

**CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD) PREDICTION  
USING MACHINE LEARNING**

**BY**

**MRIDUL DAS JOSHE  
ID: 172-15-9714**

**NAZMUL HASSAN EMON  
ID: 172-15-9703**

This Report Presented in Partial Fulfillment of the Requirements for the  
Degree of Bachelor of Science in Computer Science and Engineering:

Supervised By

**DR. Sheak Rashed Haider**  
Associate Professor & Associate Head  
Department of CSE  
Daffodil International University

Co-Supervised By

**Mr. Seikh Abujar**  
Senior Lecturer  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**MAY 2021**

## APROVAL

This project title “**CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD) PREDICTION USING MACHINE LEARNING**”, submitted by **Mridul Das Joshe** and **Nazmul Hassan Emon** ID No:172-15-9714,172-15-9703 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 31 May,2021.

### BOARD OF EXAMINERS



---

**(Name)**

**Chairman**

**Designation**

Department of CSE

Faculty of Science & Information Technology

Daffodil International University



---

**Gazi Zahirul Islam**

**Internal Examiner**

**Assistant Professor**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



---

**Raja Tariqul Hasan Tusher**

**Internal Examiner**

**Senior Lecturer**

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International university

**External Examiner**

A handwritten signature in black ink, appearing to read 'Farid', with a horizontal line drawn underneath it.

**Dr. Dewan Md. Farid**

**Associate Professor**

Department of Computer Science and Engineering

United International University

## DECLARATION

We hereby declare that this Project has been done by us under the supervision of **DR. Sheak Rashed Haider, Associate Professor & Associate Head of CSE** Daffodil International University. We also declare that neither this Project nor part of this Project has been submitted elsewhere for the award of any degree or diploma.

### Supervised by:



---

**Dr. Sheak Rashied Haider Noori**  
Associate Professor and associate Head  
Department of CSE  
Daffodil International University

### Co-Supervised by:



---

**Mr. Sheikh Abujar**  
Sr. Lecturer  
Department of CSE  
Daffodil International University

### Submitted by:



---

**Mridul Das Joshe**  
ID: -172-15-9714  
Department of CSE  
Daffodil International University



---

**Nazmul Hassan Emon**  
ID: -172-15-9703  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First, we express our heartiest and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project.

We really grateful and wish our profound our indebtedness to **Dr. Sheak Rashed Haider Noori, Associate Professor and Associate Head**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry to complete this task. His perpetual persistence, insightful direction, consistent consolation, steady and enthusiastic oversight, productive analysis, important guidance, perusing numerous sub-par draft and adjusting them at all stages have made it conceivable to finish this undertaking. We like to thank our co-supervisor **Mr. Seikh Abujar**. We might want to offer our heartiest thanks to Almighty Allah and Head, Department of CSE , for his thoughtful assistance to complete our task and furthermore to other employees and the stuff of CSE division of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while the course work.

At long last, we should recognize with due regard the consistent help and patients of us folks.

## **ABSTRACT**

In this Advanced era Chronic Inflammatory lung disease is increasing day by day. Chronic Obstructive Pulmonary Disease (COPD) is one of them which is highly conventional disease characterized by obstructed air flow from the lung. According to research, COPD occurs in people over the age of 40. The amount of affecting rates is getting increased for smoking rates, pollution, industrial pollution in the developed World. Though we believe in advance Era but we don't have sufficient awareness about COPD. People also don't know the level of COPD are currently affected. As like other disease COPD also has some state of condition like (mild, moderate, severe, very severe).

Here we implement machine learning algorithm which build a model. Model predict a level of Chronic Obstructive pulmonary Disease Level. We Learn our model with the symptom's basis data. Our system predicts the disease if have then it will give the level of the disease analyzing the symptoms provide by the user as input.

Prediction of chronic obstructive pulmonary disease is done by implementing Machine Learning Algorithms such as Logistic Regression, Decision Tree Classifier, K-Nearest Neighbor, Naïve Bayes. Therefore, we can say our model and data set have more information than past utilize. Finally, our model has standard functional testing.

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE</b>
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
List of Figure	viii
List of tables	ix

## **CHAPTER**

### **CHAPTER 1: Introduction 1-4**

1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the study	3
1.4 Research Question	3
1.5 Expected Outcomes	3
1.6 Project Management and Finance	4
1.7 Report Layout	4

### **CHAPTER 2: Background**

2.1 Preliminaries	5
2.2 Related Works	5
2.3 Comparative Analysis and Summary	6
2.4 Scope of the problem	6

<b>CHAPTER 3: Research Methodology</b>	
3.1 Introduction	7
3.2 Research Project Subject and Instruction	7
3.3 Data Collection Procedure	8-10
3.4 Applied Mechanism	12-16
3.5 Statistical Analysis	17
3.6 Implementation Requirements	17
<b>CHAPTER 4: Experimental Result and Discussion</b>	
4.1 Experimental setup	18
4.2 Experimental Result & Analysis	18-20
4.3 Discussion	20-21
<b>CHAPTER 5: Impact on Society, Environment</b>	
5.1 Impact on Society	22
5.2 Impact on Environment	22
5.3 Sustainability plan	22
<b>CHAPTER 6: Summery, Conclusion, and Implication for Future Research</b>	
6.1 Summary of the Study	23
6.2 Conclusions	23
6.3 Implication for Future Study	23
<b>FEFERENCES</b>	24
<b>APPENDICES</b>	25-27



## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figures 3.3.1: Sample Dataset	8
Figures 3.3.2: Packages	8
3.3.3: Attributes of Dataset	9
3.3.4: Positive and negative data visualization	9
Figure 3.3.5: Gender Data visualization	10
Figure 3.3.6: Age Data visualization	10
Figures 3.3.7: FEV/FVE	11
Figures 3.3.8: Data Visualization	11
Figures 3.4.1: Logistic Regression	12
Figure 3.4.2: Confusion heatmap of Logistic Regression	13
Figure 3.4.3: Prediction score of Decision Tree	14
Figure 3.4.4: Confusion heatmap of Decision Tree	14
Figure 3.4.5: Confusion heatmap of K-Nearest Neighbors	15
Figure 3.4.6: Confusion heatmap of Naïve Bayes	16
Figure3.4.7: Accuracy of all applied Algorithm	21

## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 3.2.1: Development Tools	7
Table4.2.1: Performance Table for Logistic	19
Table4.2.2: Performance Table for Decision Tree	19
Table4.2.3: Performance Table for Naïve Bayes	20
Table4.2.4: Performance Table for KNN	20

# CHAPTER 1

## Introduction

### 1.1 Introduction:

The term Machine learning is now a very well-known name. It is common for students and researchers in computer science, as well as students and researchers in other disciplines of science and engineering, to work with machine running on problem-based issues. What started in the sixties as a sub-branch of artificial intelligence came in the nineties to surpass that main branch. In fact, 80 percent of the talk today about artificial intelligence is about machine learning. Technical data and machine learning practices of Western countries in both global. Formal and informal education, speaking developed countries, now even school and college-level students learn in different ways and with the use of machine learning is going to be involved in some way. Competitions to increase the variety of machine learning hackathons held at the school-college level have gradually become very popular in those countries and they have achieved enviable success in the practical application of information at the university level as a long-term result. This was a time of turmoil coming out of the solution by using machine learning.

Machine learning is a symptom closely related to computer data files. If the predictions turn to the computer using the machine learning is learning the Statistical. Mathematical optimization in the fields of machine learning methods section provides information and application domain.

## **1.2 Motivation:**

Machine learning is an active improvement in computer algorithm acceptance and data usage. Machine learning algorithms are used, such as the different types of e-mail filtering application and computer vision algorithms used to perform the required tasks are difficult to develop or required.

Many of the world's machine learning they may also contribute to treatment and medicine. Using machine learning technology enterprises create projects that include concord protocol and increase the interest of the users. Nowadays medicine and Treatment are using Machine Learning technology to improve the performance.

Machine Learning technology has several Algorithm. Those are: Logistic Regression, K-Nearest Neighbors, Random Forest, Decision Tree Classifier, Naïve Bayes Classifier. The tools are: Knime, Accord.net, Scikit-Learn, TensorFlow, Weka, Pytorch, RapidMiner, Google Cloud, AutoML, Jupyter, Notebook, MLLIB. As our model will use several Algorithm. COPD actually happen after age of 40. Early diagnosis of COPD leads to better outcomes. There are effective medications that can help patients do more and can help lower chances of having COPD flares. Some patients may not know they have COPD. Early diagnosis is key to reduce COPD progression and symptoms.

Thus, if we can detect COPD many people will be take diagnosis as soon as possible and we will help to save many lives.

In our Project we implement machine learning so that we can predict a percentage of COPD level.

## **1.2 Rationale of the Study:**

Chronic Obstructive Pulmonary Disease (COPD) is a principal cause of fatality and calamity in less developed countries including Bangladesh. But there is no methodical review has been implemented in Bangladesh, which characterization the compiles of COPD and its risk factors.

Therefore, our research will be conducted to find a predictable accuracy of a patient. We study about Chronic Obstructive Pulmonary Disease through Bangladesh. COPD appearance is very serious in Bangladesh. It will be a great timely need for the research trough Bangladesh for making awareness and taking relevant steps for prevention and control of COPD in Bangladesh.

## **1.4. Research Questions:**

- a. What kind of disease is COPD?
- b. What are the symptoms of this disease in Bangladesh?
- c. Is this disease is relatable to Asthma?
- d. Is this Disease preventable?
- e. What are the future of is this research?
- f. How is this Project performing?

## **1.5 Expected Outcomes:**

Few numbers of studies and researches have been focusing on two common kinds of diseases like diabetes and tuberculosis. But we don't get sufficient data are available on COPD. Scince many years ago, lots of studies have been completed in our country to make awareness and identify the burden and risk factors of COPD. However, these kinds of studies remain different informatively in terms of methodology like hospital-based and population-based study, sample size, and area of the study in urban and rural area.[11].

Moreover, we tried to apply a methodology patients-based study. These kinds of scale were used to measure the different status in this study.

## **1.6 Project Management and Finance:**

The component of Machine Learning is open source so we did not need to buy any software based things. We read different article and papers from medium that costed us few dollars. For this Project we had the cost of our hardware and internet bill.

## **1.7 Report Layout:**

There are 6 chapters in this report. In Chapter 1 contains overview of our total work. There are some sections in chapter 1. They are section 1.1 Introduction, 1.2 Motivation of this research, 1.3 Rationale of the study, 1.4. Research Questions, 1.5 Expected Outcomes, 1.6 Project management and Finance, 1.7 Report layout. In Chapter 2 we have discussed the Background Studies of the research and its subsections are 2.1 Preliminaries, 2.2 Related Works, 2.3 Comparative Analysis and Summary, 2.4 Scope of the problem, 2.5 Challenges. We have discussed the Research Methodology with subsections in Chapter 3 3.1 Introduction, 3.2 Research Project Subject and Instruction, 3.3 Data Collection Procedure 3.4 Statistical Analysis, 3.5 Applied Mechanism, 3.6 Implementation Requirements. In Chapter 4, we have discussed about Experimental Results and Discussion of those applied algorithm and discussed about the best algorithm for our dataset. It's has some sections; 4.1 Experimental Setup, 4.2 Experimental Results & Analysis, 4.3 Summary. In chapter, 5 we have discussed about impact on society and Environment. It contains 5.1 Impact on Society, 5.2 Impact on Environment, 5.3 Sustainability Plan In chapter, 6 The Summary, Conclusion, and Implication for Future Research works the subsections 6.1 Summary of the Study, 6.2 Conclusions 6.3, Implication for Future Study.

## **CHAPTER 2**

### **Background**

#### **2.1 Preliminaries:**

Due to the new computer technology, in this developed era machine learning changed into advanced term. Machine learning was invented from pattern recognition and the theory that computers can learn automatically being programmed to perform specific tasks. The new information of the model is exposed as an important factor in the machine learning aspect, they are able to adapt independently.

Increasing interest in machine learning for the common reasons that have made data mining and Bayesian analysis more popular than ever past. The volume and the available data, the computational processing that is user friendly, cheaper and more powerful and the variety of affordable data collection are increasingly good things.

The standard of machine learning technology is recognized in the industry working with the largest amount of data. Every industry like health care, transportation, oil and gas, retail etc. are using Machine Algorithm.

#### **2.2 Related Works:**

Sipsita Sutradhar common and Risk Factors of Chronic Obstructive Pulmonary Disease in Bangladesh. They carried out systematic analysis with the aim to search out the prevalence of COPD. They also try to find a policy-maker to develop suitable policies [11]. Blanca E. Himes1 Prediction of Chronic Obstructive Pulmonary Disease (COPD) in Asthma Patients using Electronic Medical Records, A Bayesian network model. They used symptoms of COPD patients. Symptoms are age, sex, race, smoking history [2].

### **2.3 Comparative Analysis and Summary:**

We have introduced a methodology for predicting Chronic Obstructive Pulmonary Disease using Machine Learning Algorithm in our research. In a World Health Organization report predicts that by 2030, COPD will be over, it is the third leading cause of death and the seventh leading cause of illness worldwide. In fact, leading to COPD is one of the leading possible avoidable hospitalizations. The novel solution is that the effects of the disease on the patient's health are adequately apparent to the limited need.

### **2.4 Scope of the problem:**

We had to face some questions while working with COPD. A portion of the inquiries can emerge with respect to our applications. The inquiries and their most potential answers were given underneath.

- . What is the purpose of working with COPD despite having many different forms?
- . How to get your data because there are less such patients?

The answers of these most popular questions are given below:

Medical Treatment is one of the fundamental needs of top 5 basic human rights. It's us right to know about our disease what we are consuming. If we cannot get the treatment, it may take our lives. This technology will predict a percentage. Thus, the patient will give us their information to confirm their percentage level of COPD.

Our biggest challenge was to go from the hospital to the patient and collect the patient data. The current situation is that the harassment of Covid 19 is very high. In this situation, collecting data becomes a big challenge for us.



## CHAPTER 3

### Research Methodology

#### 3.1 Introduction:

There many types of obstructive lung diseases such as Chronic obstructive pulmonary disease (COPD) characterized by long-term breathing problems and poor airflow. Shortness of breath and cough with sputum production is the main symptoms. There were two older term such as Chronic bronchitis and emphysema used for different types of COPD.

#### 3.2 Research Project Subject and Instruction:

The name of our Project is “Chronic obstructive Pulmonary Disease prediction using Machine Learning”. This is a fresh and contemporary research topic in the technology industry in current days. The development procedure of this project work with the detailed conceptual and theoretical process is also discussed in this chapter.

A Pc with high configuration is needed for this Project. Required instruments list is given below:

Hardware and Software:

Development Tools:

Component	Description
Programing Language	Python
Python	2019,3.1
Platform	Google COLAB
Operating System	Windows 10 64 bit
Memory	8 GB

Table 3.2.1: Development Tools

### 3.3 Data Collection Procedure:

Name	Age (yrs)	Sex(m=0,f=1)	Cough(yes=1)	Breathe Short	Inhaler(yes=1)	Smoker(y)	Wheeze(y)	Chest pain	Respiratio	FEV1/FVC	FVC ( > = 80%	Occupatio	Restless	COPD(pos=1,neg=0)
Amjad	50	0	1	1	1	1	1	0	0	2	2	0	1	1
Rafiq	48	0	1	1	0	0	0	1	1	1	1	1	1	0
Sattar Mia	55	0	1	1	0	1	1	1	1	0	4	2	0	1
Salam	56	0	1	1	1	0	0	0	1	5	2	0	1	1
Alom	50	0	1	1	1	1	0	1	0	4	2	0	1	1
Biptop	30	0	1	1	0	1	1	0	1	3	1	0	0	0
Ronjit	48	0	1	1	1	1	1	0	0	5	2	0	1	1
Bablu	41	0	1	1	0	0	0	0	0	4	2	1	0	1
Parvez	46	0	1	1	1	1	1	1	1	0	2	0	1	1
Robiul	35	0	1	1	0	1	1	0	1	3	1	0	0	0
Akbor Ali	40	0	1	1	1	0	0	0	1	3	1	0	0	0
Akalu	64	0	1	1	1	0	1	1	0	5	2	0	1	1
Saleha	46	1	1	1	0	0	0	1	1	3	1	1	1	0
Edres	65	0	1	1	1	1	1	0	0	2	2	0	1	1
Rupali	46	1	1	1	0	0	1	1	3	1	0	1	1	0
Riyad	27	0	1	1	0	0	0	1	1	3	1	0	1	0
Bruse	36	0	1	1	1	1	1	1	0	5	2	0	1	1
Sanju	28	0	1	1	0	1	0	0	1	3	1	0	0	0
Subot	64	0	1	1	1	0	1	1	0	2	2	1	1	1
Kanchan	48	1	1	1	0	0	0	1	1	1	1	0	1	0
Rahman	60	0	1	1	1	1	1	1	0	5	2	0	1	0
Babul	52	0	1	1	0	1	0	0	1	3	1	0	0	0
Belal	54	0	1	1	1	1	1	1	0	5	2	0	0	1
Ahmod	50	0	1	1	0	0	1	1	0	2	2	1	1	1
Jibon	46	0	1	1	0	1	0	0	1	3	1	0	0	0
Tarabodo	60	0	1	1	0	1	1	1	0	5	2	0	1	1
Kobul	57	0	1	1	0	1	1	0	0	5	2	0	1	1

Figure 3.3.1: Sample Dataset

```
# Import all the tools we need
```

```
# Regular EDA (exploratory data analysis) and plotting libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
from google.colab import drive
drive.mount('/content/drive')
```

```
#import the dataset
df = pd.read_csv("/content/drive/MyDrive/COPD Paper/CSV FILE/COPD.csv")
df.shape
df.head()
```

Figure 3.3.2: Packages

```
#check the name of the column
df.columns

Index(['Name', 'Age (yrs)', 'Sex(m=0,f=1)', 'Cough(yes=1,no=0)',
       'Breathe Shortness(yes=1,no=0)', 'Inhaler(yes=1,no=0)',
       'Smoker(yes=1,no=0)', 'Wheeze(yes=1,no=0)', 'Chest pain(yes=1,no=0)',
       'Respiration rate(child 24-40=1 or 0,adult 12-20=1 or 0)',
       'FEV1/FVC (>70% 1, <34% 2, 60-69% 3,50-59% 4,35-49% 5)',
       'FVC ( > = 80% 1, 80%> 2)', 'Occupation(industrial place=1 for not 0)',
       'Restless', 'COPD(pos=1,neg=0)'],
      dtype='object')
```

Figure 3.3.3: Attributes of Dataset

### 3.3 Data Preprocessing and Visualization:

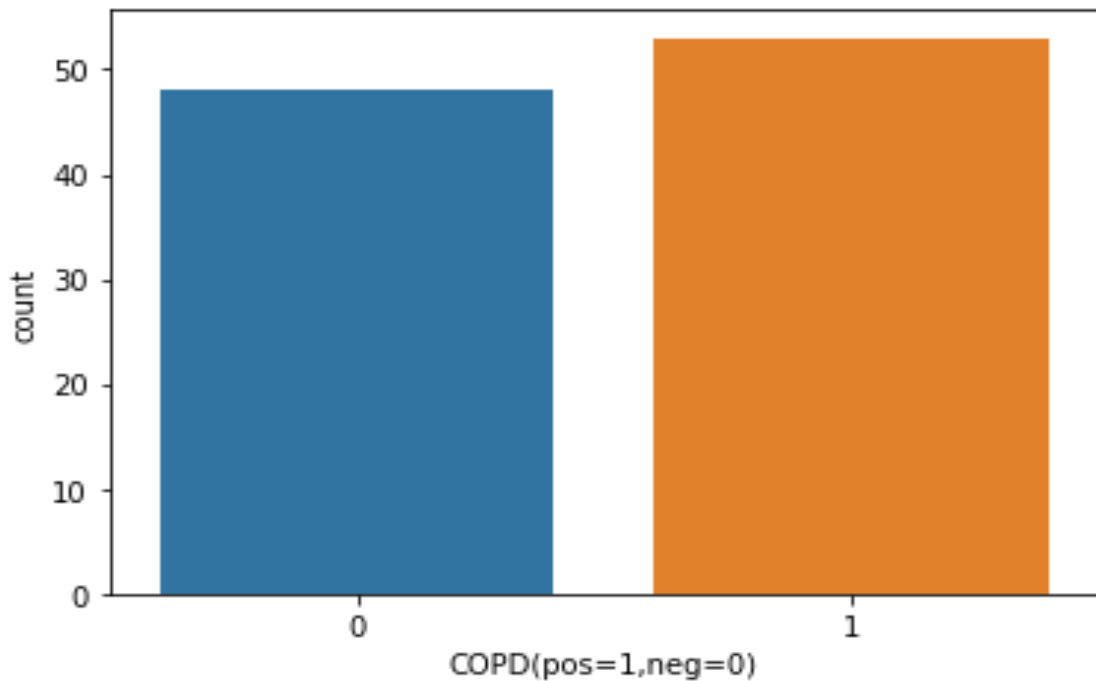


Figure 3.3.4: Positive and negative data visualization

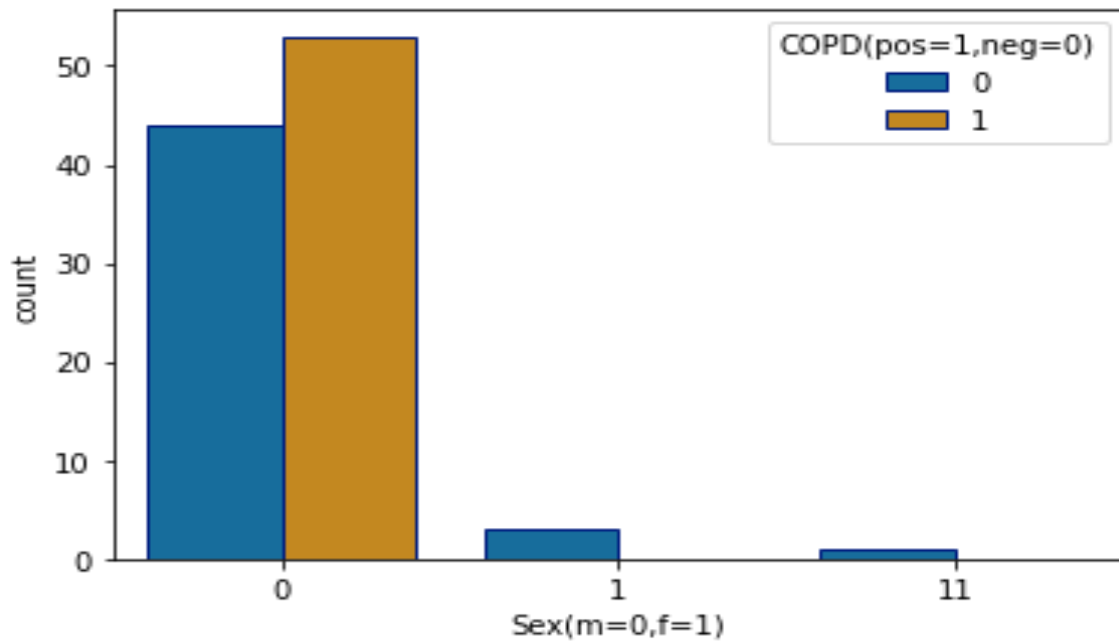


Figure 3.3.5: Gender Data visualization

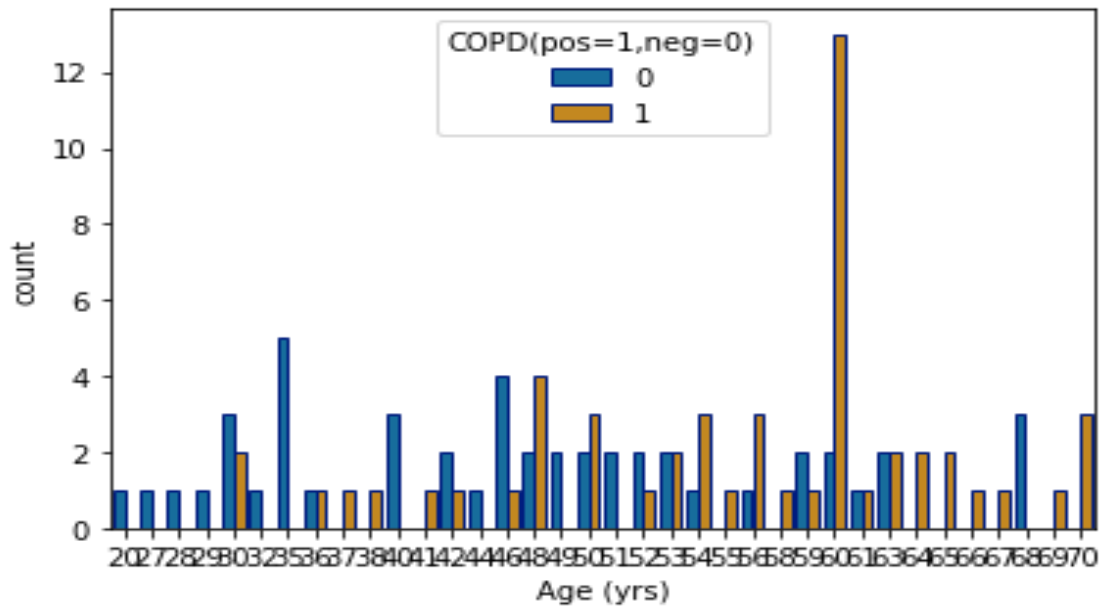


Figure 3.3.6: Age Data visualization

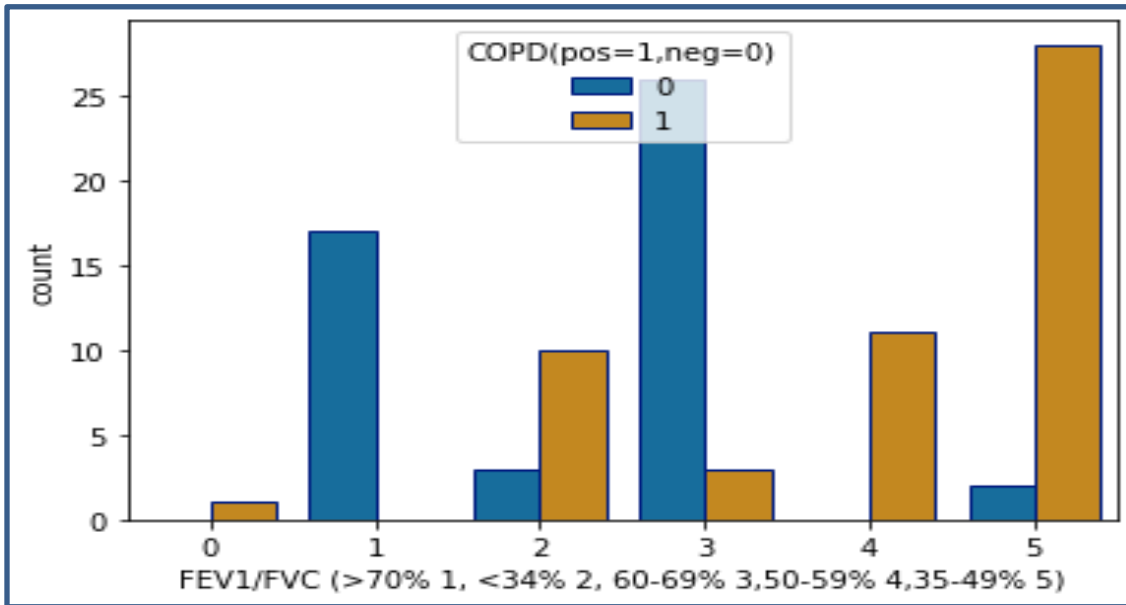


Figure 3.3.7: FEV1/FVE

Age (yrs)	sex(n=0, f=1)	Cough(yes=1, no=0)	Breathe shortness(yes=1, no=0)	Inhaler(yes=1, no=0)	Smoker(yes=1, no=0)	wheeze(yes=1, no=0)	Chest pain(yes=1, no=0)	Respiration rate(child 24-40=1 or 0, adult 12-20=1 or 0)	FEV1/FVC (>70% 1, <34% 2, 60-69% 3, 50-59% 4, 35-49% 5)	FVC (>= 80% 1, <80% 2)	occupation(industrial place=1 for not 0)	Restless	
0	50	0	1	1	1	1	1	0	0	2	2	0	1
1	48	0	1	1	0	0	0	1	1	1	1	1	1
2	55	0	1	1	0	1	1	1	1	0	4	2	0
3	30	0	1	1	0	1	1	0	1	3	1	0	0
4	56	0	1	1	1	0	0	0	1	5	2	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
96	42	0	1	0	0	1	1	0	0	2	1	1	1
97	48	0	1	1	1	1	1	0	0	5	2	0	1
98	51	0	1	0	0	1	0	0	0	2	0	1	1
99	60	0	1	1	0	1	0	1	0	4	2	0	1
100	35	0	1	1	1	0	1	0	0	5	2	1	1

101 rows x 13 columns

Figure 3.3.8: Data Visualization

### 3.4 Applied Mechanism:

We used several kinds of machine learning algorithms. We trained our dataset by using the algorithms. After that we find different accuracy for different algorithms on the testing dataset. The list of algorithms we have implemented:

- a. Logistic Regression
- b. Decision Tree Classifier
- c. K-Nearest Neighbor
- d. Navie Bayes

a. Logistic Regression: In our model first of all we implemented Logistic Regression Algorithm. There are some popular Machine Learning algorithms and Logistic regression is one of them, which works under the Supervised Learning technique. The mechanism of Logistic Regression is used for predicting the categorical dependent variable using a given set of separate variables.

Logistic Regression mainly be used to classify monitoring using different types of data and can easily figure out the variables used for the most effective classification. The below figure is showing the logistic function:

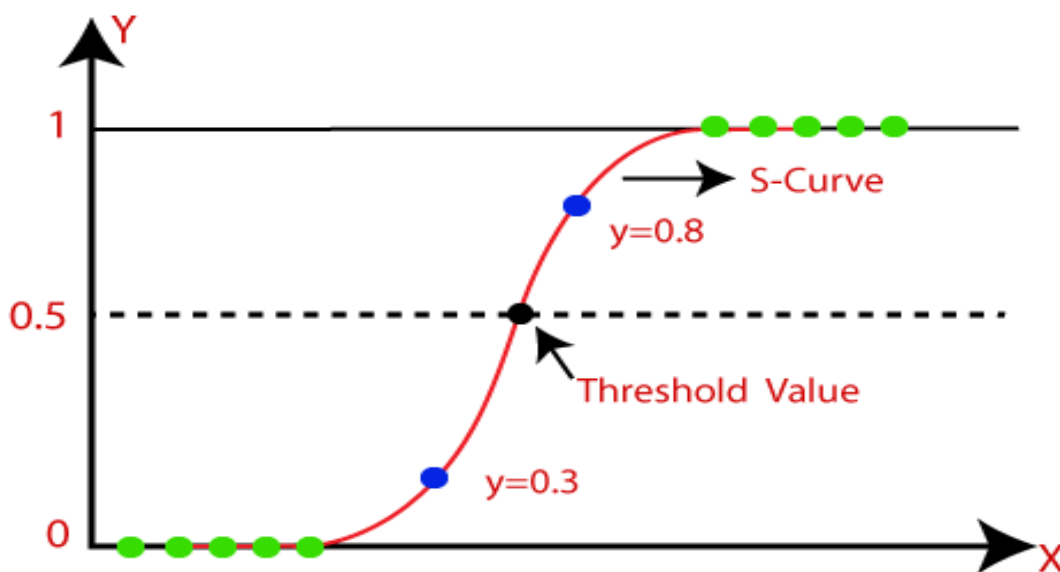


Figure 3.4.1: Logistic Regression Classification

In this case we have used the sklearn logistic regression Classifier python library to apply the algorithm and we get 95.2 % accuracy in our testing dataset using Logistic Regression algorithm. This algorithm provides max accuracy than the other algorithm. We also got our Confusion Matrix heatmap where we can measure the performance of our Logistic Regression Model. The Confusion Matrix is given below “Figure 3.4.2”.

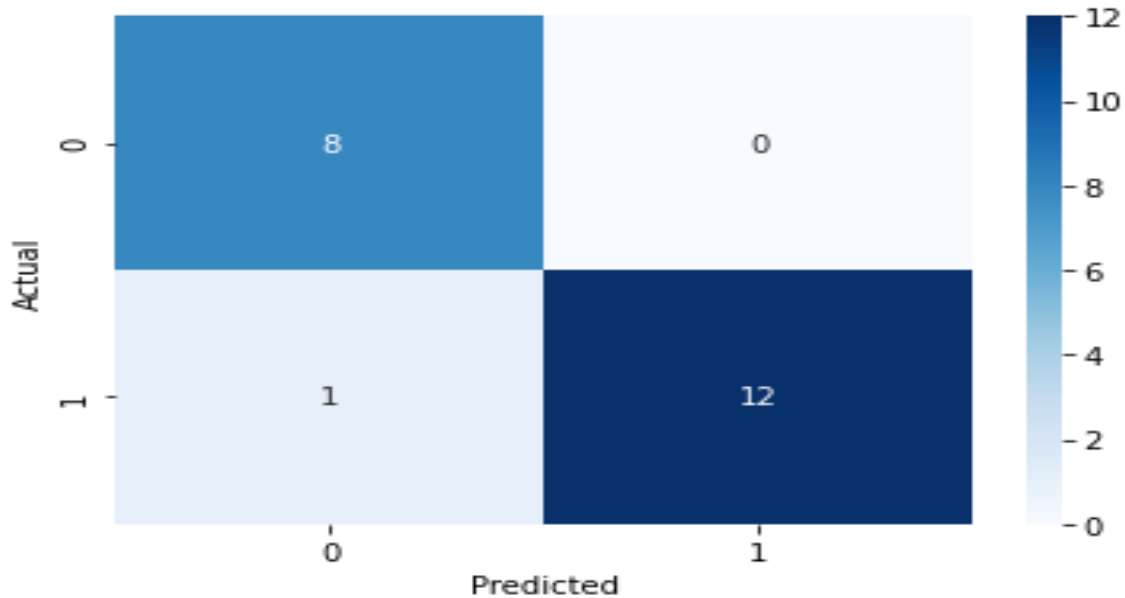


Figure 3.4.2: Confusion heatmap of Logistic Regression

The Logistic regression has performed well here. Our Logistic regression model has easily learnt from the training data and made the group of similar data points. And this Logistic regression algorithm is also designed for this kind of labels classification.

b. Decision Tree Classifier: Our second implemented algorithm is Decision Tree Classifier. We have used this algorithm for representation to find the solution of the problem in which each leaf node correlate to a class label and attributes are expressed on the internal node of the tree. The below image shows prediction score of Decision Tree:

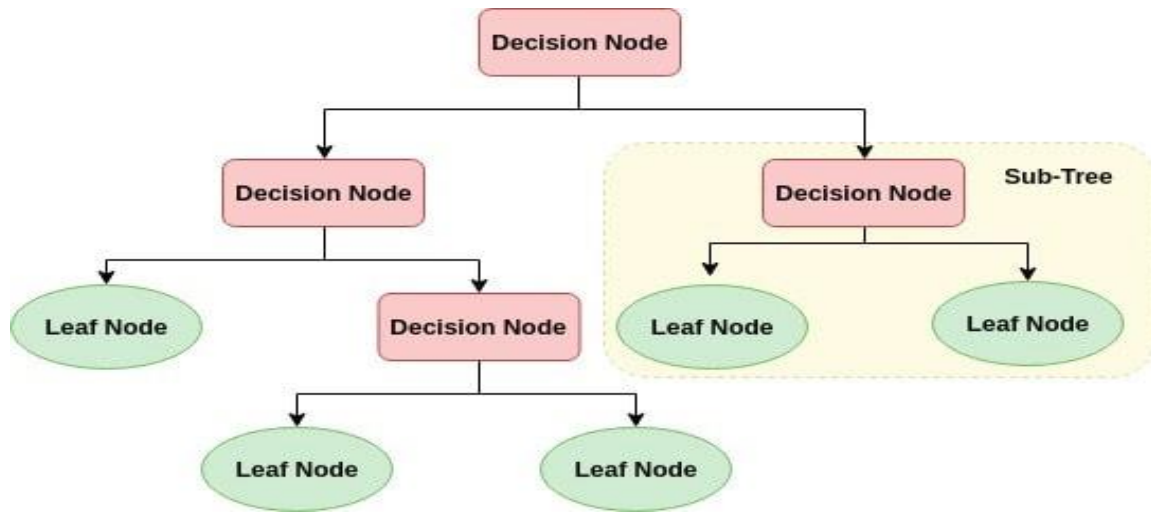


Figure 3.4.3: Prediction score of Decision Tree

In this case we have used the Decision Tree Classifier python library to apply the algorithm and we get 90.4 % accuracy in our testing dataset using Decision Tree Classifier algorithm. We also got our Confusion Matrix heatmap where we can check the output level of our Decision Tree Classifier Model. The Confusion Matrix is given below “Figure 4.1.4”.

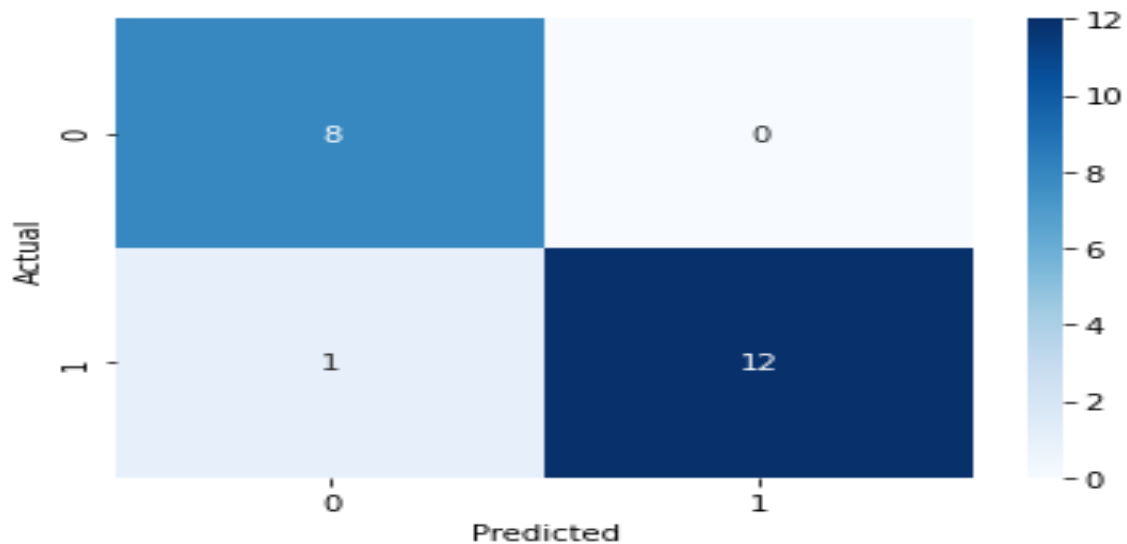


Figure 3.4.4: Confusion heatmap of Decision Tree



The Decision Tree has good prediction score. Our Logistic regression model has easily learnt from the training data and made the group of similar data points.

c. K-Nearest Neighbors (KNN): In Our third implementation algorithm, for both regression and classification problems we can use K-Nearest Neighbors (KNN) algorithm which is also a supervised machine learning algorithm. This algorithm learns from the training data based on distance metrics to classify labels on testing data.

In this case we have used the sklearn. neighbors KNeighborsClassifier python library to apply the algorithm and we achieved 62% accuracy in our testing dataset using this KNN algorithm. This algorithm provides less accuracy than the other algorithm. We also got our Confusion Matrix heatmap where we can measure the performance of our KNN model. The Confusion Matrix is given below “Figure 3.4.5”.

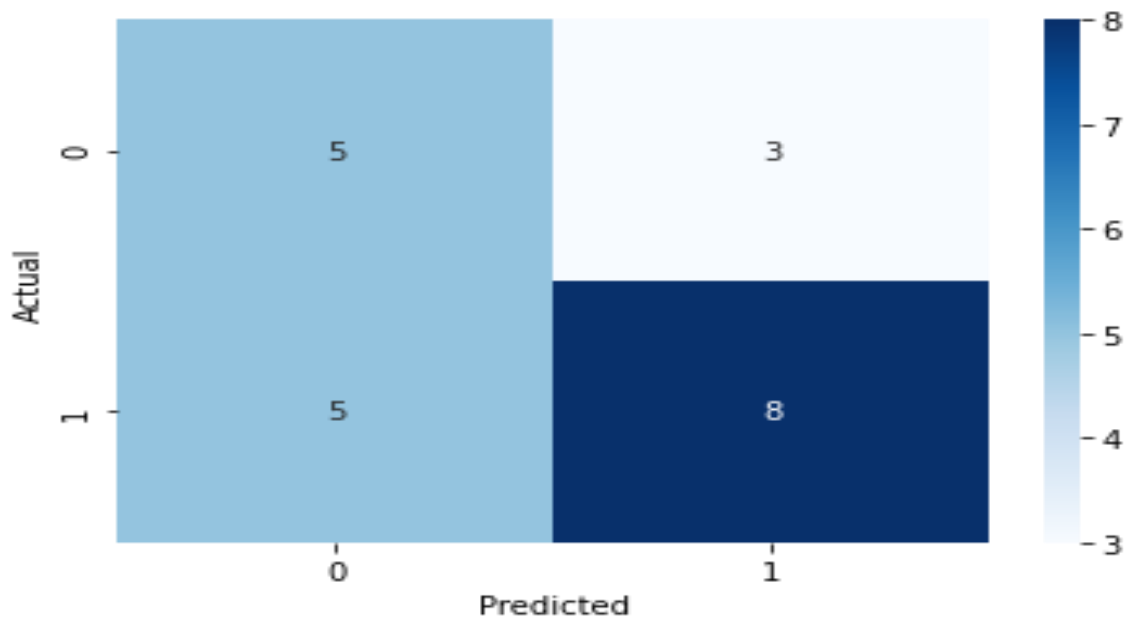


Figure 3.4.5: Confusion heatmap of K-Nearest Neighbors

d. Naive Bayes: Last of all we implemented Naïve Bayes Algorithm. We use the Naïve Bayes algorithm to predict the possibility of machine learning, it basically predicts using probability. The Naïve Bayes Classifier's aim to determine conditional probability. Naïve Bayes Algorithm works through major three steps:

1. Determine the probability of the specified class labels.
2. Consider the possibilities of specified class labels.
3. Apply Bayes Formula for calculate probability.

In this case we have used the sklearn naive\_bayes python library to apply the algorithm and we achieved 85.7 % accuracy in our testing dataset using this Naïve Bayes algorithm. This algorithm provides maximum accuracy than the previous algorithm. We also got our Confusion matrix Heatmap where we can measure the performance of Naïve Bayes. The confusion Matrix is given below “Figure 3.4.6”

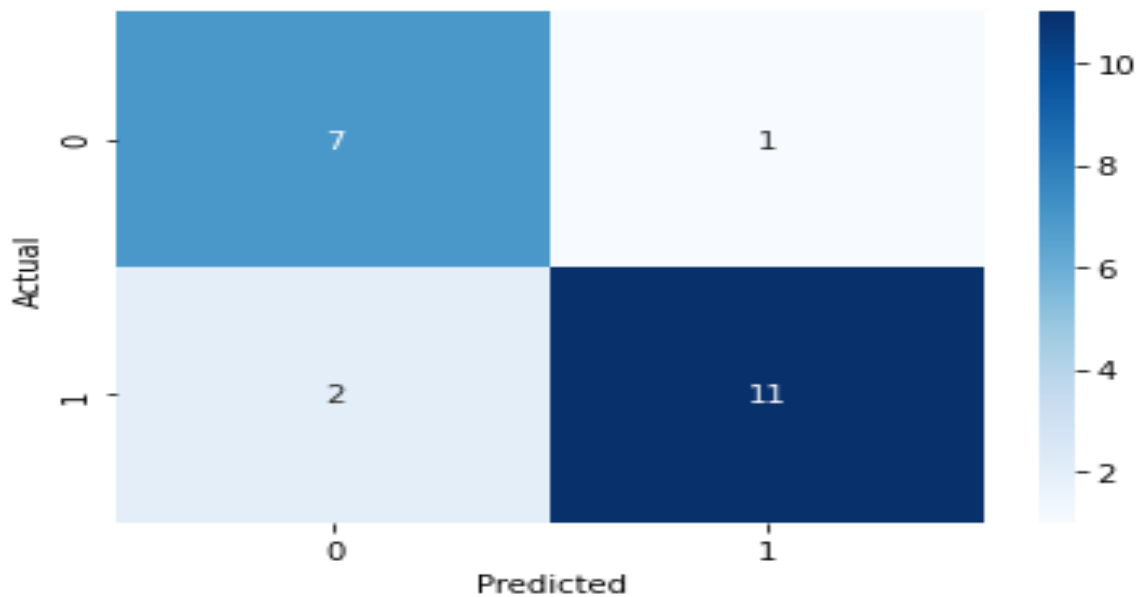


Figure 3.4.6: Confusion heatmap of Naïve Bayes

### **3.5 Statistical Analysis:**

We collected data from various kind of people. We were going many random hospitals for collecting our research. We took permission from hospital authority. They helped us as much as for our research purpose. We will keep a close eye on the data and ensure that it is not compromised in the future.

### **3.6 Implementation Requirements:**

For implementing this research project, we need some hardware and software tools. Those are mentioned in Research Project Subject and Instruction.

## CHAPTER 4

### Experimental Results and Discussion

#### 4.1 Experimental Setup:

Both supervised and unsupervised models are introduced by Machine learning, not only single model but also various types of models. Specially classification has been done by this ML model which through language processing is much more expedient to developers. Most useful four classification models we have used for finding accuracy of our model. There are Logistic Regression, Decision Tree Classifier perform with much high accuracy. But KNN Classifier and Naïve Bayes has average accuracy. Each classifier has limited discussion in the Applied Mechanism.

#### 4.2 Experimental Result & Analysis:

Logistic Regression is a momentous Machine Learning algorithm which has the ability of efficiency to provide probabilities and classify new data using continuous and discrete datasets. Already we have trained our data model in logistic Regression that has gained highest accuracy. Finally, classification report successfully predicted the outcome 95.2%. Related precision, Recall, f1-score, support is near to the accuracy of the output.

Decision Tree classifier: We come to supervised Decision Tree Classifier while we will try to create a machine learning model. Commonly it works for regression emerged problems. We have got 90.4% accuracy which is less than Logistic Regression Algorithm.

KNN is a simple classifier which train the data. Implementing KNN algorithm we have got lowest accuracy from our model. We have got 62% accuracy from this algorithm.

For probability of conditional classifier in machine learning Naïve Bayes is best. Using NB, we have trained our data model which has gained 85.7% accuracy. The classifier NB in Machine Learning which is used for coding python libraries in multinomial NB where confusion matrix is [7 11] and [1 2]. Afterwards Y test set used an accuracy which is very

perfect. Here, Related precision, recall, f1-score, support is near to the accuracy of output. In our project we used 80% of the data set for trains and 20% of data set for testing to determine model performance. The Logistic regression algorithm provides highest accuracy. The performance Table For all algorithm are given below “Table 4.2.1,4.2.2,4.2.3,4.2.4”:

	Precision	Recall	F1-Score	Support
0	0.89	1.00	0.94	8
1	1.00	0.92	0.96	13
Accuracy			0.95	21
Marco Average	0.94	0.96	0.95	21
Weighted Average	0.96	0.96	0.95	21

Table4.2.1: Performance Table for Logistic Regression

	Precision	Recall	F1-Score	Support
0	0.89	1.00	0.94	8
1	1.00	0.92	0.96	13
Accuracy			0.95	21
Marco Average	0.94	0.96	0.95	21
Weighted Average	0.94	0.95	0.95	21

Table4.2.2: Performance Table for Decision Tree

	Precision	Recall	F1-Score	Support
0	0.78	0.88	0.82	8
1	0.92	0.85	0.88	13
Accuracy			0.86	21
Marco Average	0.85	0.86	0.85	21
Weighted Average	0.96	0.95	0.86	21

Table4.2.3: Performance Table for Naïve Bayes

	Precision	Recall	F1-Score	Support
0	0.50	0.62	0.56	8
1	0.73	0.62	0.67	13
Accuracy			0.62	21
Marco Average	0.61	0.62	0.61	21
Weighted Average	0.64	0.62	0.62	21

Table4.2.4: Performance Table for KNN

### 4.3 Discussion:

In our project we applied four classification machine learning models which through we get measurable output for our dataset. These four Classifier perform well but one them do better between the classifiers.

Now begin the application of classification that creates actual performance on Logistic Regression classifier = 95.2%. So, we can say that this algorithm is best for our dataset.

The figure will so the accuracy of all algorithm:

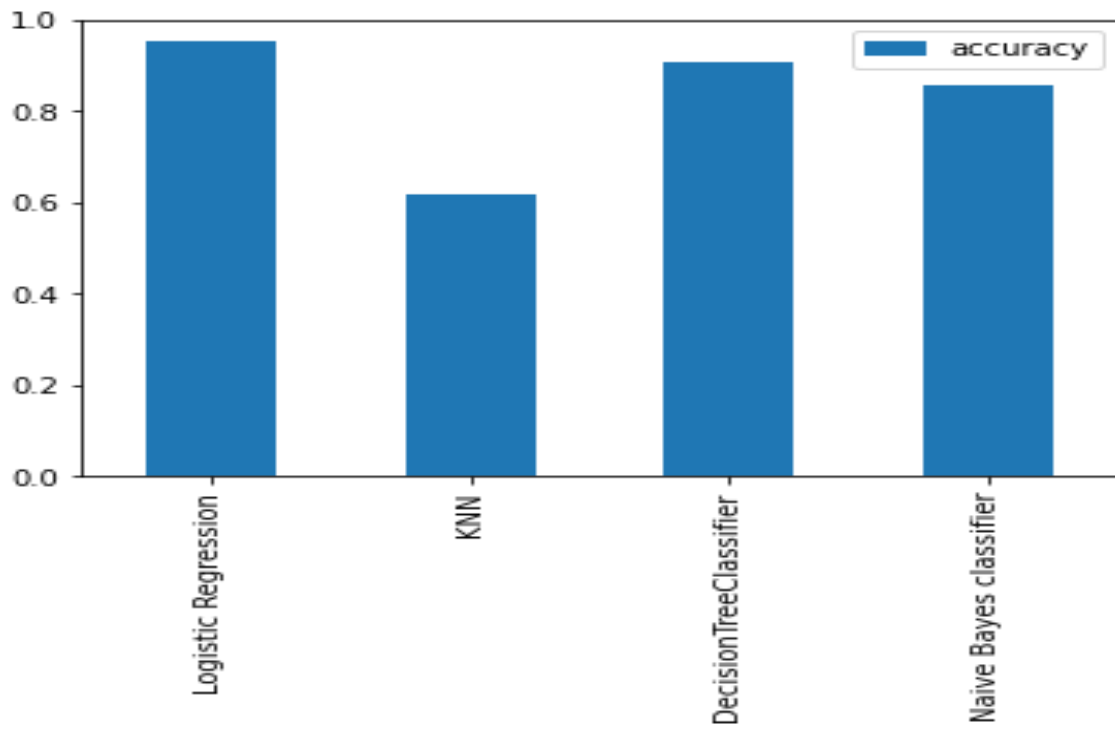


Figure3.4.7: Accuracy of all applied Algorithm

## **CHAPTER 5**

### **Impact on Society, Impact on Environment, Sustainability**

#### **5.1 Impact on Society:**

Chronic Obstructive Pulmonary Disease has pragmatic impact on society. This Disease is increasing in our society day by day. A huge part of people is suffering for this disease. COPD has many kinds of symptoms in our country basis. COPD has many kinds of symptoms in our country basis. People are leading towards to death. This Project can be implemented in health department.

#### **5.2 Impact on Environment:**

There will be no direct impact of our research on environment. We used patient's data which are always in safe. The mankind will be benefited by our Project.

#### **5.3 Sustainability Plan:**

Sustainability plan is must for any solid project idea so that can be implemented in real life. Our research mission is to predict COPD. This study aims to show no harm to humans and humans. This Project can be implemented in real life and it can be used in a bigger scale.



## CHAPTER 6

### Summary, Conclusion and Implication for future Research

#### 6.1 Summary of the study:

In this section we discussed about the experiment of our research project. To create a summary, we need all the response of the Machine Learning algorithm. All the discussion of our project in upper that provide briefly in details in above. As for now we are convinced by the fact that COPD can be predict by machine learning as our dataset. Similar work has not been done yet in Bangladesh.

#### 6.2 Conclusions:

In our research, from many other methods we choose a machine learning based approach. We wanted to start a process for peoples that they will know and aware of COPD. We are hopeful for our work as we wanted to create something new and impactful for general well-being.

#### 6.3 Implication for Future Study:

Technology is updating every other minute worldwide. To be updated and cope up with that we always need to go with the flow and keep progressive. Now, it is on a small scale and we are hopeful for further development of this Project for industrial use. We have also some ideas regarding this Project which we can implement in the future development.

- a. We will build a user input, doctors will easily sure a predictable percentage to detect COPD.
- b. In future we will a system which will used by the Government.

## References

- [1] Swaminathan, S., Qirko, K., Smith, T., Corcoran, E., Wysham, N. G., Bazaz, G., ... & Gerber, A. N. (2017). A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PloS one*, 12(11), e0188532.
- [2] Himes, B. E., Dai, Y., Kohane, I. S., Weiss, S. T., & Ramoni, M. F. (2009). Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, 16(3), 371-379.
- [3] Sanchez-Morillo, D., Fernandez-Granero, M. A., & Leon-Jimenez, A. (2016). Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review. *Chronic respiratory disease*, 13(3), 264-283.
- [4] Chang, C. L., Sullivan, G. D., Karalus, N. C., Mills, G. D., McLachlan, J. D., & Hancox, R. J. (2011). Predicting early mortality in acute exacerbation of chronic obstructive pulmonary disease using the CURB65 score. *Respirology*, 16(1), 146-151.
- [5] Spathis, D., & Vlamos, P. (2019). Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health informatics journal*, 25(3), 811-827.
- [6] Amaral, J. L., Lopes, A. J., Jansen, J. M., Faria, A. C., & Melo, P. L. (2012). Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. *Computer methods and programs in biomedicine*, 105(3), 183-193.
- [7] Shah, S. A., Velardo, C., Farmer, A., & Tarassenko, L. (2017). Exacerbations in chronic obstructive pulmonary disease: identification and prediction using a digital health system. *Journal of medical Internet research*, 19(3), e69.
- [8] Postma, D. S., Löfdahl, C. G., Stephen Lam, M. D., Man, S. P., Donald Tashkin, M. D., & Wise, R. A. (2016). Individualized prediction of lung-function decline in chronic obstructive pulmonary disease.
- [9] Blumenthal, J. A., Smith, P. J., Durham, M., Mabe, S., Emery, C. F., Martinu, T., ... & Palmer, S. (2016). Biobehavioral prognostic factors in chronic obstructive pulmonary disease: results from the INSPIRE-II trial. *Psychosomatic medicine*, 78(2), 153.
- [10] McCarthy, B., Casey, D., Devane, D., Murphy, K., Murphy, E., & Lacasse, Y. (2015). Pulmonary rehabilitation for chronic obstructive pulmonary disease. *Cochrane database of systematic reviews*, (2).
- [11] Sutradhar, I., Gupta, R. D., Hasan, M., Wazib, A., & Sarker, M. (2019). Prevalence and risk factors of chronic obstructive pulmonary disease in Bangladesh: a systematic review. *Cureus*,

## Chronic\_Obstructive\_pulmonary\_Disease\_prediction\_Using\_...

### ORIGINALITY REPORT

<b>21</b> %	<b>15</b> %	<b>8</b> %	<b>9</b> %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

### PRIMARY SOURCES

<b>1</b>	Nushrat Jahan Ria, Sharun Akter Khushbu, Mohammad Abu Yousuf, Abu Kaisar Mohammad Masum, Sheikh Abujar, Syed Akhter Hossain. "Toward an Enhanced Bengali Text Classification Using Saint and Common Form", 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020 Publication	<b>4</b> %
<b>2</b>	Submitted to Daffodil International University Student Paper	<b>3</b> %
<b>3</b>	<a href="https://dspace.daffodilvarsity.edu.bd/8080">dspace.daffodilvarsity.edu.bd:8080</a> Internet Source	<b>3</b> %
<b>4</b>	<a href="http://www.cureus.com">www.cureus.com</a> Internet Source	<b>2</b> %
<b>5</b>	Submitted to University of West Florida Student Paper	<b>1</b> %
<b>6</b>	<a href="http://www.sas.com">www.sas.com</a> Internet Source	<b>1</b> %

[ymgsapo.com](http://ymgsapo.com)

7	Internet Source	1 %
8	www.featureranking.com Internet Source	1 %
9	en.wikipedia.org Internet Source	1 %
10	journals.plos.org Internet Source	1 %
11	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	<1 %
12	Submitted to University of Buraimi Student Paper	<1 %
13	iacthealth.com Internet Source	<1 %
14	in.springboard.com Internet Source	<1 %
15	Submitted to TechKnowledge Student Paper	<1 %
16	"Cyber Security and Computer Science", Springer Science and Business Media LLC, 2020 Publication	<1 %
17	Rejwana Haque Pial, Mohammad Rashidul Hashan, Sherief Ghazy, Mahmoud Dibas, Amr	<1 %

Ehab El-Qushayri, Mohamed M. Abdel-Daim.  
"Comparative study on respiratory function  
among rural women using biomass fuel and  
non-biomass fuel: evidence of a cross-  
sectional survey in Bangladesh",  
Environmental Science and Pollution  
Research, 2020  
Publication

---

18 [www.coursehero.com](http://www.coursehero.com)  
Internet Source

---

<1%

---

Exclude quotes On

Exclude matches < 10 words

Exclude bibliography On