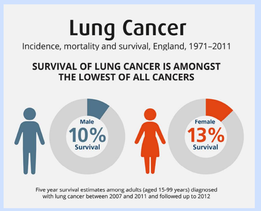
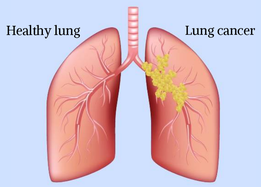
Problem Statement

Lung cancer is the leading cause of cancer-related **deaths worldwide**, accounting for the **highest mortality rates** among both men and women





Estimate **2.2 million new lung cancer cases** occurred in 2020
+ Factors why this number is still high regardless of how developed the healthcare system is:

- **Limited transportation**
- **Inaccessible facility**
- **Limited capacity**
- **Cost**

=> Our aiming for this program is to **facilitate less advance healthcare system** by **predict the level of severity of a patient's lung cancer**.

Methodology

1. Preparing the data

	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
0	P1	33	1	2	4	5	4	3	2	2	...	3	4	2	2	3	1	2	3	4	Low
1	P10	17	1	3	1	5	3	4	2	2	...	1	3	7	8	6	2	1	7	2	Medium
2	P100	35	1	4	5	6	5	5	4	6	...	8	7	9	2	1	4	6	7	2	High
3	P1000	37	1	7	7	7	7	6	7	7	...	4	2	3	1	4	5	6	7	5	High
4	P101	46	1	6	8	7	7	7	6	7	...	3	2	4	1	4	2	4	2	3	High
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	P995	44	1	6	7	7	7	7	6	7	...	5	3	2	7	8	2	4	5	3	High
996	P996	37	2	6	8	7	7	7	6	7	...	9	6	5	7	2	4	3	1	4	High
997	P997	25	2	4	5	6	5	5	4	6	...	8	7	9	2	1	4	6	7	2	High
998	P998	18	2	6	8	7	7	7	6	7	...	3	2	4	1	4	2	4	2	3	High
999	P999	47	1	6	5	6	5	5	4	6	...	8	7	9	2	1	4	6	7	2	High

The data consists of the symptoms and habits of the patient

- 22 rows of the 1000 patients' symptoms
- 3 rows of the patient ID, gender, and the level of their cancer

(ranging from low, medium to high)

Firstly, we have to start change the level of the patient which is a string value into and integers. **"Low"** will be interpreted as **zero, "Medium"** will be interpreted as **one**, and **"High"** will be interpreted as **two.**

"Level" consists of:

- Low: 303
- Medium: 332
- High: 365

=> Thus there's no need for data resampling.

Level
Low
Medium
High

Level
0
1
2

2. Creating Model

We'll be using two method, ***logistic regression*** and ***k-mean clustering*** which we will then compare the accuracy of the data and choose the best one

a. Logistic regression

+ **Creating a model** by inserting **the gender, age and symptoms as independent variables,** the **levels as the observed data.**
+ Using the model from ***sklearn***.
+ Start from **dissecting the data into 10 groups 8 of them**, X_train and y_train will be used to train the model and the other 2, X_test and y_test will be used to test the accuracy of the model.
+ The **model is trained 10 times**, with different random state to ensure the best accuracy.

***Outcome:**

A **confusion matrix** consists of the amount of predicted value **by the model and the type** of the data being predicted, precision correct, prediction of the model over total prediction from the model in percentage and recall the value predicted correctly over the total value predicted.

```
Classification report:
              precision    recall  f1-score   support

           0       0.79      0.93      0.86        67
           1       0.89      0.69      0.78        58
           2       0.97      1.00      0.99        75

    accuracy                           0.89       200
   macro avg       0.89      0.87      0.87       200
weighted avg       0.89      0.89      0.88       200

Confusion matrix:
 [[62  5  0]
 [16 40  2]
 [ 0  0 75]]
Random State 0
 Accuracy = 0.89
```

```
Random State 0:
 Accuracy = 0.99
Random State 1:
 Accuracy = 0.98
Random State 2:
 Accuracy = 0.99
Random State 3:
 Accuracy = 0.99
Random State 4:
 Accuracy = 0.99
Random State 5:
 Accuracy = 0.98
Random State 6:
 Accuracy = 0.97
Random State 7:
 Accuracy = 0.98
Random State 8:
 Accuracy = 0.99
Random State 9:
 Accuracy = 0.99
the highest accuracy obtainable is 0.99
```

*the accuracies from the ten different trainings and test sets

In addition we also calculate the **T-values** and **P-values** of the independents variable which quantify the difference between the population means, thus can confirm the validity of **null hypothesis.**

	Column	P-value
0	Gender	9.076422e-25
17	Swallowing Difficulty	1.379085e-198
18	Clubbing of Finger Nails	1.635227e-205
21	Snoring	2.297729e-211
9	Smoking	3.194101e-221
14	Weight Loss	5.536682e-228
16	Wheezing	4.095510e-228
20	Dry Cough	4.052102e-253
19	Frequent Cold	1.978174e-259
13	Fatigue	1.811820e-260
15	Shortness of Breath	2.216475e-274
2	Alcohol use	5.030563e-289
1	Air Pollution	1.955552e-296
10	Passive Smoker	5.287787e-299

	Column	T-value
3	Dust Allergy	85.877541
4	OccuPational Hazards	71.480384
8	Obesity	70.759239
6	chronic Lung Disease	70.056787
5	Genetic Risk	67.007306
7	Balanced Diet	65.202314
12	Coughing of Blood	64.520602
11	Chest Pain	57.359846
10	Passive Smoker	54.076207
1	Air Pollution	53.647767
2	Alcohol use	52.420639
15	Shortness of Breath	50.034215
13	Fatigue	47.811502
19	Frequent Cold	47.647283
20	Dry Cough	46.653602

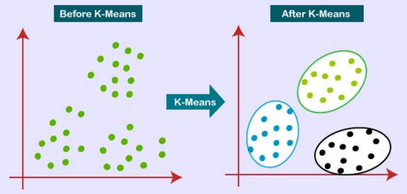
+ P-value would tell you whether the hypothesis is likely true
+ T-value would tell would tell whether it's likely false, both values are inversely proportional

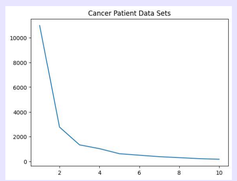
=> We can deduce that **dust allergies** symptom is **the most** relate to the level of cancer, and **gender** is **least** related to the level of cancer.

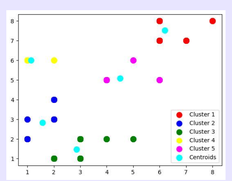
b. K-means clustering

The K-means clustering algorithm tries to minimise the distance of the points in a cluster with their centroid.

The class K-means is imported from ***sklearn***.cluster. Within Cluster Sum of Squares was used to obtain number of clusters. The "Elbow Method" was also used.







Reflection

AI application

From the use of the AI we're able to determine the severity of the condition of the patients. Hence this would significantly help the doctors or clinic to be able to prioritise the aids needed for each patients in a shorter period of time. If this method can be used with other type of disease, it will surely improve the medical industry.

Using local data

One factor that hasn't been mainly considered and highlighted in the data frame is the environmental factor of the patients.

Using a data collected from Singapore citizen would significantly improve the compatibility of the model with the targeted communities' data

Overfitting

Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead.

Hence, further improve in accuracy of the model may cause the model to performs very well for training data but has poor performance with test data (new data).

Citation:

Data set:
www.kaggle.com%2Fdatasets%2Frishidamarla%2Fcancer-patients-data%2Fcode

Nawawat Klinngam
Mohgiendren Tiruvarasu
Nguyen Ngoc Khanh An