Lung Cancer Prediction

University Name

Course Name

Student Name

Contents

[Introduction 3](#_Toc120809468)

[Data Collection 3](#_Toc120809469)

[Data Preprocessing 3](#_Toc120809470)

[Acquiring The Dataset 4](#_Toc120809471)

[Import Packages Required 4](#_Toc120809472)

[Import the Datasets 4](#_Toc120809473)

[Encoding the Categorical data 4](#_Toc120809474)

[Data partitioning 5](#_Toc120809475)

[Data visualization and Analysis 5](#_Toc120809476)

[Results and Conclusion 8](#_Toc120809477)

[References 9](#_Toc120809478)

[Appendix 10](#_Toc120809479)

# Introduction

With recent advances in deep learning, research has made a significant leap to help identity, classify, and quantify patterns in medical images. Notably, improvements in computer vision inspired its use in medical image analysis, such as image segmentation, image registration, image fusion, image annotation, computer-aided diagnosis and prognosis, lesion/landmark detection, and microscopic imaging analysis, to name a few. For instance, 2-D skin cancer images were successfully classified by a single CNN architecture in (Zhou & Arandian, 2021).

In this same research paper, I selected a dataset from Kaggle focused on improving lung health through lung cancer detection due to smoking. In recent months, sophisticated attempts to solve this problem have been released. Beginning with AlexNet and continuing with the model described by one of the winners of the Kaggle Data Science Bowl competition as a foundational framework, we propose a simple and elegant approach to the challenge. The effectiveness of the cancer prediction system helps people to know their cancer risk at a low cost and also allows people to make the appropriate decision based on their cancer risk status. (Joshua, Chakkravarthy, & Bhattacharyya, 2020).

# Data Collection

Dataset used is collected from the Kaggle website, a worldwide data science platform where almost all datasets for practice and challenges events on data science and Big Data take place. We found an existing repository on Kaggle with weather forecasting data already collected and which was to be used for the challenge. We downloaded the dataset into our pc to work on the pc as the working directory. The link to the dataset is provided at <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer> . The effectiveness of the cancer prediction system helps people to know their cancer risk at a low cost and allows people to make the appropriate decision based on their cancer risk status. The data is collected from the website online lung cancer prediction system (Radhika, Nair, & Veena, 2019).

**What is the AL problem you are trying to solve and why(motivation)?**

I am trying to solve an Artificial Intelligence problem under machine learning.

**Why are the methods used to solve the problem?**

1. Random Forest.
2. Decision Trees.
3. Naïve Bayes.

# Data Preprocessing

This is the first step that marks the initiation of classification-building techniques. In most cases, the real-world dataset we collect is incomplete, inconsistent, and inaccurate. This means that this dataset, in any case, must contain errors/ outliers as a result which were caused during data entry processes. Or, we may find that the dataset needs specific attr1qibute values/trends. To work on all this, data preprocessing is a step that helps the analysts bring the dataset back to completeness, consistency, and accuracy. This is achieved by performing data cleaning, formatting, organizing, and structuring the raw data collected. This is the only preparation for implementing the classification techniques. The following are the various steps implemented during data preprocessing to achieve a prepared dataset ready for classification;

# Acquiring The Dataset

Having collected data, loading the dataset into the current working environment is the first step of data preprocessing. We must acquire relevant data sets to build appropriate models using classification techniques. This data is composed of data gathered from multiple sources. These are combined in one proper format to form a dataset. In our case, data collected will only be relevant if it contains information such as humidity, rainfall, temperature, and so on for different latitudes and longitude. In simple terms, a weather forecasting dataset must contain weather-related data. Several online sources have been provided where one can download historical data to use. Apart from downloading data, one can also use Python APIs and many ways to scrape data from online platforms. However, in our case, we downloaded data from Kaggle, our data source for this project. The link to our dataset has been provided in the introduction section of this report paper. Our dataset was stored in comma-separated values (CSV) format.

# Import Packages Required

Before working on the downloaded data set, various libraries have to be imported to make it possible to load the data and start working over it. The following are the core python data preprocessing libraries that will be used in our project implementation;

Pandas – loading different file formats, data manipulation, and analysis.

Matplotlib – for 2D plotting of figures

Seaborn - alternative 2D plotting library to matplotlib.

Scikit learn - for implementation of classification techniques.

# Import the Datasets

In this step, the downloaded dataset must be imported into the working environment. Therefore, we first set the current directory as the working directory, then using panda’s library “the read\_csv()” function, import the CSV dataset files.

# Encoding the Categorical data

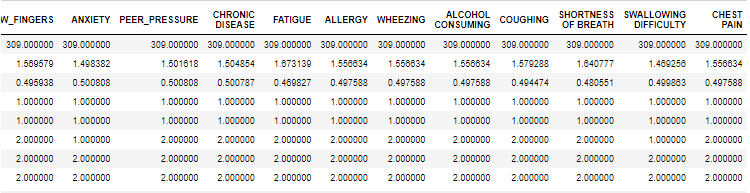
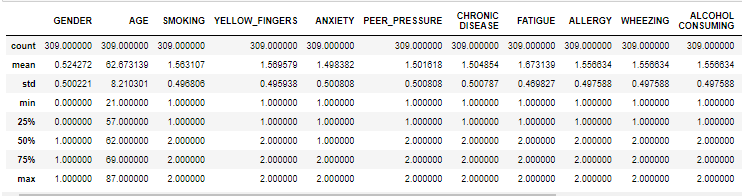
The label/categorical data variables refer to the information with specific categories. In our dataset, there is one categorical variable, the Gender, Lung cancer variables. Since the classification techniques we want to work with are machine learning models, these models are primarily based on mathematical equations. Therefore, the classification techniques can’t read and understand data in such a format. Therefore, keeping categorical data in the equation without encoding will cause specific issues; hence only numbers are needed in the equations. Consequently, we encoded our weather condition variable into numbers for our models to understand the data and work with. This is achieved by using the label Encoder() function from the sci-kit learn.

# Data partitioning

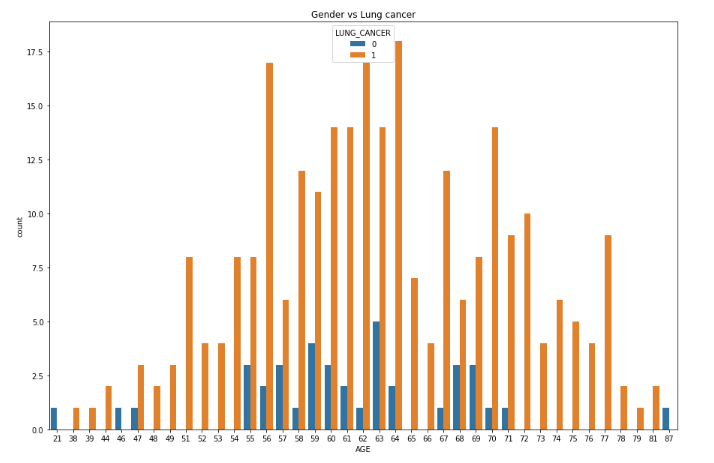
Next step is data partitioning/splitting. To enforce any technique in machine learning, every dataset must be partitioned into two sets – the training set and the test set. The training set denotes the subset of data used for training the model. Here, we are already aware of the output. On the other hand, a test subset of data is used for testing the model trained. The model uses this test set to predict outcomes. Data is usually split into 70:30 or 80:20, indicating that 70% or 80% is for training while 30% or 20% is for the testing set. In our case, the splitting done was 70:30. However, it is essential to note that the splitting process continuously varies according to the shape and size of the dataset. This is achieved using the train\_test\_split() function in the sci-kit learn library.

# Data visualization and Analysis

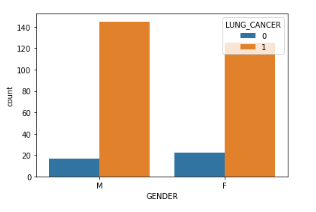
Data description table.



The table above shows the mean, maximum, minimum, quartiles, and standard deviation. Using the max, min, and mean values, we can find the outliers in the data if found in the dataset, then we work on removing them. By looking keenly, we find out that there are no outliers in our dataset, since the maximum, minimum and mean value for each feature is close.

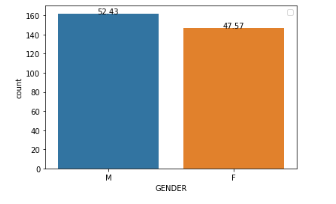


Gender and Lung cancer histogram showing the count



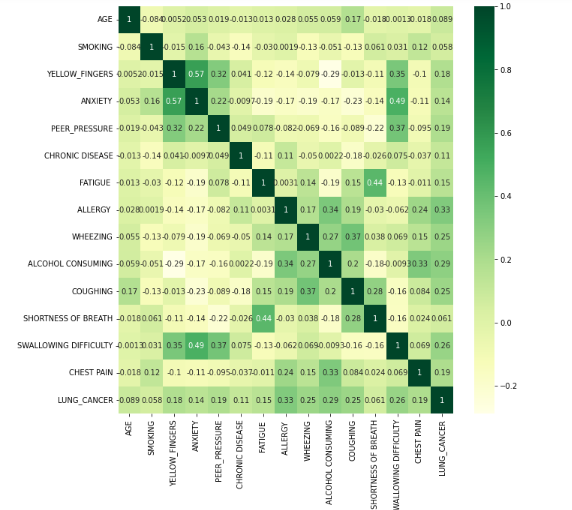
Its clear that male are the most affected gender with lung cancer disease.

Gender Histogram to show maximum count for both male and female



Male is the dominance gender in our dataset with the total count of 52.43, then female follows as the second with 47.57.

Correlation matrix for the selected features of our dataset



The above diagram shows the correlation of the data set features explained as follows. From the scatter plot above, we observe that, there is a positive high correlation between anxiety and wallowing difficulty. This means that, whenever someone gets anxious, he/she is likely to experience difficulty in wallowing. There is a positive high correlation between anxiety and yellow fingers, the same case applies between fatigue and shortness of breath. On the other hand, Our dataset has a medium to a high negative correlation between alcohol consumption and fatigue. There is a negative correlation between coughing and chronic diseases.

# Results and Conclusion

NULL HYPOTHESIS: Male is the leading gender with lung cancer diseases.

ALTERNATE HYPOTHESIS: Male is the leading gender with lung cancer diseases

NULL HYPOTHESIS: People who smoke are the leading group with lung cancer.

ALTERNATE HYPOTHESIS: People who smoke are not the leading group with lung cancer.

NULL HYPOTHESIS: People between the age of 40-50 are the leading age group with lung cancer disease.

ALTERNATE HYPOTHESIS: People between the age of 40-50 are not the leading age group with lung cancer disease.

From the previous slide, it’s evident that Random forest classifier out performs Naïve Bayes and Decision tree classifier with an accuracy of 93.5%, followed by Naïve Bayes with 91.9% and then decision tree closes with 87.1%.

This means that both three models can predict correct results for almost all the cases. But, the accuracy score of Random forest model is better than the one for naïve Bayes and decision trees which implies that the Random forest model is better than Naïve Bayes and Decisions tree.

In conclusion, We observe that, in lung cancer detection, the choice of dataset features impacts the results much more. Higher the number of features, more is the training time required to train the model. We also confirm that, smoking plays a big role cancer development.

# References

Joshua, E. S., Chakkravarthy, M., & Bhattacharyya, D. (2020). An Extensive Review on Lung Cancer Detection Using Machine Learning Techniques. *A Systematic Study. Rev. d'Intelligence Artif.*, 351-359.

Radhika, P. R., Nair, R. A., & Veena, G. (2019). A comparative study of lung cancer detection using machine learning algorithms. *In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* , 1-4.

Zhou, B., & Arandian, B. (2021). An improved CNN architecture to diagnose skin cancer in dermoscopic images based on wildebeest herd optimization algorithm. . *Computational Intelligence and Neuroscience*.

# Appendix

“Lung\_Cancer\_Prediction.py” file contains project code. The code which includes dataset loading, preprocessing, analysis, modelling and evaluation of the techniques modelled.

“presentation.pptx” file that contains the project presentation.

“survey lung cancer.csv” file that contains the collected data.