Dataset Exploration Project Part 1 (DE pt1) – Dataset Description & FINER Questions

Tanbir Singh Sodhi

Student id: 200532155

Georgian College

22F Math for Data Analytics - 02 BDAT1005-22F-10903

Prof. Jonathan Gladstone

About the Dataset / Dataset Description:

The dataset is about the usefulness of lung-cancer predication system in U.S. that helps individuals to understand their current health conditions and warns them about the possible risks that may lead to lung cancer while spending low cost and aids them to make appropriate decisions based on their results. In other words, this dataset tells the possibilities of a person getting lung disease by knowing that the factors are absent or present. The dataset has been taken from an online website for lung-cancer prediction system. This dataset contains 21 columns(variables) with 1210 rows(records). The data dictionary defines all the variables/factors that are used in the dataset.

Data-Dictionary:

Variables / COLUMNS	DESCRIPTION	Range	Limitations
GENDER	Either of the two sexes (male and female).	Male = M, Female = F	
AGE	Age of the patient in years.	21 - 87	Excluded people below 21 and above 87 years of age.
Race	Skin Color of the patient.	white or black	The Survey was Limited to two races only.
BMI	Body Mass Index of the patient.	Below 18.5 = underweight, 18.5 – 24.9 = healthy weight, 25.0 – 29.9 = overweight, 30.0 and Above = obesity.	
SMOKING	Did the Patient smokes cigarettes regularly.	yes or no.	Does not tell the number of cigarettes the patient took regularly and also the interval.
YELLOW_FINGEERS	Did the patient has yellow fingers or not.	yes or no.	

ANXIETY	Did the patient has anxiety issues or not.	yes or no.	
CHRONIC_DISEASE	Did the patient has long term/chronic disease.	yes or no.	type of disease is not known.
FATIGUE	Did the patient feels tiredness.	yes or no.	
ALLERGY	the patient is allergic to anything.	yes or no.	Type of allergy is not known.
ALCOHOL_CONSUMING	Did the patient consumes alcohol regularly.	yes or no.	Amount of alcohol consumption is unknown.
COUGHING	Did the patient has couching symptoms.	yes or no.	
SHORTNESS_OF_BREATH	Did the patient feels shortness of breath symptoms.	yes or no.	
SWALLOWING_DIFFICULTY	Did the patient feels pain/difficulty in swallowing anything.	yes or no.	
CHEST_PAIN	Did the patient feels chest pain.	yes or no.	
PhysicalActivity	Did the person does any physical activity in the past.	yes or no.	
GenHealth	general health of the patient before illness.	extremely poor, poor, fair, good, excellent	
SleepTime	number of hours the patient sleep in a day.	1 – 18	
Asthma	Did the patient has asthma.	yes or no.	
Obesity_levels	the patient's obesity level (in percentage).	3% - 67%	
LUNG_CANCER	Did the patient has lung cancer.	yes or no.	

Research questions:

After briefly analyzing the dataset and it's variable, The following questions might be answered using the dataset:

Q1. Does smoking cause lung cancer?

Q2. Does coughing and chest pain leads to asthma?

Q3. Does allergy have an effect on the patients that causes yellow fingers?

Q4. Does anxiety lead to a smaller number of sleep time?

Q5. Does alcohol consumption lead to coughing and chest pain?

Q6. Does smoking, alcohol consumption and asthma cause lung cancer?

Description of how these questions were developed:

Q1. Does smoking cause lung cancer?

In this question, I wanted to check the fact that smoking does causes cancer with an actual dataset that represent the same output.

Q2. Does coughing and chest pain leads to asthma?

In this question, I wanted to test the hypothesis that whether chest pain and coughing lead to asthma by visualizing the dataset.

Q3. Does allergy have an influence on the patients that causes yellow fingers?

In this question, the hypothesis is also checked if allergy has an effect on patients which causes their fingers to turn yellowish or some other factors are working behind the scenes for which I may need more information or data along the way.

Q4. Does anxiety lead to a smaller number of sleep time?

In this question, I wanted to test the assumption that the patient suffering from anxiety disorder has an influence on their sleep cycle or not. So, by plotting some visuals and chats, it would be cleared if that's the case or not.

Q5. Does alcohol consumption lead to coughing and chest pain?

This question is derived from a previous question I wanted to know the answer off (does coughing and chest pain leads to asthma?). My hypothesis is that they relate to each other and to what extend-it will be cleared further in the project.

Q6. Does smoking, alcohol consumption and asthma cause lung cancer?

In this question, the general observation of the dataset is deduced to three possible leading factors which causes lung cancer. Further It will become clearer if smoking, regular alcohol consumption and asthma develop into lung cancer or not.

References:

The dataset has been taken from the following website:

Lung Cancer | Kaggle