Lung Cancer Susceptibility

Creating a lung cancer risk predictor.

Selected Topic & Reason for Topic

Topic:

- Susceptibility to lung cancer based on lifestyle and demographic parameters such as age, gender, alcohol use, genetic risk, and smoking.

Reason for Topic:

- Being able to define a patient's risk level of developing lung cancer can help encourage lifestyle changes to reduce risk.
- Early detection is key to survival. Detected in its earliest stages it is most treatable, with a cure rate as high as 80-90%.

Description of Data Source

This dataset was sources from Kaggle, a community form of datasets. It shows the demographic and lifestyle data of 1000 lung cancer patients.

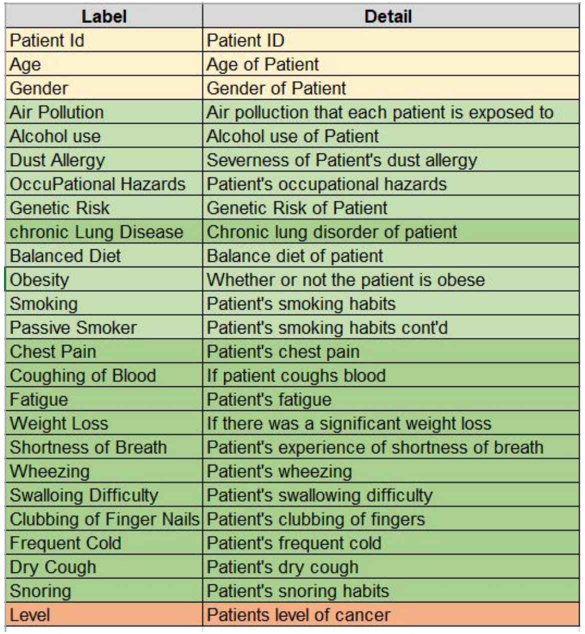
Questions to Answer

- Which combination of a patient's lifestyle would make them most susceptible to lung cancer?
- Which combination of symptoms would indicate the level of a patient’s cancer?

Description of Data Exploration Phase

- After defining the questions and topic, we searched for relevant datasets on Google Dataset Search and Kaggle.
- We were specifically looking for data on cancer patients and their lifestyle choices leading up to them contracting cancer.
- We also searched for data on cancer patient demographics.

Data Exploration Analysis



Label	Detail
Patient Id	Patient ID
Age	Age of Patient
Gender	Gender of Patient
Air Pollution	Air pollution that each patient is exposed to
Alcohol use	Alcohol use of Patient
Dust Allergy	Severness of Patient's dust allergy
OccuPational Hazards	Patient's occupational hazards
Genetic Risk	Genetic Risk of Patient
chronic Lung Disease	Chronic lung disorder of patient
Balanced Diet	Balance diet of patient
Obesity	Whether or not the patient is obese
Smoking	Patient's smoking habits
Passive Smoker	Patient's smoking habits cont'd
Chest Pain	Patient's chest pain
Coughing of Blood	If patient coughs blood
Fatigue	Patient's fatigue
Weight Loss	If there was a significant weight loss
Shortness of Breath	Patient's experience of shortness of breath
Wheezing	Patient's wheezing
Swalloing Difficulty	Patient's swallowing difficulty
Clubbing of Finger Nails	Patient's clubbing of fingers
Frequent Cold	Patient's frequent cold
Dry Cough	Patient's dry cough
Snoring	Patient's snoring habits
Level	Patients level of cancer

Patient Id = P1 - P999

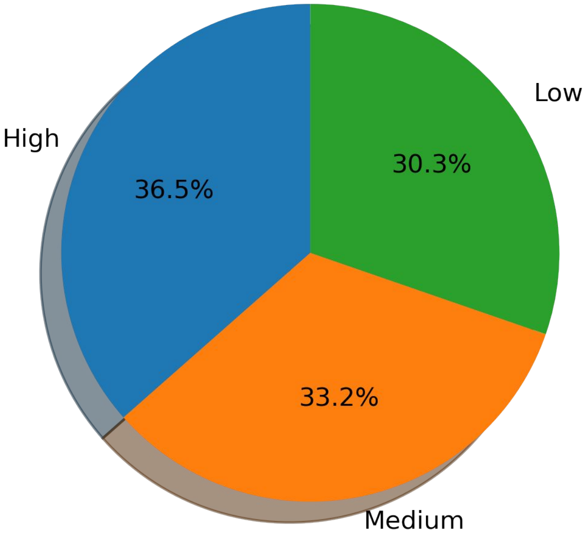
Age = range is 14 - 73

Gender = Male and Female

21 Risk Characteristics = ranked from 1 - 10

Level = “Low”, “Medium”, “High”

Level

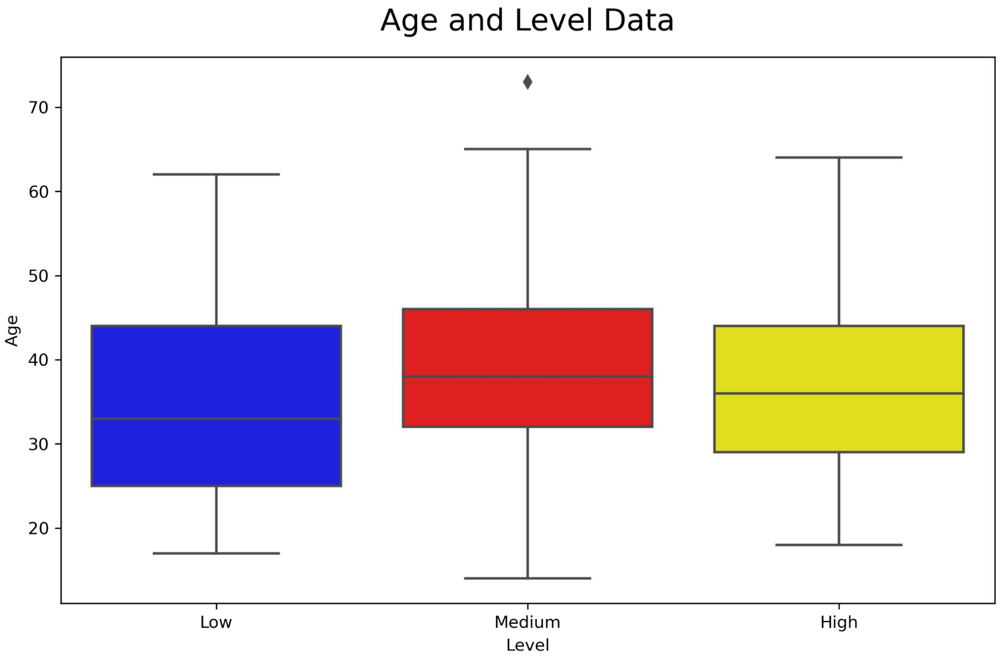


Low = 303 patients

Medium = 332 patients

High = 365 patients

Age



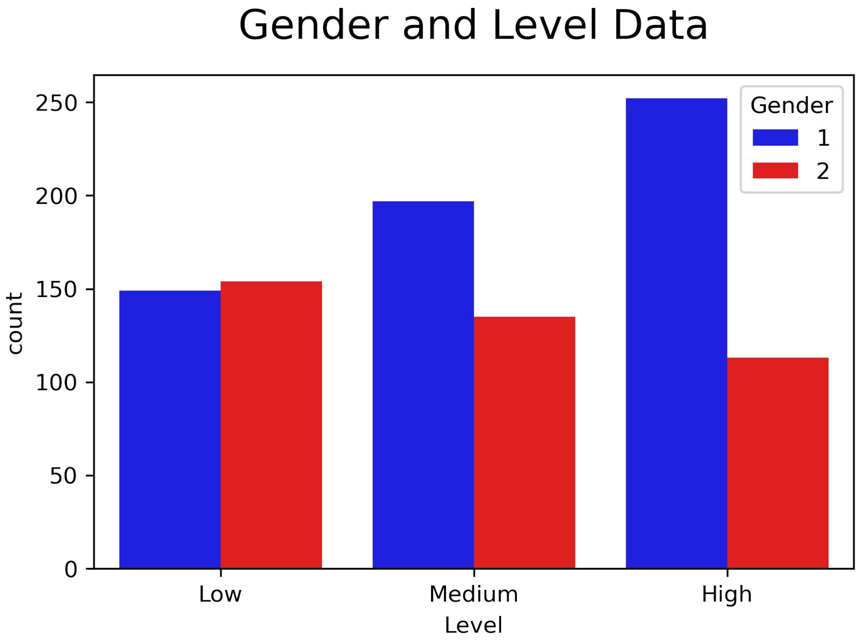
Minimum Age = 14

Maximum Age= 73

Mean = 37

- 134 patients are over the age of 50

Gender



1 = Male and **2** = Female

The data set consisted of:

- 598 male patients
- 402 female patients

Data Visualisation, Cleaning and Preprocessing

- Visualisation using Tableau
- Decide to use all the data
- Re-organise the data using SQL

Technologies, Languages, Tools, Algorithms Used

- Python
- SQL
- HTML
- Tableau
- Supervised Machine Learning
- Support Vector Classifier
- Logistic Regression
- Decision Tree Classifier

Description of Data Analysis Phase

- The dataset was prepared for the machine learning model by cleaning, removing unnecessary information, and converting all data to numerical values.
- Statistical information was extracted using python and Excel.
- The dataset was visualized using python and Tableau to find and display trends.
- The dataset was put through a supervised machine learning algorithm to create a predictive model.
- Applied Grid Search to model.

Results of Analysis

- Chest pain and coughing of blood were the most prevalent symptoms among patients with high level lung cancer.
- A combination of alcohol abuse, bad diet, occupational hazards, and smoking were the most prevalent lifestyle attributes of patients with high level lung cancer.
- Created a model that can be used to accurately predict cancer susceptibility based on a patient's lifestyle and symptom information.

Recommendations for Future Analysis

- Limited dataset. As more data is introduced, the model might have to be reevaluated.
- Change scale to 0-5 instead of 0-10.

Anything The Team Would Have Done Differently

- Try to make age and gender more relevant factors to make our model more useful.