**LUNG CANCER PREDICTION**

A Course Project report submitted

in partial fulfilment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

in

**COMPUTER SCIENCE & ENGINEERING**

by

**D.SAKEETH (2003A51086)**

**B. SOWMYA (2003A51114)**

**J. BHAVANA (2003A51102)**

**M. SANDEEP (2003A51092)**

Under the guidance of

**Mr. RAMESH DADIR**

Assistant Professor, Department of CSE.

**

**Department of Computer Science and Artificial Intelligence**

**

**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that this project entitled **“LUNG CANCER PREDICTION**” is the bonafide work carried out by **D. Sakeeth, B. Sowmya, J. Bhavana, M. Sandeep** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **COMPUTER SCIENCE & ENGINEERING** during the academic year 2021-2022 under our guidance and Supervision.

**Mr. RAMESH DADIR Dr. M.Sheshikala**

Assistant Professor, Department of CSE Assoc. Prof. & HOD (CSE),

S R University, S R University,

Ananthasagar, Warangal. Ananthasagar, Warangal.

**ACKNOWLEDGEMENT**

We express our thanks to Course co-coordinator **Mr.** **RAMESH DADIR, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude toHead of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, C. V. Guru Rao, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**ABSTRACT**

Cancer has identified a diverse condition of several various subtypes. The timely screening and course of treatment of a cancer form is now a requirement in early cancer research because it supports the medical treatment of patients. Many research teams studied the application of ML and Deep Learning methods in the field of biomedicine and bioinformatics in the classification of people with cancer across high- or low-risk categories. These techniques have therefore been used as a model for the development and treatment of cancer. As, it is important that ML instruments are capable of detecting key features from complex datasets. Many of these methods are widely used for the development of predictive models for predicating a cure for cancer, some of the methods are artificial neural networks (ANNs), support vector machine (SVMs) and decision trees (DTs). While we can understand cancer progression with the use of ML methods.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **CHAPTER NAME** | **PAGE NO** |
| 1 | Introduction | 6 |
| 2 | Literature Survey | 7 |
| 3 | Methodology | 9 |
| 4 | Implementation | 12 |
| 5 | Analysis & Results | 23 |
| 6 | Conclusion | 24 |
| 7 | References | 25 |

**1.INTRODUCTION**

The main weight of ailment overall is as Lung malignancy that is the most inescapable disease in the two men and women. A few other reports estimate some 221,200 new cases of pulmonary cancer occur and represent approximately 13% of all cancer diagnoses in 2015. Approximately 27 percent of all cancer deaths are attributed to lung cancer. Lung nodules must therefore be closely examined and monitored when at an early stage. The predictive models discussed here are based on different supervised ML techniques, input and data samples.

In order to predict the various types of diseases, different deep learning & machine learning algorithms are used , such as Support vector machine (SVM), Neural Network (NN), LR, Nevin biases (NB), Fuzzy logic, transfer learning, ensemble learning, Transduction learning, KNN

**2.LITERATURE SURVEY**

A process of finding hidden knowledge from large volumes of data in databases, which perhaps comprise of data mining approaches, to recognize and make use of lung patterns and connection surrounded by a large amount of inconstant, and they allowed prognosticating the result of a disease using the previous instances kept in a set of data.

**2.1 Self Organizing Map**

A Review of Lung cancer Prediction System using Data Mining Techniques and Self Organizing Map (SOM) Bharathi et.al suggested that a prototype lung cancer system extracts hidden knowledge from a historical lung cancer disease database According to him, the most effective model to predict patients with Lung cancer disease appears to be the self-organizing map (SOM). The strength of predicting input extent on devices in which other forms are developed with a short-dimensional systematic framework that can be productively employed to traverse the effects of the data. When the numeral of SOM components is great, to ease the significant analysis of the map and the data needs to be clustered.

**2.2 CT Scan Images by Watershed algorithm**

In Lung Cancer Detection using CT Scan Images authors Suren Makajua et.al suggested the strength of determining cancer, a new system is suggested. The advanced system is used to notice the cancer Analysis of Lung Cancer 5 patient tumor from the lung CT scan image by tomography in which computer performs the moving of cells using the strength of watershed segmentation which is the mathematical morphology for observation.

**2.3 CT Scan by Artificial Neural Networks**

In Lung Cancer Detection using Machine Learning Sasikala et.al suggested in their paper that a convolutional neural network formed arrangement was applied to discern a term for diseases by defining the strength in which abnormal cells divide without control and can invade nearby tissues apprise in the chest CT scan which contains more detailed type of chest x-ray that takes many detailed pictures of your lungs and inside of your chest . They used an algorithm for supervised learning of artificial neural networks using gradient descent named as a Back-proportion algorithm by using distance and angular spatial relationship features.

**2.4 Deep CNN Algorithm**

In Lung Cancer Detection and Classification Using Deep CNN Kawsar Ahmed et.al suggested that the exploratory consequences are divided into two parts that are predicting tools to represent lung cancer and the discovery of periodic patterns. Operating the data from a large store of data accumulated from a wide range of sources, the remarkable designs are drawn for predicting Lung cancer.

**2.5 Traditional Lumbered Central Method**

In Early Detection of Lung Cancer Risk Using Data Mining F. Leena et.al In this study, a new methodology of genetic algorithm is launched to excel in the restriction of traditional lumbered central method and their strength is to search parallel from a population of points and uses probabilistic selection rules. The approach was quite established because the genetic algorithm has the ability to avoid being trapped in a locally optimal solution.

**2.6 Genetic Algorithm**

In Prediction of Lung Cancer using Data Mining Techniques Joey Mark Diaz et.al, this study prosperously arranged lung cancer position of patients on the premise of gene expression data using his strength that Genetic Algorithms to manage a process that uses data mining and probability to forecast outcomes. The representations allowed to keep the exact divination uniform accompanied by the extremely nominal numeral of characteristics picked up. To summarize, representing affiliation surrounded by the picked characteristics and inspecting how categorization show escorted by gene series is approved for later testing.

**2.7 K-Means and Farthest Clustering Algorithm**

In Lung Cancer Data Analysis by k-means and Farthest First Clustering Algorithms authors, Dharmarajan et.al the time will differ from one machine that processes to another machine that processes. His strength was algorithm Farthest first traversal that is suitable for large data sets but it creates a nonuniform cluster of a bounded metric space in a sequence and the k-Means clustering algorithm has been executed here. To conclude, the k-Means algorithm is structured for the lung cancer dataset with an Attribute relation file format. It is thoroughly convenient for the necessity clustering of cancer linked healing implementations.

**2.8 Double Convolutional Network**

Using Double Convolution Neural Network for Lung Cancer Stage Detection Goran Jakimovski et.al suggested that the initial originality in our paper is utilizing the K-means algorithm to pre-distinguish the images into a stack of identical portion images, where deep neural network contains a definite measurement of complication accompanied by further two layers. The next originality is the further complex surface with edge enhancing device for removing impurities, prior to rigorously explore for cancer.

**2.9 Data Mining Techniques**

Improves Treatment Programs of Lung Cancer Using Data Mining Techniques Zakaria et.al suggested that multi-paradigm numerical computing environment computing MATLAB used in implementing our project. Medical image permits biotic forms to be separated but at present, there is no categorization plan for advanced image processing with accuracy. They strengthen their focus on Artificial Neural Networks that contain the pieces of a computing system designed to simulate the way the human brain analyses and processes information

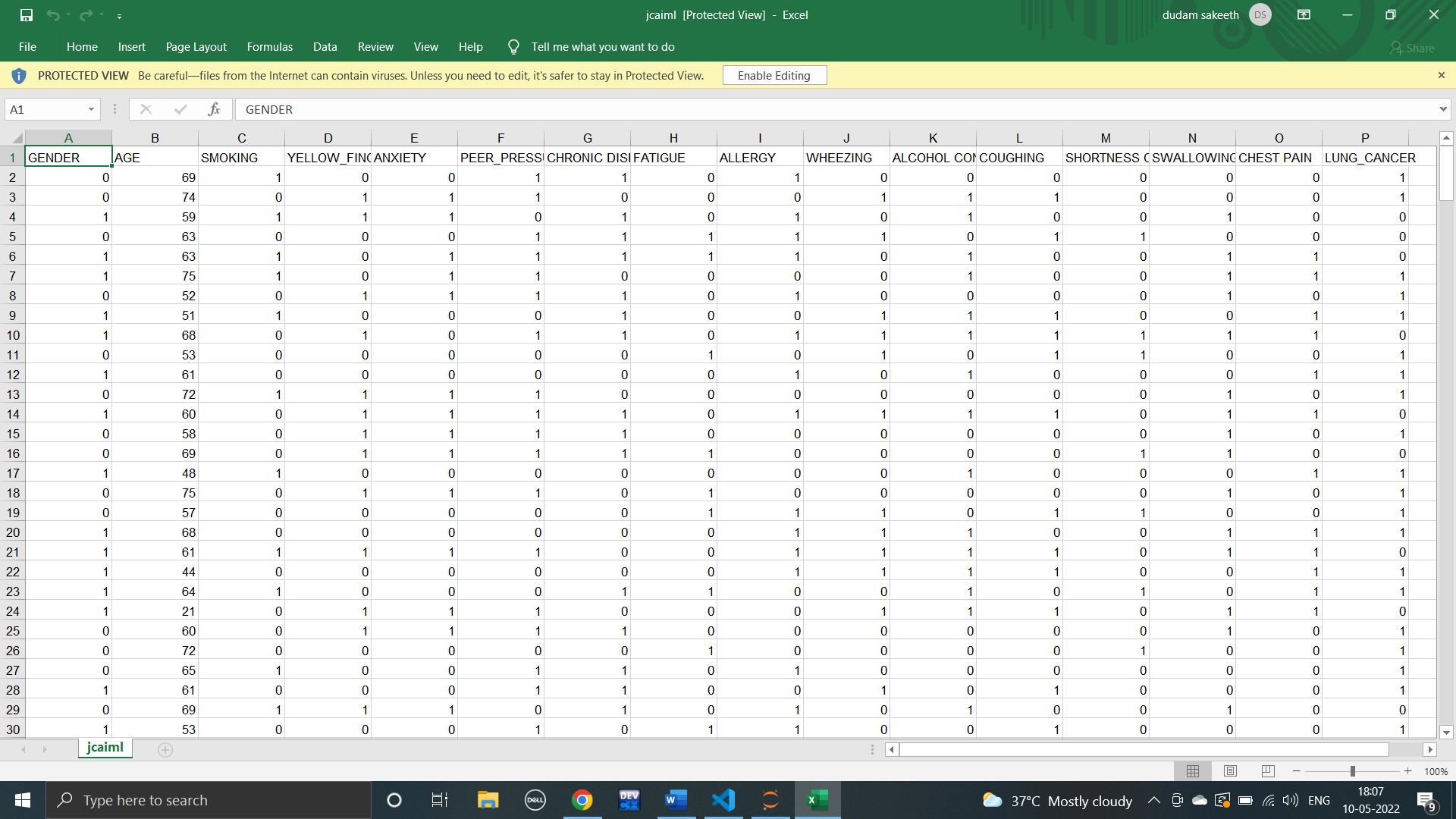
**3.METHODOLOGY**

**3.1 DATASET**

### The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status. The data is collected from the website online lung cancer prediction system.

Total no. of attributes:16  
No of instances:284  
Attribute information:

1. Gender: M(male), F(female)
2. Age: Age of the patient
3. Smoking: YES=2, NO=1.
4. Yellow fingers: YES=2, NO=1.
5. Anxiety: YES=2, NO=1.
6. Peer\_pressure: YES=2, NO=1.
7. Chronic Disease: YES=2, NO=1.
8. Fatigue: YES=2, NO=1.
9. Allergy: YES=2, NO=1.
10. Wheezing: YES=2, NO=1.
11. Alcohol: YES=2, NO=1.
12. Coughing: YES=2, NO=1.
13. Shortness of Breath: YES=2, NO=1.
14. Swallowing Difficulty: YES=2, NO=1.
15. Chest pain: YES=2, NO=1.
16. Lung Cancer: YES=0, NO=1.



**3.2 METHODS/MODELS USED**

DATA VISUALIZATION

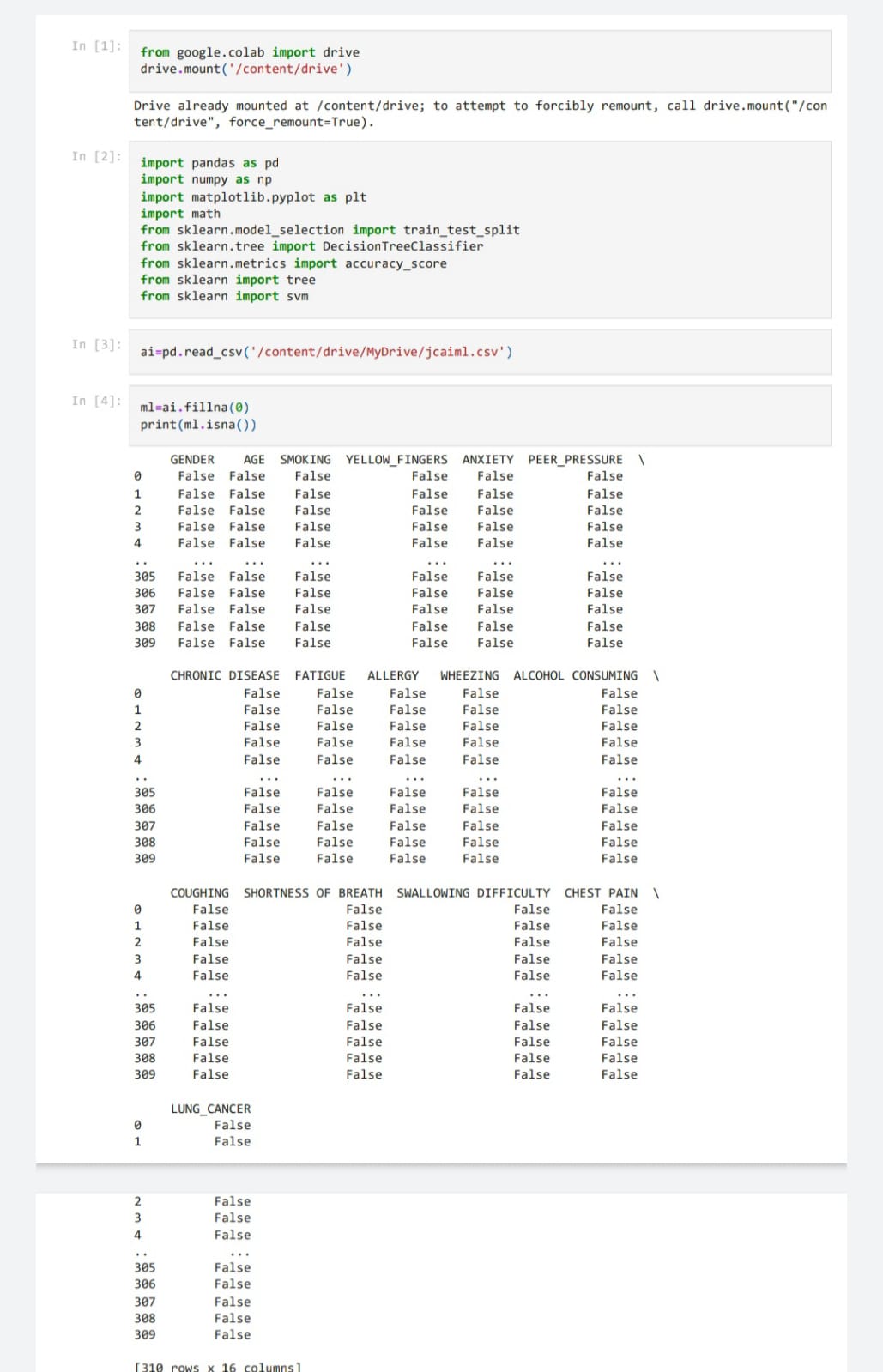
SUPPORT VECTOR MACHINE

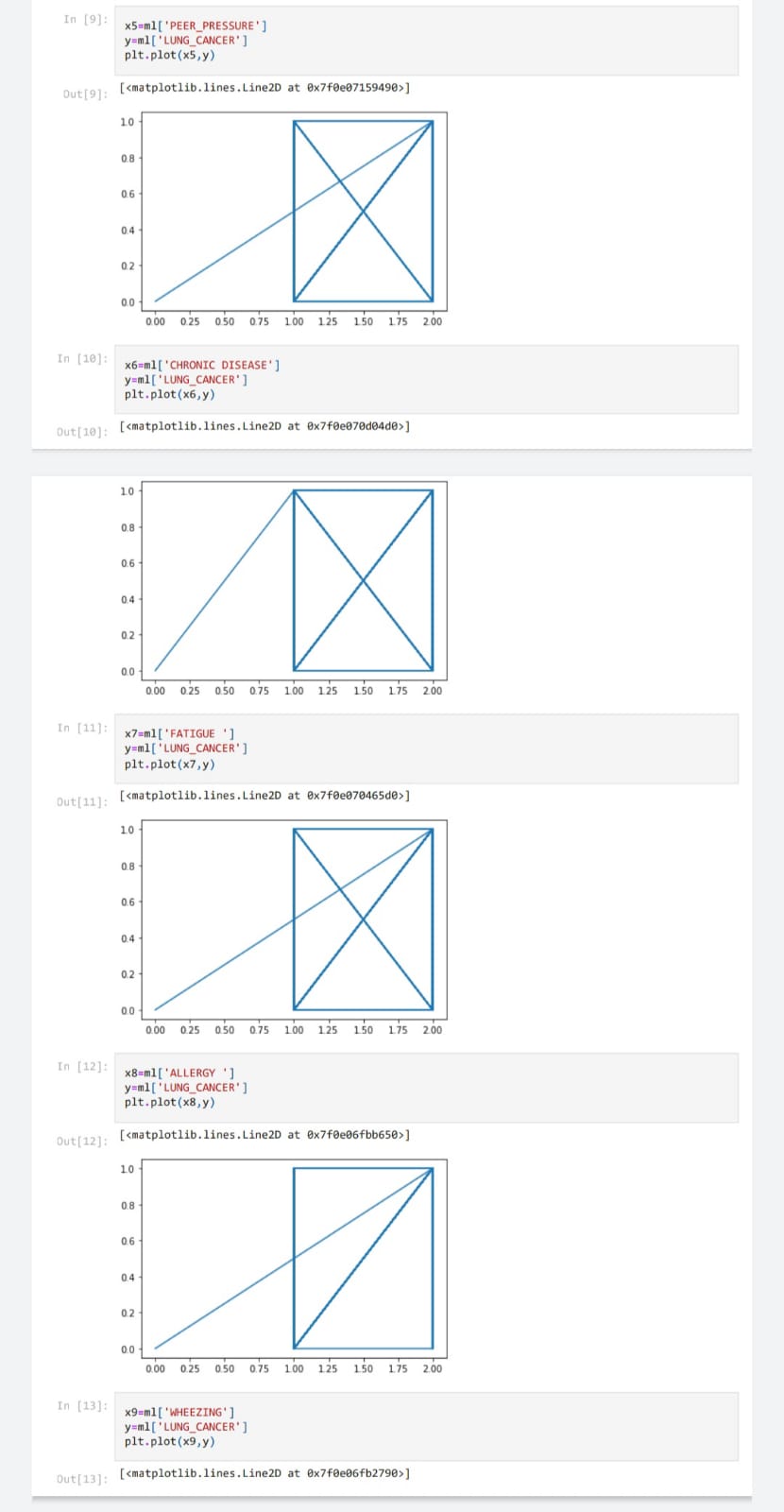
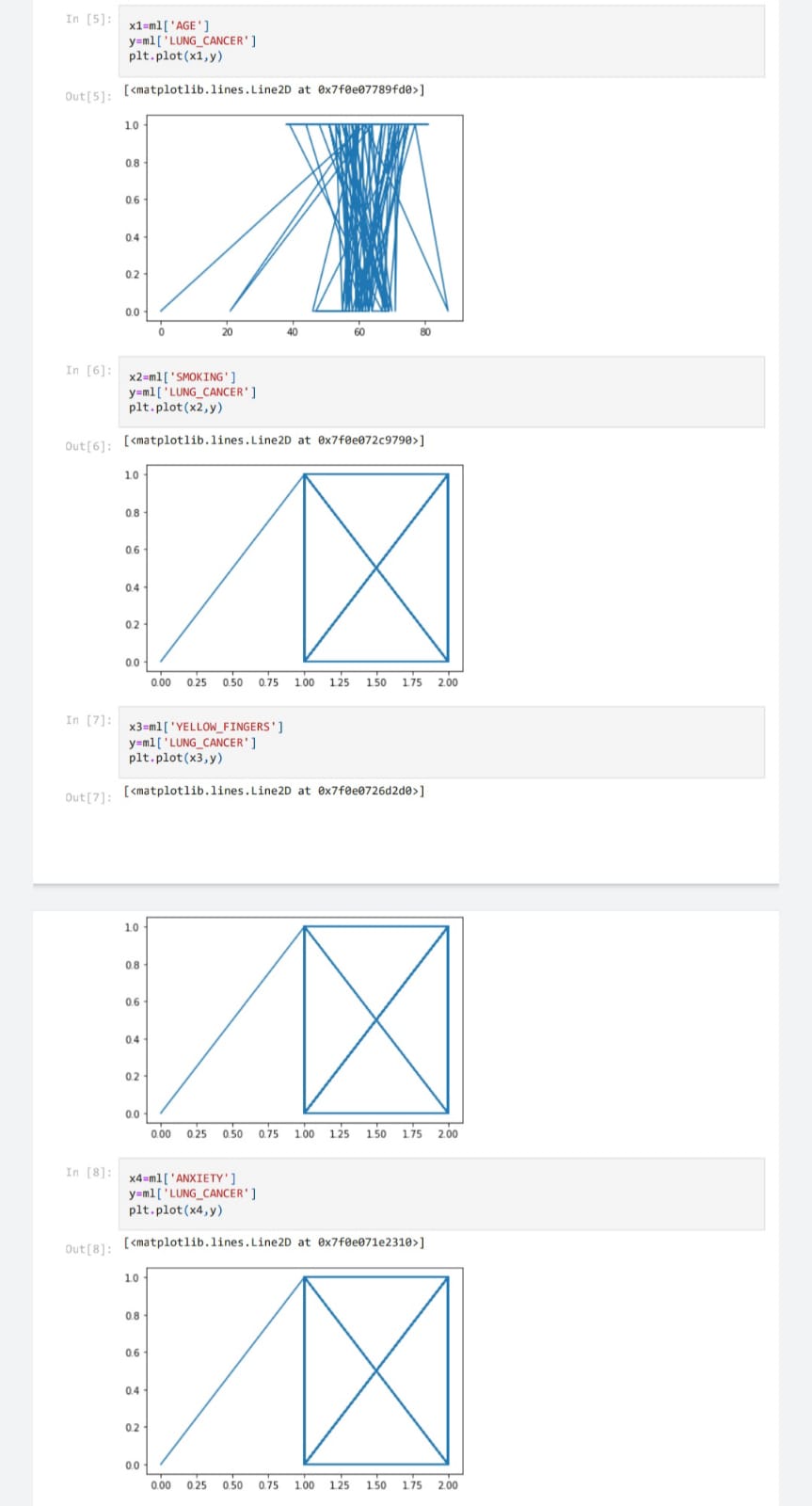
DECISION TREE

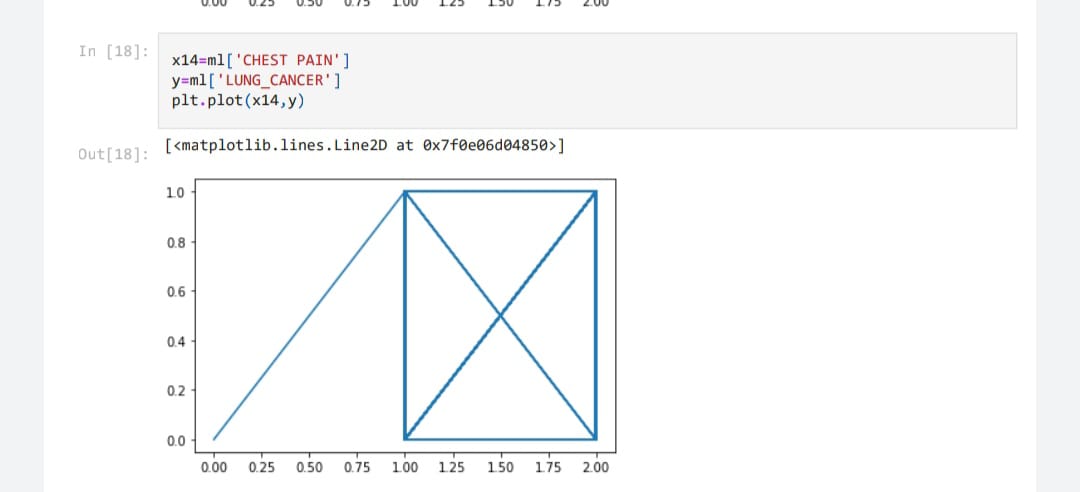
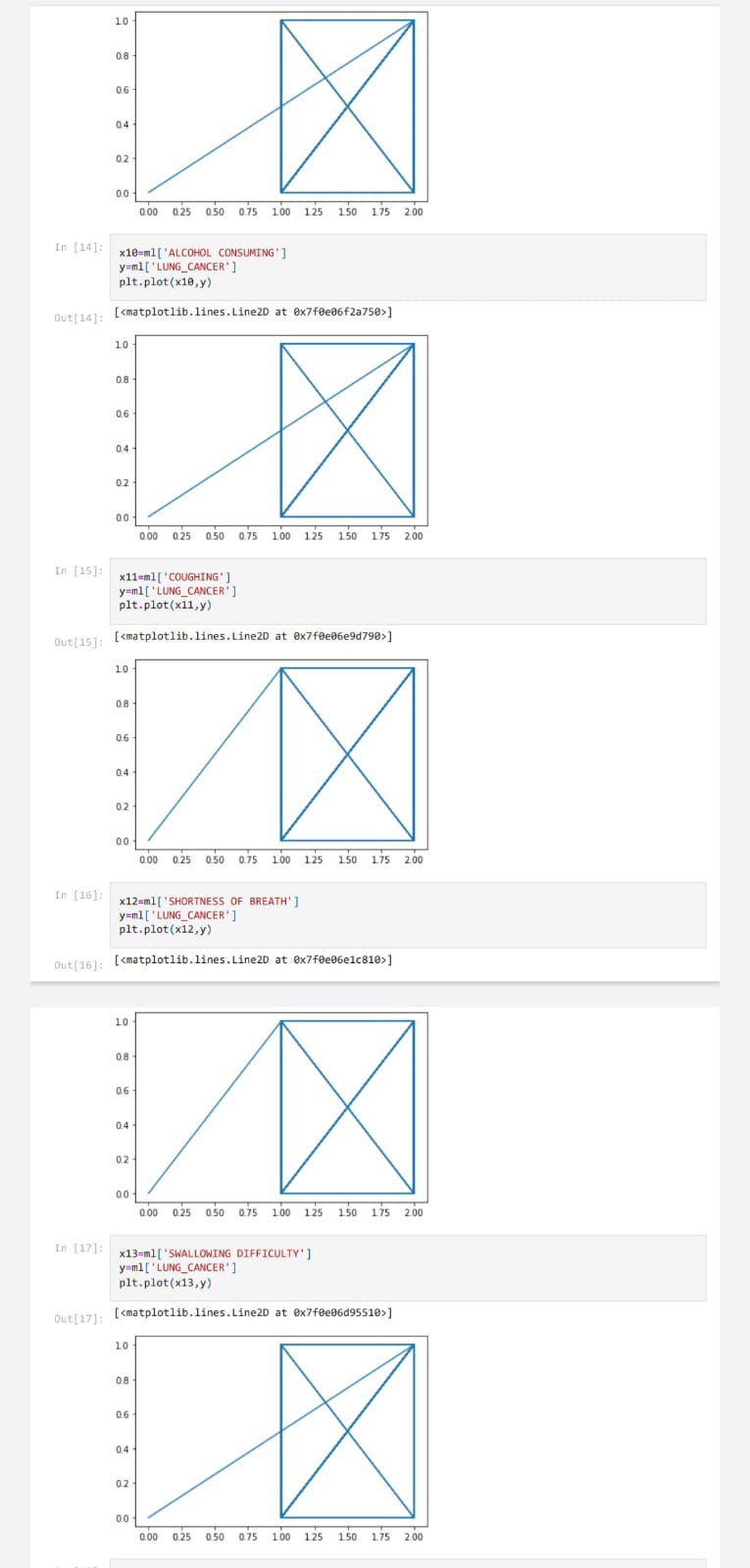
TRAINING THE DATA

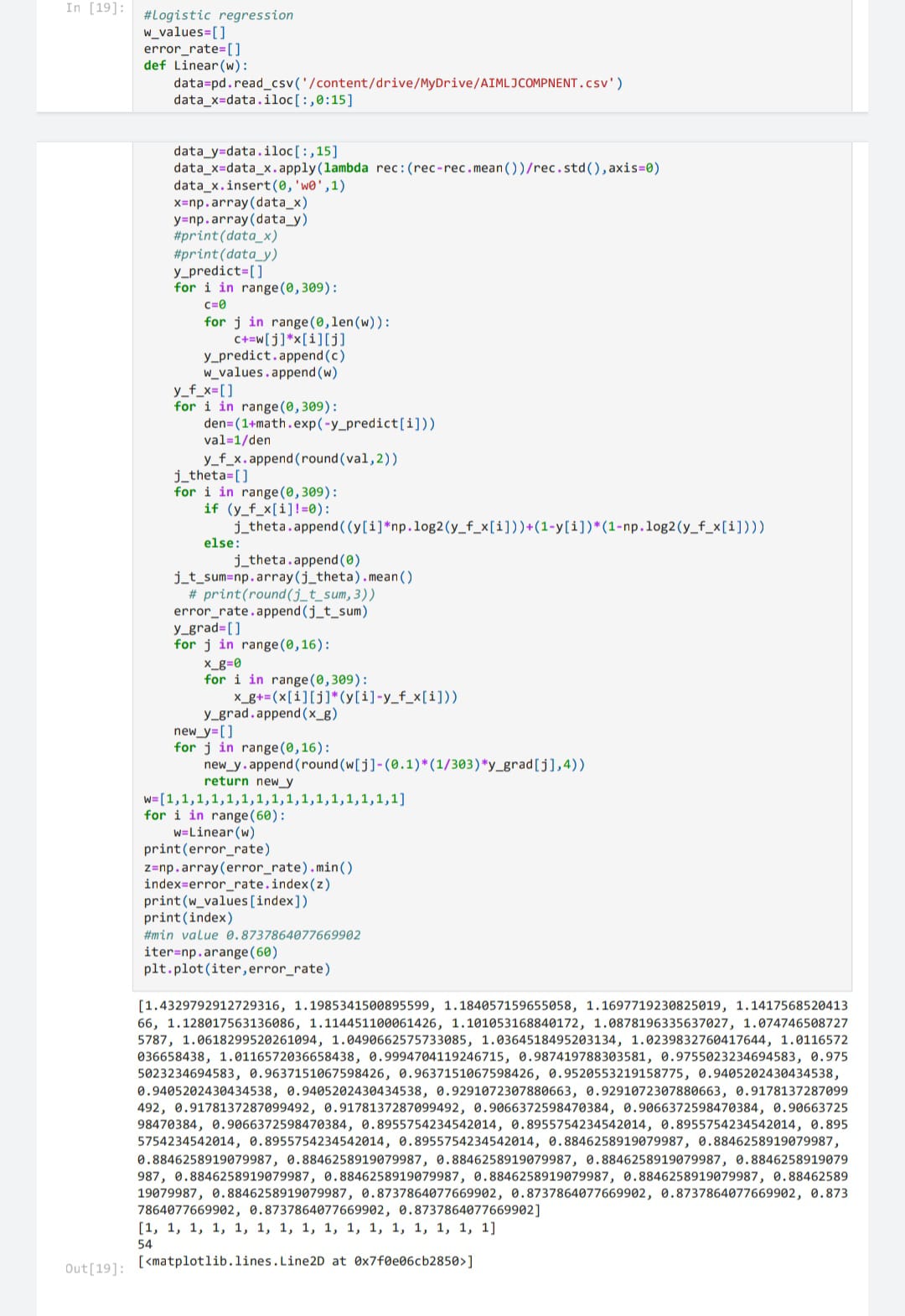
LOGISTIC REGRESSION

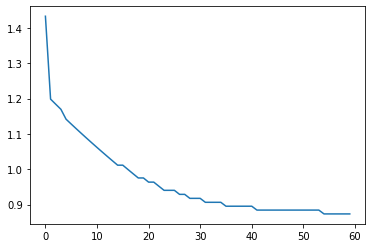
**4.IMPLEMENTATION**

**4.1 DATA VISUALIZATION**

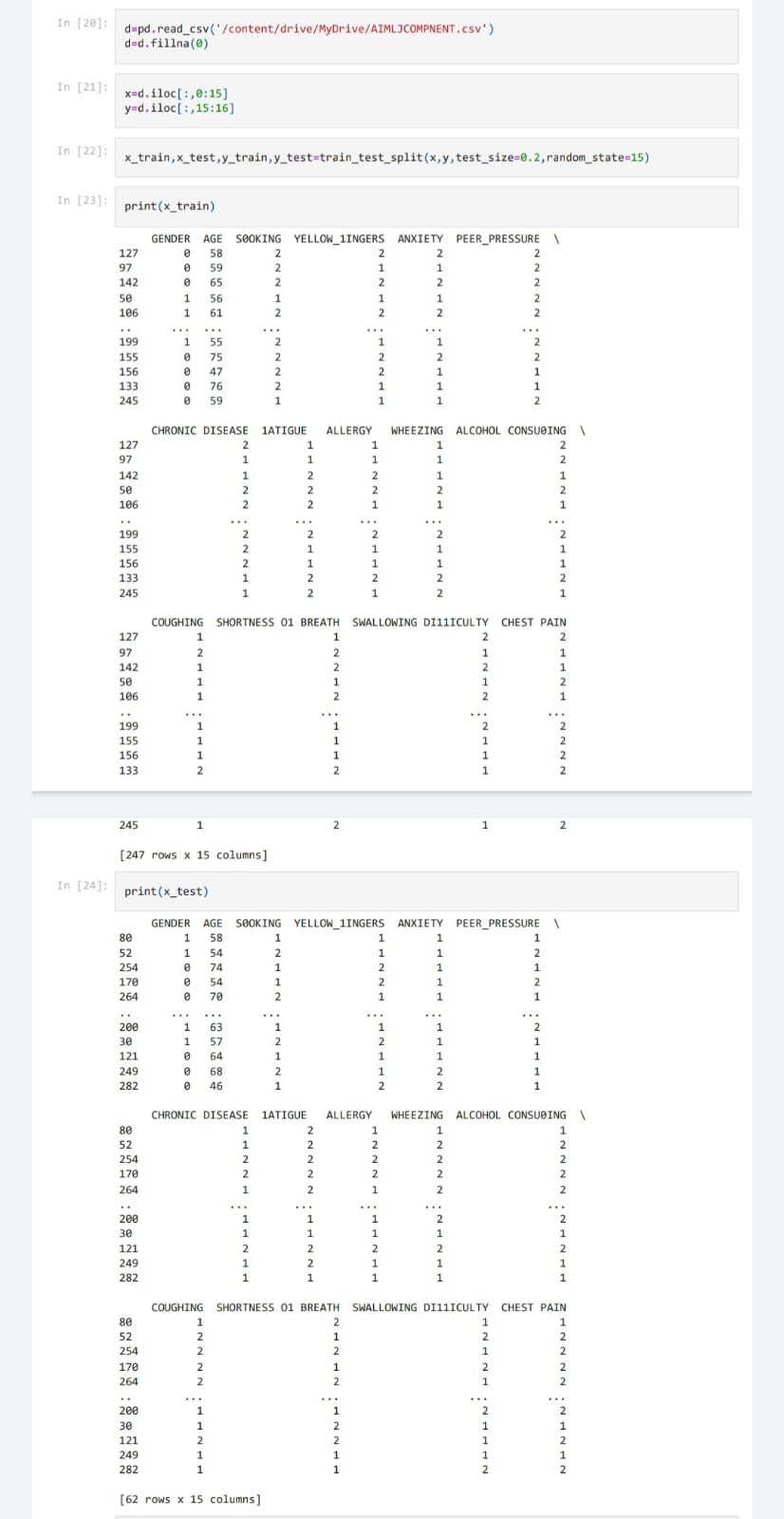
****

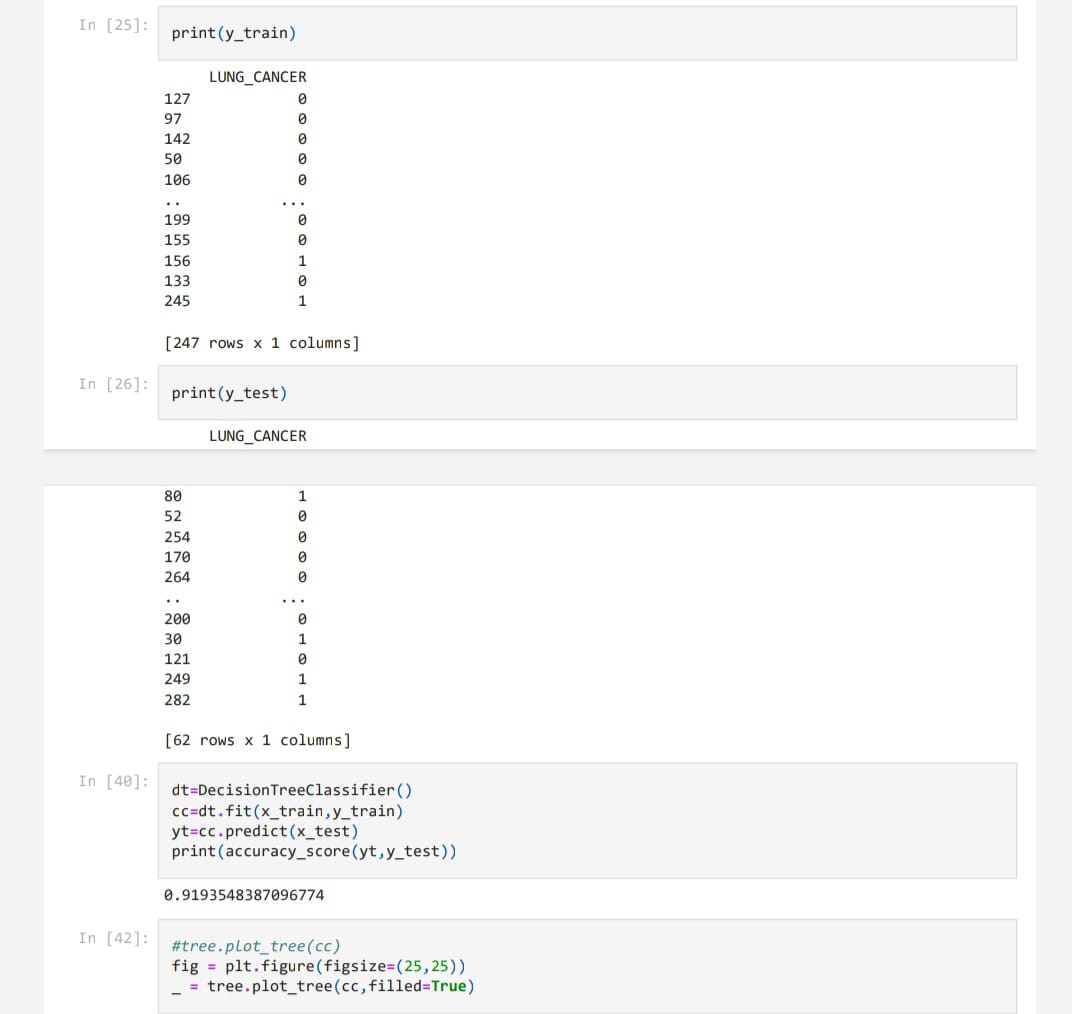
****

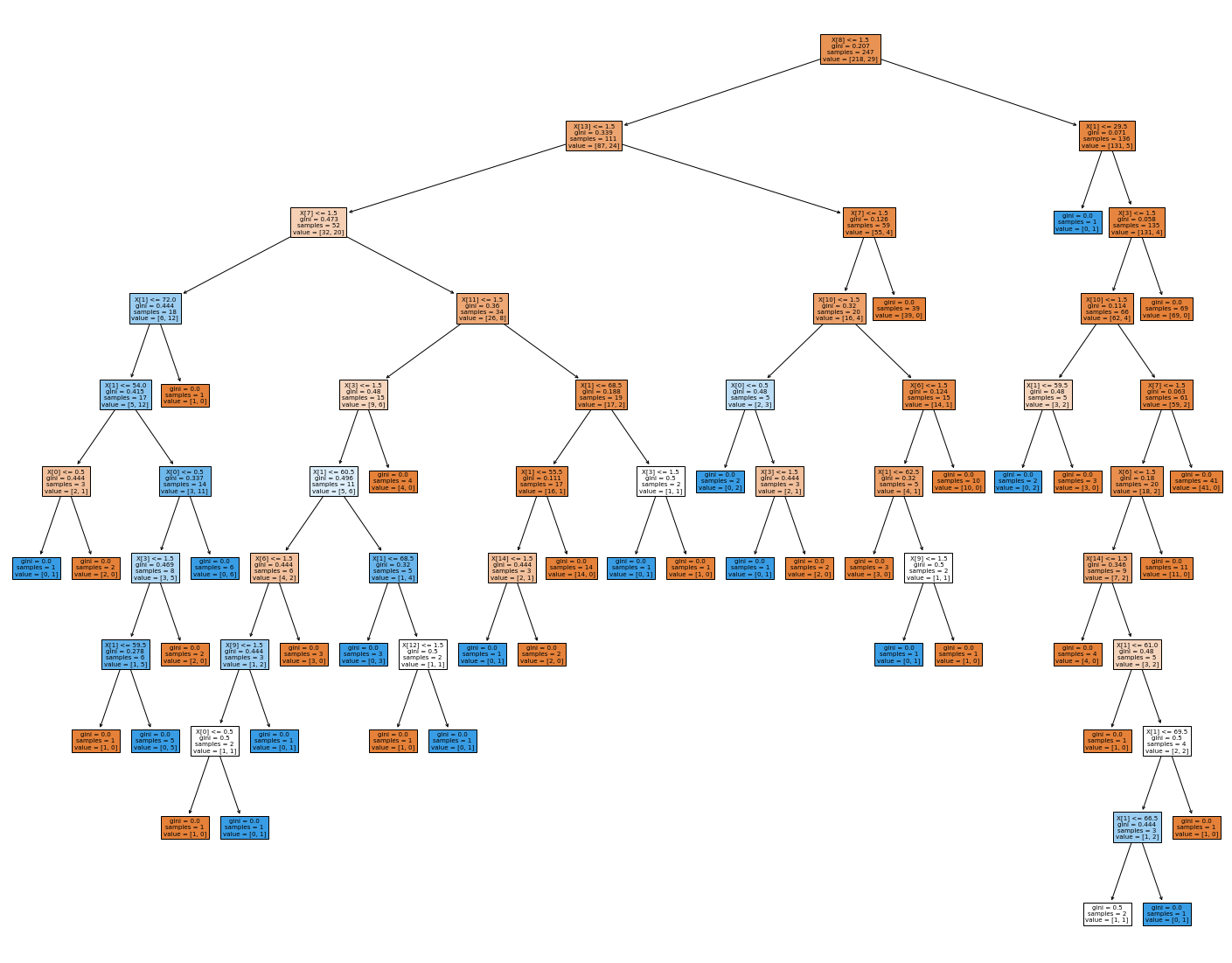
**4.2 LOGISTIC REGRESSION**

****

**4.3 DECISION TREE**

****

****

****

**4.4 SUPPORT VECTOR MACHINE**

**5.ANALYSIS AND RESULT**

In this review, we discussed the machine learning algorithm used to predict the chances of an individual getting lung cancer within the future. The algorithms that we have chosen to study are logistic regression and SVM. Based on the analysis of their result:

• The logistic regression algorithm showed an accuracy of 0.9193548387096774.

• SVM algorithm showed an accuracy of 0.9193548387096774.

i.e., the machine learning algorithm we implemented can predict the presence of lung cancer with 91% accuracy.

**6.CONCLUSION**

Lung cancer is one of the most dangerous diseases and the most common cause of death, the severity of the disease lies in the difficulty of diagnosing it in the early stages. This paper tries to endeavor to investigate of two supervised learning models to find the best classifier could classify lung cancer in early stage.

**7.REFERENCES**

Dataset: <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>

Environment: <https://research.google.com/colaboratory/>

Github: <https://github.com/dsakeeth/aiml_jcomponent/blob/main/AIMLJC_LUNG_CANCER_PREDICTION.ipynb>