1 survey_lung_cancer.csv

https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer/data

1.1 Description

The effectiveness of cancer prediction system helps the people to know their cancer risk with low cost and it also helps the people to take the appropriate decision based on their cancer risk status. The data is collected from the website online lung cancer prediction system .

1.2 Features

#	feature	description
1	Gender	M(male), F(female)
2	Age	Age of the patient
3	Smoking	YES=2 , NO=1.
4	Yellow fingers	YES=2 , NO=1.
5	Anxiety	YES=2 , NO=1.
6	Peer_pressure	YES=2 , NO=1.
7	Chronic Disease	YES=2 , NO=1.
8	Fatigue	YES=2 , NO=1.
9	Allergy	YES=2 , NO=1.
10	Wheezing	YES=2 , NO=1.
11	Alcohol	YES=2 , NO=1.
12	Coughing	YES=2 , NO=1.
13	Shortness of Breath	YES=2 , NO=1.
14	Swallowing Difficulty	YES=2 , NO=1.
15	Chest pain	YES=2 , NO=1.
16	Lung Cancer	YES , NO.

2 cancer_patient_data_sets.csv

ref: https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-r

2.1 Description

Lung Cancer Prediction Air Pollution, Alcohol, Smoking & Risk of Lung Cancer

2.1.1 About this dataset

This dataset contains information on patients with lung cancer, including their age, gender, air pollution exposure, alcohol use, dust allergy, occupational hazards, genetic risk, chronic lung disease, balanced diet, obesity, smoking, passive smoker, chest pain, coughing of blood, fatigue, weight loss ,shortness of breath ,wheezing ,swallowing difficulty ,clubbing of finger nails and snoring.

2.1.2 How to use the dataset

Lung cancer is the leading cause of cancer death worldwide, accounting for 1.59 million deaths in 2018. The majority of lung cancer cases are attributed to smoking, but exposure to air pollution is also a risk factor. A new study has found that air pollution may be linked to an increased risk of lung cancer, even in nonsmokers.

The study, which was published in the journal Nature Medicine, looked at data from over 462,000 people in China who were followed for an average of six years. The participants were divided into two groups: those who lived in areas with high levels of air pollution and those who lived in areas with low levels of air pollution.

The researchers found that the people in the high-pollution group were more likely to develop lung cancer than those in the low-pollution group. They also found that the risk was higher in nonsmokers than smokers, and that the risk increased with age.

While this study does not prove that air pollution causes lung cancer, it does suggest that there may be a link between the two. More research is needed to confirm these findings and to determine what effect different types and levels of air pollution may have on lung cancer risk

2.1.3 Research Ideas

predicting the likelihood of a patient developing lung cancer identifying risk factors for lung cancer determining the most effective treatment for a patient with lung cancer

2.2 Features

Column name	Description
Age	The age of the patient. (Numeric)
Gender	The gender of the patent. (Categorical)
Air Pollution	The level of air pollution exposure of the patient. (Categorical)
Alcohol use	The level of alcohol use of the patient. (Categorical)
Dust Allergy	The level of dust allergy of the patient. (Categorical)
OccuPational Hazards	The level of occupational hazards of the patient. (Categorical)
Genetic Risk	The level of genetic risk of the patient. (Categorical)
chronic Lung Disease	The level of chronic lung disease of the patient. (Categorical)
Balanced Diet	The level of balanced diet of the patient. (Categorical)
Obesity	The level of obesity of the patient. (Categorical)
Smoking	The level of smoking of the patient. (Categorical)
Passive Smoker	The level of passive smoker of the patient. (Categorical)
Chest Pain	The level of chest pain of the patient. (Categorical)
Coughing of Blood	The level of coughing of blood of the patient. (Categorical)
Fatigue	The level of fatigue of the patient. (Categorical)
Weight Loss	The level of weight loss of the patient. (Categorical)
Shortness of Breath	The level of shortness of breath of the patient. (Categorical)
Wheezing	The level of wheezing of the patient. (Categorical)
Swallowing Difficulty	The level of swallowing difficulty of the patient. (Categorical)
Clubbing of Finger Nails	The level of clubbing of finger nails of the patient. (Categorical)

3 covid_dataset.csv

ref: `https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker`

3.1 Description

These data will help to identify whether any person is having a coronavirus disease or not based on some pre-defined standard symptoms. These symptoms are based on guidelines given by the World Health Organization (WHO) and the Ministry of Health and Family Welfare, India.

Disclaimer: The results or analysis of these data should be taken as medical advice.

The dataset contains seven major variables that will be having an impact on whether someone has coronavirus disease or not, the description of each variable are as follows:

- Country: List of countries person visited.
- Age: Classification of the age group for each person, based on WHO Age
- Group Standard Symptoms: According to WHO, 5 are major symptoms of COVID-19, Fever, Tiredness, Difficulty in breathing, Dry cough, and sore throat.
- Experience any other symptoms: Pains, Nasal Congestion, Runny Nose, Diarrhea and Other.
- Severity: The level of severity, Mild, Moderate, Severe Contact: Has the person contacted some other COVID-19 Patient

3.2 Features

The normalized dataset can be described in that way:

Column Number	Column Name	Non-Null Count
0	Fever	316800
1	Tiredness	316800
2	Dry-Cough	316800
3	Difficulty-in-Breathing	316800
4	Sore-Throat	316800
5	None_Sympton	316800
6	Pains	316800
7	Nasal-Congestion	316800
8	Runny-Nose	316800
9	Diarrhea	316800
10	None_Experiencing	316800
11	Age_0-9	316800
12	Age_10-19	316800
13	Age_20-24	316800
14	Age_25-59	316800
15	Age_60+	316800
16	Gender_Female	316800
17	Gender_Male	316800
18	Gender_Transgender	316800
19	Severity_Mild	316800
20	Severity_Moderate	316800
21	Severity_None	316800
22	Severity_Severe	316800
23	Contact_Dont-Know	316800
24	Contact_No	316800
25	Contact_Yes	316800
26	Country	316800

4 dataset_integrated.csv

All datasets above integrated in a such way that a R script can described in the scripts/integrate_datasets.r can be explained.

4.1 Features

Index	Column Name	Non-Null Count
0	AGE	1309
1	SMOKING	318109
2	YELLOW_FINGERS	1309
3	ANXIETY	309
4	PEER_PRESSURE	309
5	CHRONIC_DISEASE	309
6	FATIGUE	318109
7	ALLERGY	309
8	WHEEZING	1309
9	ALCOHOL_CONSUMING	309
10	COUGHING	318109
11	SHORTNESS_OF_BREATH	318109
12	SWALLOWING_DIFFICULTY	318109
13	CHEST_PAIN	318109
14	LUNG_CANCER	318109
15	SOURCE	318109
16	AGE_0_9	318109
17	AGE_10_19	318109
18	AGE_20_24	318109
19	AGE_25_59	318109
20	AGE_60_	318109
21	GENDER_FEMALE	318109
22	GENDER_MALE	318109
23	COLD_SYMPTOMNS	318109
24	RESPIRATORY_SYMPTOMNS	318109
25	OTHER_SYMPTOMS	318109
26	SNORING	318109
27	SEVERITY	318109
28	INDEX	1000
29	PATIENT_ID	1000
30	AIR_POLLUTION	1000
31	ALCOHOL_USE	1000
32	DUST_ALLERGY	1000
33	OCCUPATIONAL_HAZARDS	1000
34	GENETIC_RISK	1000
35	CHRONIC_LUNG_DISEASE	1000
36	BALANCED_DIET	1000

Continued on next page

Continued from previous page

Index	Column Name	Non-Null Count
37	OBESITY	1000
38	PASSIVE_SMOKER	1000
39	COUGHING_OF_BLOOD	1000
40	WEIGHT_LOSS	1000
41	FREQUENT_COLD	1000
42	DRY_COUGH	1000
43	FEVER	316800
44	NONE_SYMPTON	316800
45	PAINS	316800
46	NASAL_CONGESTION	316800
47	RUNNY_NOSE	316800
48	DIARRHEA	316800
49	NONE_EXPERIENCING	316800
50	GENDER_TRANSGENDER	316800
51	SEVERITY_MILD	316800
52	SEVERITY_MODERATE	316800
53	SEVERITY_NONE	316800
54	SEVERITY_SEVERE	316800
55	CONTACT_DONT_KNOW	316800
56	CONTACT_NO	316800
57	CONTACT_YES	316800
58	COUNTRY	316800