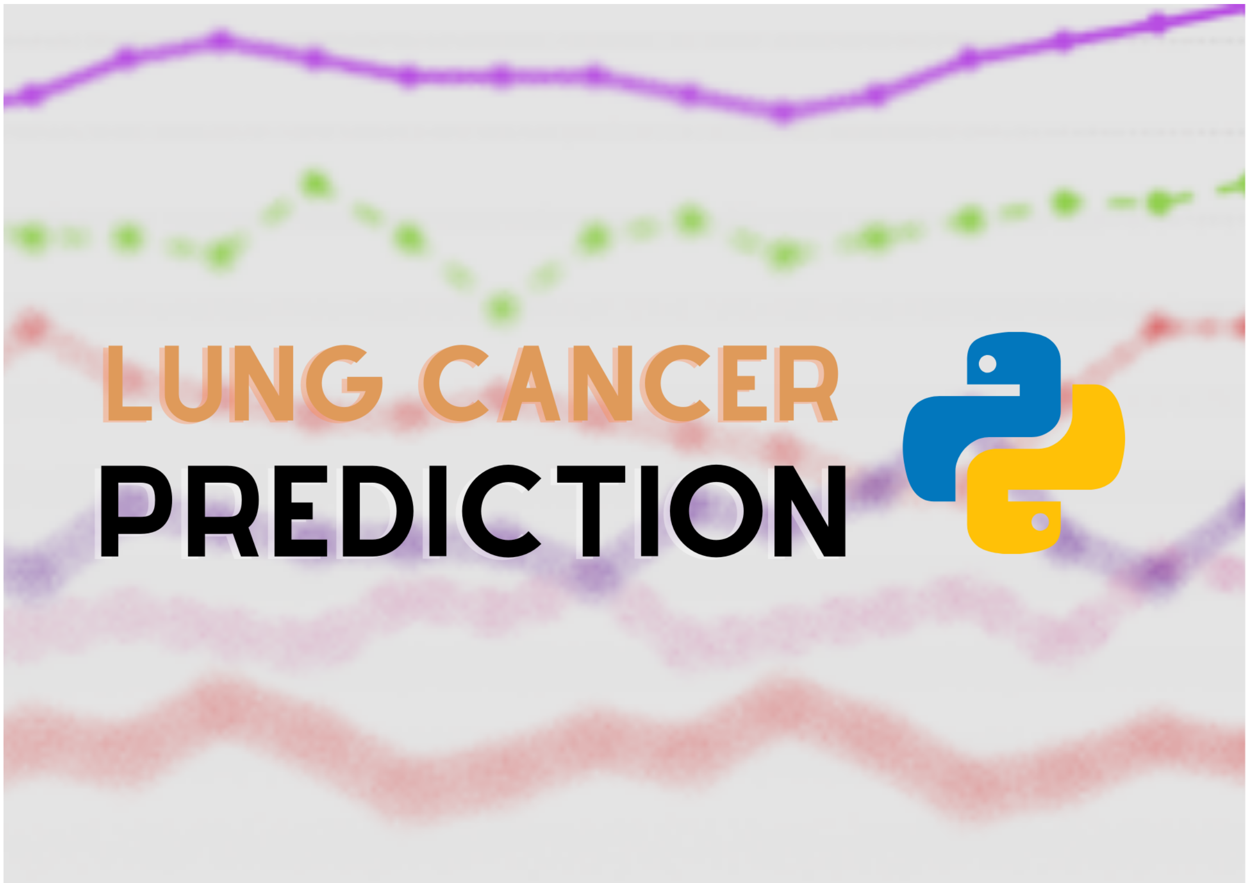
PYTHON PROJECT REPORT

IT Workshop - Final Semester Project



Alastair D'Souza	**(BT20CSE027)**
Aditi Yadav	**(BT20CSE055)**
Sahil Kothiwala	**(BT20CSE125)**
Yogesh Sewada	**(BT20CSE179)**
Prateek Niket	**(BT20CSE211)**

10.11.2021

2nd Year (IIIrd Sem) CSE

INTRODUCTION

Lung cancer is one of the leading causes of cancer deaths in both men and women. Manifestation of Lung cancer in the body of the patient reveals early symptoms in most of the cases.

Primary **prevention activities** include cigarette Smoking, diet modification, and chemoprevention. Screening is reasonably secondary prevention.

Pre-diagnosis should identify or narrow down the possibility of screening for lung cancer disease. **Symptoms** and **risk factors** (smoking, alcohol consumption, obesity and insulin resistance) had a significant effect in the pre-diagnosis stage.

OBJECTIVE: Analysing the dataset and predicting the most prominent causes/symptoms of Lung cancer in patients based on the dataset used.

PYTHON LIBRARIES USED

- **NumPy**

  NumPy is a Python library used for working with arrays. It also has functions for working in the domain of linear algebra, Fourier transform, and matrices.

- **Pandas**

  Pandas stand for "Python Data Analysis Library ". It is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data.

- **SciKit Learn**

  Scikit-learn is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

  Support vector machines (**SVMs**) are a set of supervised learning methods used for classification, regression and outlier detection.

- **Matplotlib**

  Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

  - **Seaborn**
    - Seaborn is a Python data visualization library based on matplotlib.
    - It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily.
  - **Pyplot**
    - Pyplot is a plotting library used for 2D graphics in a python programming language. It can be used in python scripts, shell, web application servers and other graphical user interface toolkits.
    - Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

FUNCTIONS USED

- **Plotter Function**

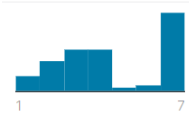
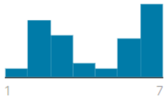
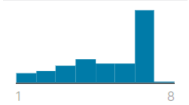
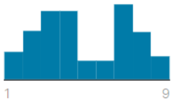
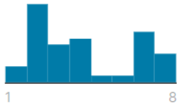
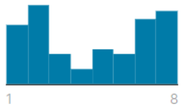
  Helper function to make a quick consistent plot with few easy changes for aesthetics.

  Input:

  - **plot**: sns or matplot plotting function
  - **x_label**: x_label as string
  - **y_label**: y_label as a string
  - **x_rot**: x-tick rotation, default=None, can be int 0-360
  - **y_rot**: y-tick rotation, default=None, can be int 0-360
  - **fontsize**: size of plot font on-axis, default=12, can be int/float
  - **fontweight**: Adding character to font, default=None, can be 'bold'
  - **legend**: Choice of including legend, default=True, bool
  - **save**: Saves image output, default=False, bool
  - **save_name**: Name of output image file as .png. Requires Save to be True. default=None, string: 'Insert Name.png'

- ○ **Output**: A customized plot based on given parameters and an output file

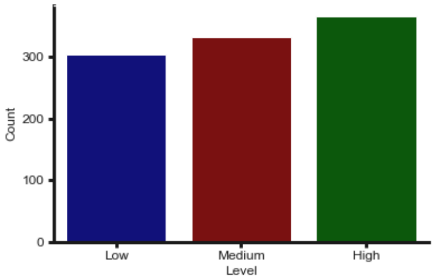
DATASET USED

OBESITY Whether or not the patient is obese	BALANCED DIET A balanced diet of the patient	DUST ALLERGY Severeness of Patient's dust allergy	COUGHING of BLOOD If the patient coughs blood	PASSIVE SMOKER The patient's smoking habits continued	ALCOHOL USE Alcohol use of Patient
					

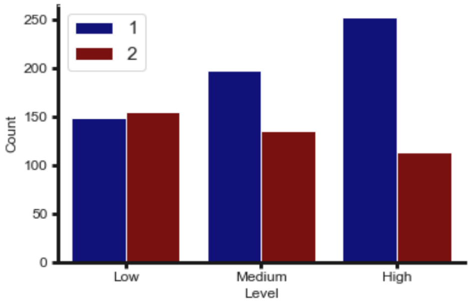
The dataset which we have used has data of patients on over 25 different parameters and symptoms. The above few mentioned are the most prominent ones. The rest can be viewed in the code.

PLOTS GENERATED

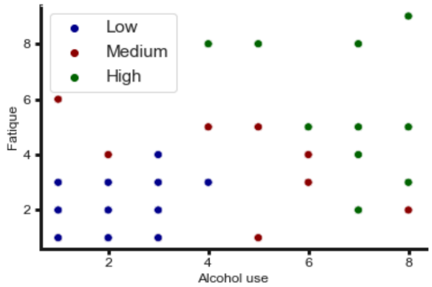
- BAR Graph



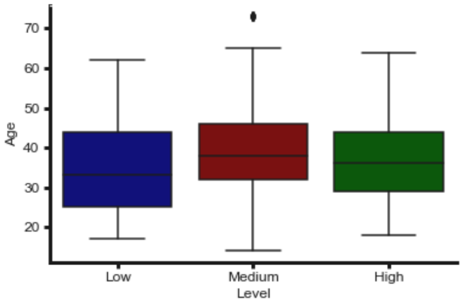
- COLUMN Graph



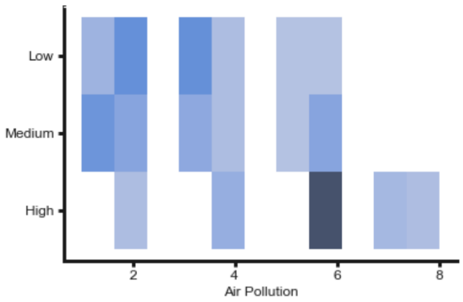
- SCATTER Plot



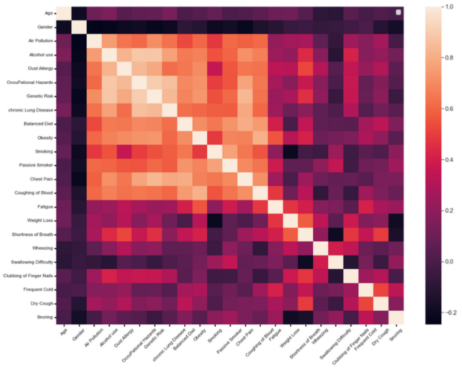
- BOX Plot



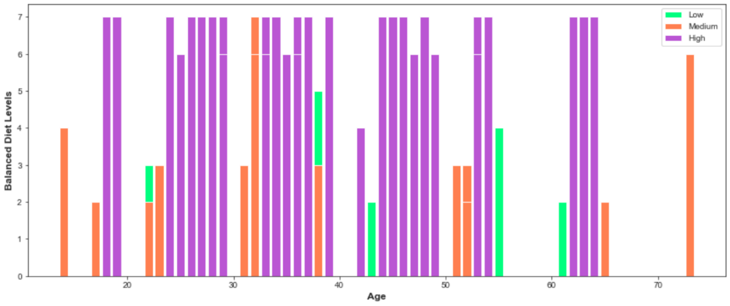
- HISTOGRAM Plot



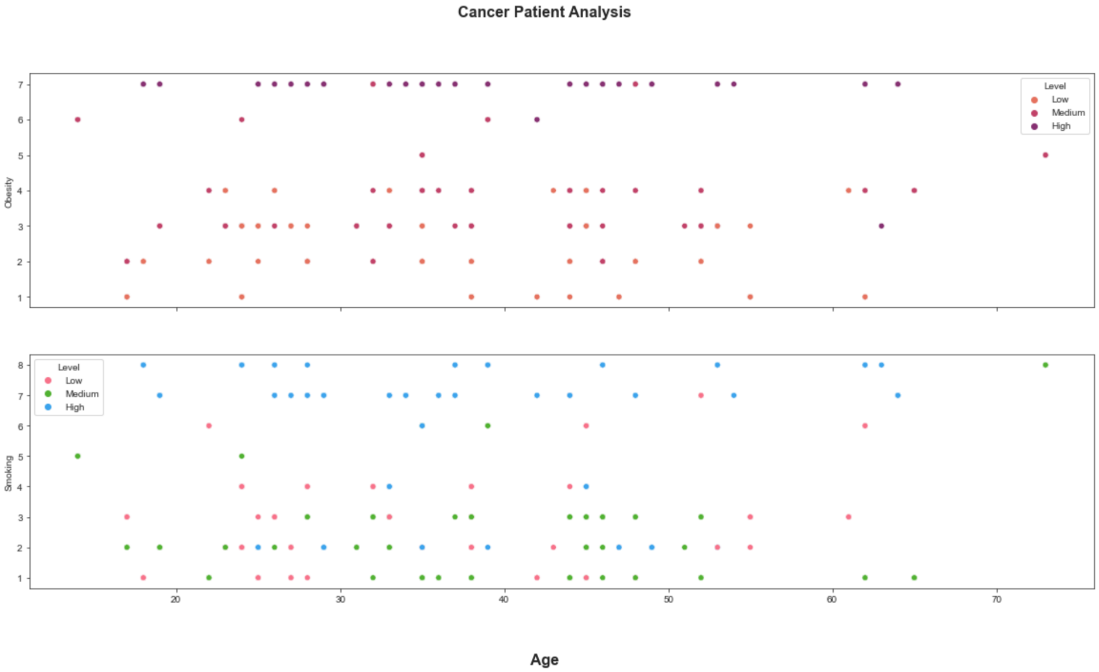
- HEATMAP Plot



- BAR Graph



- FINAL ANALYSIS Plot



RESULTS

Using the created model we could predict the most prominent causes/symptoms of Lung cancer in patients based on the dataset used.

On examining different symptoms & possible causes of Lung cancer based on our dataset we plotted multiple graphs and plots (bar graph, scatter plot, count plot,etc.) to draw various conclusions.

CONCLUSION

This predictive model will help health professionals/individuals know about the possible causes and early symptoms of Lung Cancer.

The 25 selected features in our dataset provided 99.5% accuracy when modeled on support vector machine(SVM) classifier.