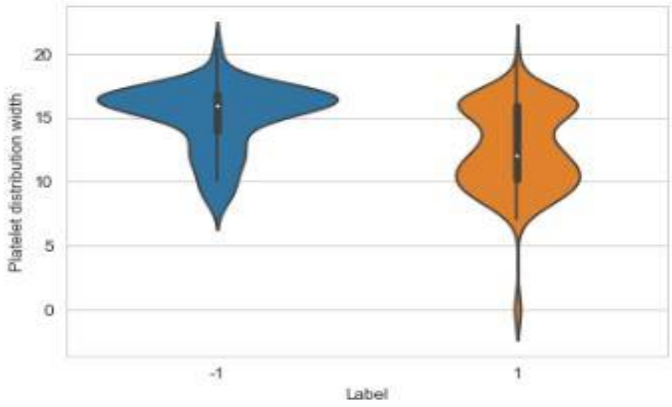


Figure 22 Violin plot of Platelet distribution width

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having Platelet distribution width ≤ 10.
- 50^{th} percentile (Median) of the cancer patients' values are having Platelet distribution width ≤ 12.
- 25^{th} percentile of the non-cancer patients' values are having Platelet distribution width ≤ 14.
- 50^{th} percentile (Median) of the non-cancer patients' values are Platelet distribution width ≤ 16.

After reading the plot, we can say that if we choose large Platelet distribution width ≤ 15.0 as our threshold for classification then we can see that for class-label '-1' more than 50% of the values have level greater than this threshold and for class label '1' more than 65% of the values less than this threshold.

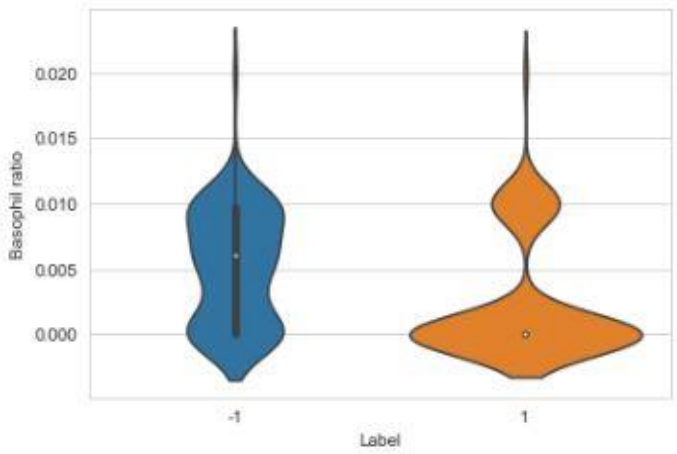


Figure 23 Violin plot of Basophil ratio

OBSERVATIONS

- 25^{th} percentile of the non-cancer patients' values are having Basophil ratio ≤ 0.
- 50^{th} percentile (Median) of the non-cancer patients' values are having Basophil ratio ≤ 0.0060.
- 75^{th} percentile of the non-cancer patients' values are having Basophil ratio ≤ 0.0095.
- 100^{th} percentile of the non-cancer patients' values are having Basophil ratio ≤ 0.0200.

After reading the plot, we can say that if we choose large Basophil ratio ≤ 0 as our threshold for classification then we can see that for class-label '-1' more than 75% of the values have level greater than this threshold and for class label '1' it contains whiskers for this threshold.

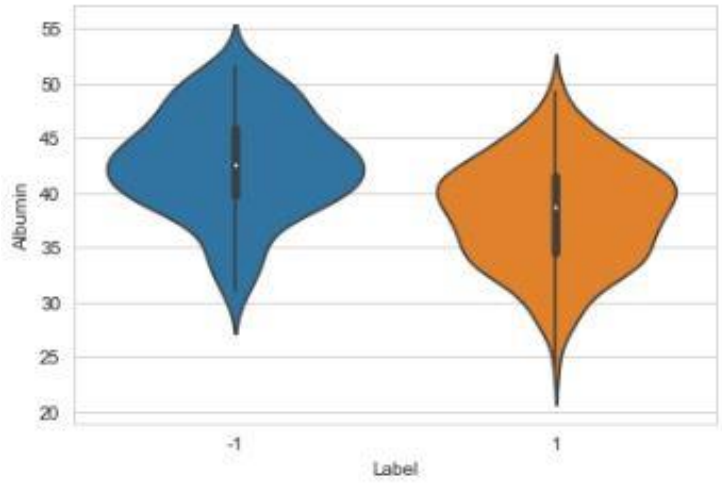


Figure 24 Violin plot of Albumin

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having Albumin ≤ 34.50.
- 50^{th} percentile (Median) of the cancer patients' values are having Albumin ≤ 39.
- 25^{th} percentile of the non-cancer patients' values are having Albumin ≤ 40.
- 50^{th} percentile (Median) of the non-cancer patients' values are Albumin ≤ 42.50.

After reading the plot, we can say that if we choose Albumin ≤ 40 as our threshold for classification then we can see that for class-label '-1' approximately 75% of the values have level greater than this threshold and for class label '1' more than 50% of the values less than this threshold.

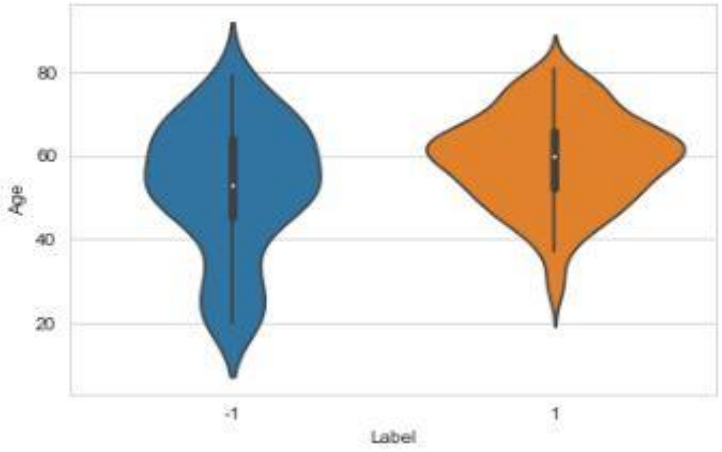


Figure 25 Violin plot of Age

After reading the plot, we can say that if we choose Age ≤ 50 as our threshold for classification then we can see that for class-label '-1' more than 50% of the values have level greater than this threshold and for class label '1' more than 25% of the values less than this threshold.

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having Age ≤ 52.
- 50^{th} percentile (Median) of the cancer patients' values are having Age ≤ 60.
- 25^{th} percentile of the non-cancer patients' values are having Age ≤ 45.
- 50^{th} percentile (Median) of the non-cancer patients' values are Age ≤ 53.

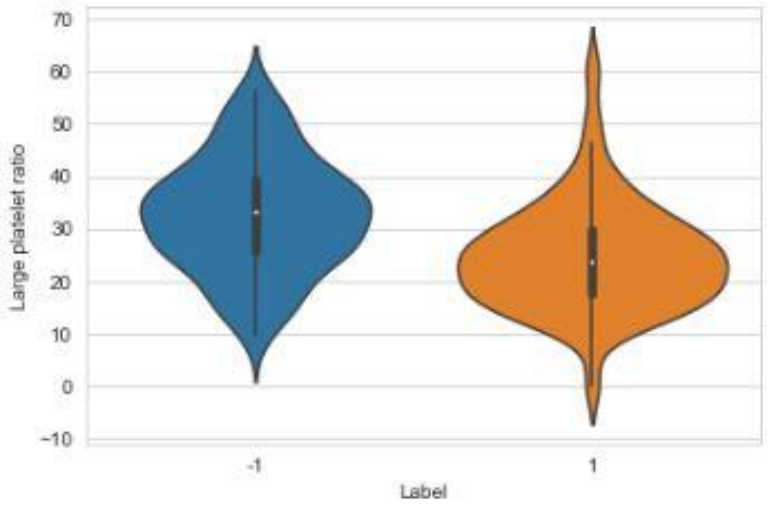


Figure 26 Violin plot of Large platelet ratio

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having large platelet ratio ≤ 18.
- 50^{th} percentile (Median) of the cancer patients' values are having large platelet ratio ≤ 24.
- 25^{th} percentile of the non-cancer patients' values are having large platelet ratio ≤ 26.

- 50^{th} percentile (Median) of the non-cancer patients' values are large platelet ratio ≤ 34.

After reading the plot, we can say that if we choose large platelet ratio ≤ 30 as our threshold for classification then we can see that for class-label '-1' more than 50% of the values have level greater than this threshold and for class label '1' approximately 60% of the values less than this threshold.

3.1.2 BIVARIATE EXPLORATORY DATA ANALYSIS

In Bivariate EDA, we only take into account two variables or features at a time. In the Univariate Analysis, we were unable to distinguish different groups because of overlapping, and we were not able to know how many points are there. Hence we have applied this Bivariate EDA. We graphed the plot 2-D Scatter plot and Pair plot into Bivariate Analysis.

2-D Scatter Plot

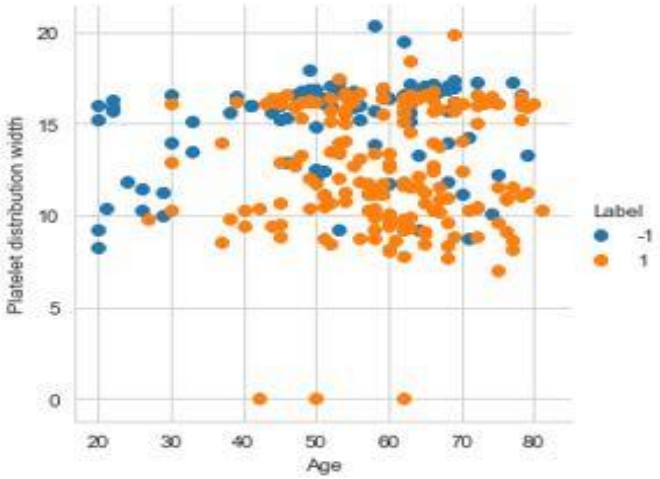


Figure 27 2-D Scatter Plot of Age with Platelet distribution width

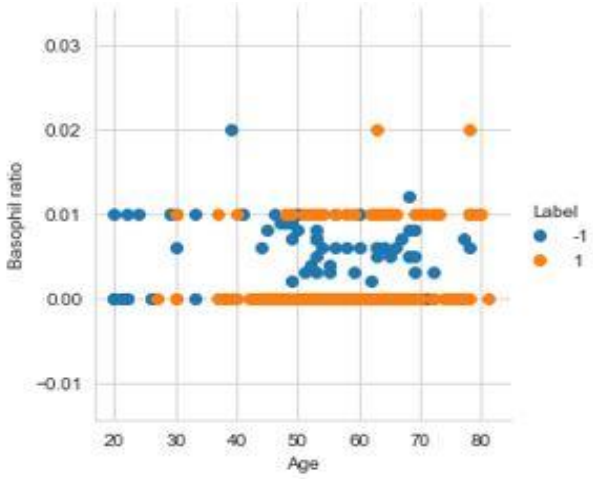


Figure 28 2-D Scatter Plot of Age with Basophil ratio

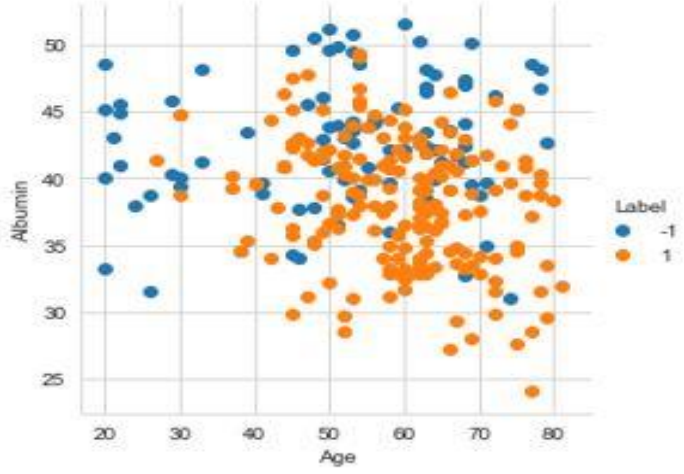


Figure 29 2-D Scatter Plot of Age with Albumin

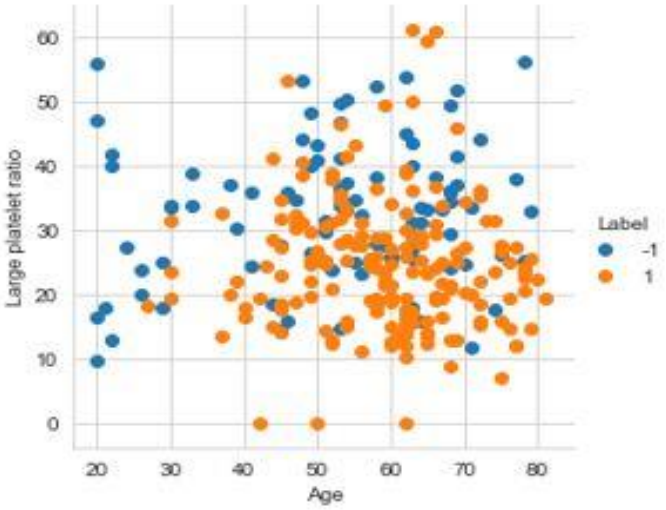


Figure 30 2-D Scatter Plot of Age with Large platelet ratio

OBSERVATIONS

- Separating cancer patients from non-cancer patients is much harder as they have considerable overlap.

As can be seen from the above figures, the cancer and non can cancer patients have overlapping scatter plots. Therefore these features are not capable or sufficient enough to use as standalone features to classify patients

Pair Plot

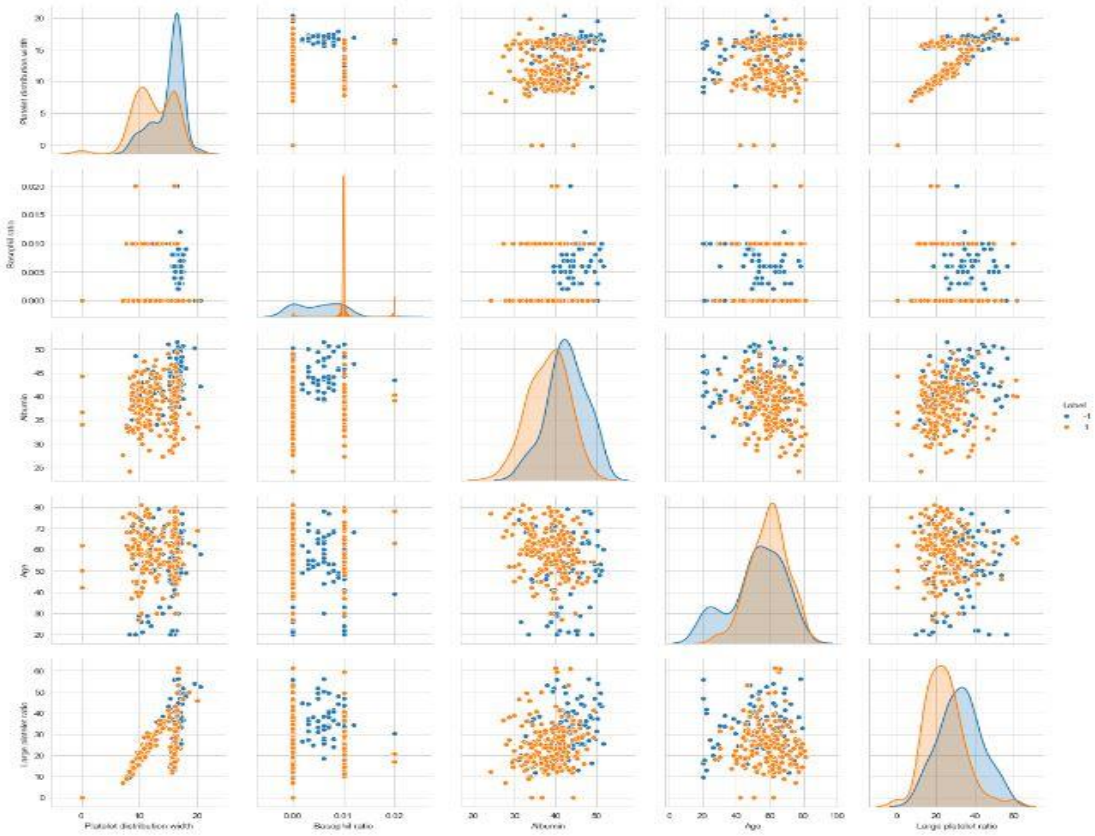


Figure 31 Pair plot of Platelet distribution width, Basophil ratio, Albumin, Age, Large platelet ratio

OBSERVATIONS

- 25^{th} percentile of the cancer patients' values are having large platelet ratio ≤ 18.
- 50^{th} percentile (Median) of the cancer patients' values are having large platelet ratio ≤ 24.

As can be seen from the above figures, the cancer and non can cancer patients have overlapping pair plots. Therefore these features are not capable or sufficient enough to use as standalone features to classify patients.

In this 2-D pair-plot graph, we can get P (5, 2) = 20 scattered graphs because here we have five dimensions in which ten graphs will be unique. We were also unable to make observations and distinguish between different classes in the 2-D plot graph while the diagonal graph is PDF, as can be seen in Figure 31 Pair Plot graph.

CHAPTER 4

SELECTION OF FEATURES

Features are the selected columns of dataset using which we can determine to which class/class-label a particular data point should belong to. Selection of features should be performed in such a way that our training model is given maximum information.

We had used these methods for the selection of features to gain maximum knowledge

1. SelectKBest method: It is one of the algorithms in Scikit-learn library for feature selection. In this algorithm, K features with highest value score from the dataset are selected. For our model, Chi2 has been used as a scoring function for SelectKBest.

2. Tree based method: It is also one of the mainly used algorithms for feature ranking in the Scikit-learn library. It uses multiple techniques to improve model performance.

   ExtraTreeClassifier: It is an ensemble learning technique, also known as extremely randomized trees classifier. Multiple de-correlated decision trees are generated in this classifier and the aggregates of its results are used as classification result.

Table 1 Features ranked from ExtraTreeClassifier method.

RANK	INDEX	REFERENCE RANGE
1	Age	20-81
2	ALB/GLB	0.590-2.660
3	Glucose	2.590-27.060
4	Aspartate aminotransferase	7.000-136.000
5	Uric acid Uric acid	86.000-821.000
6	Alanine transaminase	3.000-177.000
7	AST/ALT	0.230-6.250

Table 2 Blood indices top ranking features (Wu, Jiangpeng et al. 2019)

RANK	INDEX	REFERENCE RANGE
1	Basophil ratio	0.00-0.01
2	Creatine kinase isoenzymes (U/L)	0.0-25.0
3	Platelet large cell ratio (%)	17.0-45.0
4	Albumin (g/L)	30.0-55.0
5	Platelet distribution width (fl)	9.0-17.0
6	Neutrophilic granulocytes (10^9/L)	2.00-7.00
7	White blood cell count (10^9/L)	4.00-10.00
8	Albumin Globulin ratio	1.10-2.50
9	Monocytes (10^9/L)	0.12-1.20
10	Monocyte ratio	0.03-0.08
11	Lymphocyte ratio	0.20-0.40
12	Neutrophile granulocyte ratio	0.50-0.70
13	Lactate dehydrogenase (U/L)	0.0-240.0
14	Carbamide (mmol/L)	1.80-8.00
15	Eosinophil cells (10^9/L)	0.02-0.50
16	Mean corpuscular volume (fl)	80.0-100.0
17	Alkaline phosphatase (U/L)	0.0-120.0
18	Mean corpuscular hemoglobin (pg)	27.0-34.0
19	Creatine kinase (U/L)	0-195

Table 1 and Table 2 shows 26 features; our classification model used these 26 features. Table 1, ranked the features using ExtraTreeClassifier and Table 2, features were selected from (Wu, Jiangpeng et al. 2019); in total 26 features were chosen for our final model. These 26 selected features helped us to achieve better performance in terms of evaluation parameters as compared to (Wu, Jiangpeng et al. 2019).

CHAPTER 5

METHODOLOGY

Improved Lung Cancer detection (ILCD) using Machine Learning techniques, datasets was split into training set and test set i.e. 226 data points was used for training the models (training the dataset) and the remaining 51 data points was used for the evaluation of model performance (testing the dataset). The training data set was standardized for logistic regression and KNN, all the models except these uses non standardized data set. The models were evaluated on the parameters discussed in the section ahead.

ILCD uses cross validation. We have used 10- fold cross validation for model development and hyper-parameter tuning of model was also done.

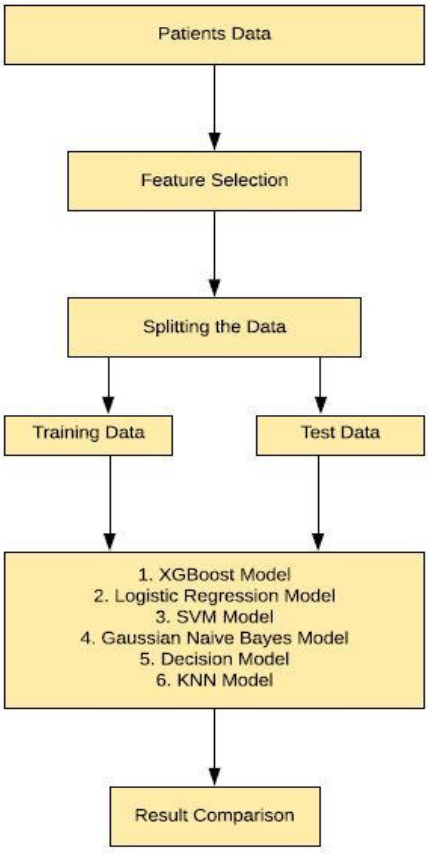


Figure 32 Overall Methodology

5.1 EVALUATION PARAMETERS

The parameters that were used to evaluate Improved Lung Cancer detection (ILCD) using Machine Learning models are as follow:

- Accuracy: Accuracy is the ratio of the correct predictions made by the model to the total observations.
- Confusion Matrix: Confusion Matrix is an NxN matrix, here N is the number of classes in the dataset. For example ILCD has two classes Lung Cancer patients (or '1') and without Lung Cancer patients (or '-1') so the confusion matrix will be of the size 2x2.
- AUC-ROC: AUC is the area under the ROC curve. ROC curve is the plot between TPR and FPR.
- FNR: FNR is the ratio of the person who is having the disease, but the test results say Negative. It is also known as Miss Rate (MR).

Classification methods such as XGBoost, Decision Tree, KNN, Logistic Regression, Gaussian Naïve Bayes and SVM had been used on the lung cancer dataset. All the models' results were analyzed, and it was found that XGBoost provided the best results.

The final classification model is built using XGBoost classification method. Speed and accuracy are the two principal advantage of using the XGBoost model. Classification, regression, user-defined prediction and ranking are the problems that can be solved using XGBoost method. We first used XGBoost hyperparameter tuning using GridSearchCV in our research work. GridSearchCV is a standard way to do hyper parameter tuning on any classification method. Grid of parameters are defined for our classification method and then grid search was run on this grid of parameters. 10-fold Cross-validation was also used with GridSearchCV.

The other parameters that were also found to test performance of our models are Sensitivity (TPR), Matthews correlation coefficient (MCC), Specificity (TNR), False Positive, True Positive, True Negative and False Negative.

- Sensitivity = TP / (TP+FN)
- Specificity = TN / (TN+FP)
- False Negative Rate = FN / (TP+FN)
- Accuracy = (TP+TN) / (TP+FP+FN+TN)
- $MCC = (TP*TN - FP*FN) / \sqrt{((TP + FN) * (TN + FP) * (TN + FN))}$

The methodological steps are summarized below:

1. Training set and test set were created from the dataset.
2. Select the features as discussed in Chapter 4 (selection of features).
3. Class weight is applied wherever applicable as the dataset is unbalanced and hyperparameter tuning using GridSearchCV is done.
4. Use best parameters to create model.
5. Validation of the model is done using 10-fold Cross Validation.
6. Find the evaluation parameters like accuracy, confusion matrix. We will be getting TP, TN, FP and FN from the confusion matrix.

The XGBoost method is better, according to the accuracy matrix. The FNR weightage is quite high, FNR for the model is low, which is good for our application domain. The FNR is the ratio of the person who has the disease, but the results of the test say negative. It is also called Miss Rate. The closest this value is to 0, the better it will be. The AUC-ROC curve value can vary from [0, 1]. The prediction performance of lung cancer will be better, closer the AUC of ROC is to 1. For the medical domain, FNR and Sensitivity or TPR are the better parameters. If we try to decrease or reduce FNR, then it means we are increasing Sensitivity or TPR. Sensitivity or TPR should be maximized because if it is low, then it means there is a high chance of classifying a patient with lung cancer as not having cancer and for the patient, this could be fatal. That is why it is essential to minimize FNR or maximize Sensitivity or TPR.

CHAPTER 6

RESULT

Table 3 compares the performance of various classification methods on lung cancer dataset. The XGBoost model's performance achieved better results compared with the RBLC Model [11]. 10-fold Cross-validation with GridSearchCV was performed on each of the machine learning models that were used on the lung cancer dataset. Hyper parameter tuning of models was also done using Scikit-learn GridSearchCV where ever applicable. Accuracy, AUC-ROC Curve, FNR and confusion matrix are the evaluation parameters. The comparison with the base paper and with other models are shown in Table 3 for FNR, Table 4 for the performance of models and Table 5 for AUC-ROC Curve. If the false-negative rate is high, that means a diagnosis can endanger a patient's life because the true value is misclassified as false.

The FNR score was only less than 4%, AUC of ROC score was 95% greater than 90% while the accuracy was 92.16% that is also higher than 90%.

False Negative Rate = 3.33%, Accuracy = 92.16%, Sensitivity = 96.67%, Specificity = 85.71%, MCC = 83.86%, AUC = 95%.

Table 3 False Negative Rate (FNR) of Models

MACHINE LEARNING TECHNIQUES	FALSE NEGATIVE RATE (FNR)
RBLC MODEL [11]	14.29%
ILCD MODEL	3.33%

Table 4 Performance of Models

MACHINE LEARNING TECHNIQUES	TRAIN ACCURACY	TEST ACCURACY
RBLC MODEL [11]	-	88.24%
ILCD MODEL	99.01%	92.16%
LOGISTIC REGRESSION MODEL	87.71%	86.27%
SVM MODEL	88.67%	84.31%
GAUSSIAN NAÏVE BAYES MODEL	76.60%	72.55%
DECISION TREE MODEL	82.55%	76.47%
KNN MODEL	100%	74.51%

Table 5 AUC-ROC Score

MACHINE LEARNING TECHNIQUES	AUC-ROC SCORE
RBLC MODEL [11]	0.9016
ILCD MODEL	0.95
LOGISTIC REGRESSION MODEL	0.91
SVM MODEL	0.91
GAUSSIAN NAÏVE BAYES MODEL	0.81
DECISION TREE MODEL	0.80
KNN MODEL	0.77

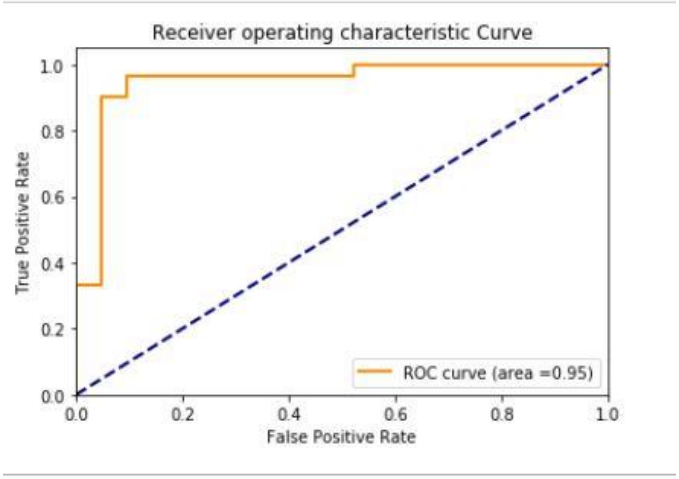


Figure 33 ROC curve

Actual	0	1	All
Predicted			
0	18	1	19
1	3	29	32
All	21	30	51

Figure 34 Confusion Matrix

Comparison of RBLC model [11] performance with all models that were applied is shown in Table 4. The FNR for XGBoost Model is one order of magnitude smaller than the RBLC model [11] as shown in Table 3. The AUC-ROC Score of RBLC model [11] with other models is compared in Table 5. For test data, the AUC of the ROC curve is 95% as shown in Figure 33. For XGBoost model, it is much better than the RBLC model [11] who's AUC of the ROC curve was 90%. In Figure 34 Actual refers to Actual Values while Predicted refers to Predicted Values. In Figure 34 Matrix (0, 0) refers to TN, matrix (0, 1) refers to FN, matrix (1, 0) refers to FP and matrix (1, 1) refers to TP.

CHAPTER 7

CONCLUSION

XGBoost provides the best accuracy, sensitivity, AUC-ROC Curve and FNR for lung cancer, i.e. 92.16%, 96.67%, 0.95 and 3.33%. In healthcare, the less false-negative rate is desirable, as it means that true value is misclassified as false less often. Some models perform well for a certain parameter, and some not well for other parameters. On the basis of accuracy we obtained, it gave better results than the basic paper implementation. The parameter tuning helped us to increase the accuracy of our models. Parameter tuning increased the accuracy of our models. All the results above show that Machine Learning techniques can be useful in the identification of lung cancer. These methods can help doctors diagnose lung cancer, which can be verified by screening tests that can help us save a human being's precious life.

REFERENCES

[1] S. Roan, ‘Blood Test Can Accurately Detect Early-Stage Lung Cancer’, Everyday Health, 2018. [Online]. Available: https://www.everydayhealth.com/

lung-cancer/blood-test-can-accurately-detect-early-stages/. [Accessed: 26- May – 2020]

[2] ‘Lung cancer’, Mayo Clinic, 2020. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/lung-cancer/diagnosis-treatment/ drc-20374627. [Accessed: 26-May-2020]

[3] ‘Cancer statistics’, National Cancer Institute, 2018. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/statistics. [Accesses: 26-May-2020]

[4] ‘International Agency for Research on Cancer’, World Health Organization, Available: https://gco.iarc.fr/. [Accessed: 26-May-2020]

[5] Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 68, no. 6 (2018): 394-424.

[6] Clark, Robert D., and Daniel J. Webster-Clark. "Managing bias in ROC curves." Journal of computer-aided molecular design 22, no. 3-4 (2008): 141-146.

[7] Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." cell 144, no. 5 (2011): 646-674.

[8] Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. "Machine learning applications in cancer

prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

[9] Leng, Shaoyi, Jianjun Zheng, Yinhua Jin, Hongbin Zhang, Ya Zhu, Jing Wu, Yan Xu, and Puhong Zhang. "Plasma cell-free DNA level and its integrity as biomarkers to distinguish non-small cell lung cancer from tuberculosis." Clinica Chimica Acta 477 (2018): 160-165.

[10] Fradkin, Dmitriy, Dona Schneider, and Ilya Muchnik. "Machine learning methods in the analysis of lung cancer survival data." DIMACS Technical Report 2005–35 (2006).

[11] Wu, Jiangpeng, Xiangyi Zan, Liping Gao, Jianhong Zhao, Jing Fan, Hengxue Shi, Yixin Wan, E. Yu, Shuyan Li, and Xiaodong Xie. "A Machine Learning Method for Identifying Lung Cancer Based on Routine Blood Indices: Qualitative Feasibility Study." JMIR medical informatics 7, no. 3 (2019): e13476.

[12] Lynch, Chip M., Behnaz Abdollahi, Joshua D. Fuqua, R. Alexandra, James A. Bartholomai, Rayeanne N. Balgemann, Victor H. van Berkel, and Hermann B. Frieboes. "Prediction of lung cancer patient survival via supervised machine learning classification techniques." International journal of medical informatics108 (2017): 1-8.

[13] Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability." arXiv preprint arXiv:1206.1121 (2012).

[14] Lynch, Chip M., Victor H. van Berkel, and Hermann B. Frieboes. "Application of unsupervised analysis techniques to lung cancer patient data." PloS one 12, no. 9 (2017).

[15] 'Lung Cancer 101', lungcancer.org, Available: https://www.lungcancer.org/find_information/publications/163-lung_cancer_101/265-what_is_lung_cancer. [Accessed: 28-May-2020]

[16] Agrawal, Ankit, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, and Alok Choudhary. "A lung cancer outcome calculator using ensemble data mining on SEER data." In Proceedings of the tenth international workshop on data mining in bioinformatics, pp. 1-9. 2011.

[17] Szpechcinski, A., J. Chorostowska-Wynimko, R. Struniawski, W. Kupis, P. Rudzinski, R. Langfort, E. Puscinska, P. Bielen, P. Sliwinski, and T. Orlowski. "Cell-free DNA levels in plasma of patients with non-small-cell lung cancer and inflammatory lung disease." British journal of cancer 113, no. 3 (2015): 476-483.

[18] Dheda, K., and C. E. Barry 3rd. "3rd, Maartens G." Lancet 387 (2016): 1211-26.

[19] Zumla, Alimuddin, Andrew George, Virendra Sharma, Rt Hon Nick Herbert, Aaron Oxley, and Matt Oliver. "The WHO 2014 global tuberculosis report—further to go." The Lancet Global Health 3, no. 1 (2015): e10-e12.

[20] Travis, William D., Elisabeth Brambilla, Andrew G. Nicholson, Yasushi Yatabe, John HM Austin, Mary Beth Beasley, Lucian R. Chirieac et al. "The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification." Journal of thoracic oncology 10, no. 9 (2015): 1243-1260.

[21] Singh, Vikas Kumar, Subhash Chandra, Sachin Kumar, Ghanshyam Pangtey, Anant Mohan, and Randeep Guleria. "A common medical error: lung cancer misdiagnosed as sputum negative tuberculosis." Asian Pac J Cancer Prev 10, no. 3 (2009): 335-8.

[22] Esposito, Angela, Carmen Criscitiello, Dario Trapani, and Giuseppe Curigliano. "The emerging role of “Liquid Biopsies,” circulating tumor cells, and circulating cell-free tumor dna in lung cancer diagnosis and identification of resistance mutations." Current oncology reports 19, no. 1 (2017): 1.

[23] Paci, Massimiliano, Sally Maramotti, Enrica Bellesia, Debora Formisano, Laura Albertazzi, Tommaso Ricchetti, Guglielmo Ferrari et al. "Circulating plasma DNA as diagnostic biomarker in non-small cell lung cancer." Lung cancer 64, no. 1 (2009): 92-97.

[24] Dimitoglou, George, James A. Adams, and Carol M. Jim. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability." arXiv preprint arXiv:1206.1121(2012).

[25] Yoon, Kyong-Ah, Sohee Park, Sang Hee Lee, Jin Hee Kim, and Jin Soo Lee. "Comparison of circulating plasma DNA levels between lung cancer patients and healthy controls." The Journal of Molecular Diagnostics 11, no. 3 (2009): 182-185.

[26] Lum, Caroline, and Muhammad Alamgeer. "Technological and Therapeutic Advances in Advanced Small Cell Lung Cancer." Cancers 11, no. 10 (2019): 1570.

[27] van der Drift, Miep A., Bernard EA Hol, Corné HW Klaassen, Clemens FM Prinsen, Yvonne AWG van Aarssen, Rogier Donders, Jos WJ van der Stappen, PN Richard Dekhuijzen, Henricus FM van der Heijden, and Frederik BJM Thunnissen. "Circulating DNA is a non-invasive prognostic factor for survival in non-small cell lung cancer." Lung Cancer 68, no. 2 (2010): 283-287.

[28] Jiang, Tao, Changyun Zhai, Chunxia Su, Shengxiang Ren, and Caicun Zhou. "The diagnostic value of circulating cell free DNA quantification in non-small cell lung cancer: A systematic review with meta-analysis." Lung Cancer 100 (2016): 63-70.

[29] Gautschi, Oliver, Colette Bigosch, Barbara Huegli, Monika Jermann, Arthur Marx, Eveline Chassé, Daniel Ratschiller et al. "Circulating deoxyribonucleic acid as prognostic marker in non–small-cell lung cancer patients undergoing chemotherapy." Journal of Clinical Oncology 22, no. 20 (2004): 4157-4164.

[30] Kumar, Sachin, Randeep Guleria, Vikas Singh, Alok C. Bharti, Anant Mohan,

and Bhudev C. Das. "Efficacy of circulating plasma DNA as a diagnostic tool for advanced non-small cell lung cancer and its predictive utility for survival and response to chemotherapy." Lung Cancer 70, no. 2 (2010): 211-217.

[31] Chiappetta, Caterina, Marco Anile, Martina Leopizzi, Federico Venuta, and Carlo Della Rocca. "Use of a new generation of capillary electrophoresis to quantify circulating free DNA in non-small cell lung cancer." Clinica Chimica Acta 425 (2013): 93-96.

[32] Wang, Brant G., Han-Yao Huang, Yu-Chi Chen, Robert E. Bristow, Keyanunoosh Kassauei, Chih-Chien Cheng, Richard Roden, Lori J. Sokoll, Daniel W. Chan, and Ie-Ming Shih. "Increased plasma DNA integrity in cancer patients." Cancer research 63, no. 14 (2003): 3966-3968.

[33] Shen, Shu Yi, Rajat Singhania, Gordon Fehringer, Ankur Chakravarthy, Michael HA Roehrl, Dianne Chadwick, Philip C. Zuzarte et al. "Sensitive tumour detection and classification using plasma cell-free DNA methylomes." Nature 563, no. 7732 (2018): 579-583.

[34] M. Araujo, 'HOW MACHINE LEARNING WORKS', feedzai, 2020. [Online]. Available: https://feedzai.com/blog/how-machine-learning-works/. [Accessed: 29-May-2020]

[35] 'Machine Learning in MATLAB', MathWorks, Available: https://www.mathworks.com/help/stats/machine-learning-in-matlab.html. [Accessed: 29-May-2020]

LIST OF PUBLICATIONS OF THE CANDIDATE'S WORK

[1] Puneet, and Anamika Chauhan. "A Study of Black hole attacks in Delay Tolerant Network" In *International Conference on Emerging Wireless Communication Technologies and Information Security*, EWICS 2020. Springer, Cham, 2020.

[2] Puneet, and Anamika Chauhan. "Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices" In *2020 IEEE International Conference for Innovation in Technolog*y, INOCON2020.