**Bonertz, Brian**

**Project 3**

**Week 9 Project Proposal:**

**Title: Lung Cancer Data Analysis**

**Introduction**

Lung Cancer is a condition affecting millions of people globally, typically associated with smoking. Examining features associated with individuals throughout the world that may contribute to lung cancer can be useful for medical research, predictive analysis, and improved global policy. This project aims to analyze these features to determine their predictive associations with lung cancer as well as evaluate clusters or patters that may contribute to lung cancer.

**Problem Statement**

Using predictive analysis and classification modeling, can individuals be identified as high risk for lung cancer, and can groups be identified as higher risk of lung cancer through classification modeling.

**Objectives**

1. To analyze the dataset to identify key factors affecting the diagnosis of lung cancer.
2. To develop prediction models for identifying individuals at high risk of lung cancer.
3. Evaluate variables that contribute to diagnosis of lung cancer.
4. To examine classification models to determine variables that correlate with individuals with lung cancer.

**Data Description**

The dataset named “Lung Cancer Risk & Trends Across 25 Countries” was sourced from Kaggle website and consists of over 220,632 rows of data and 17 features containing both categorical and numerical data types. Data features are listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | ID | Unique identifier per individual | Numerical (INT) |
| 2 | Country | Home country of individual sampled | Categorical |
| 3 | Population Size | Number of individuals sampled in each country | Numerical (INT) |
| 4 | Age | Age of individual | Numerical (INT) |
| 5 | Gender | Gender of individual | Categorical |
| 6 | Smoker | Determination smoker or not | Categorical |
| 7 | Years\_of\_Smoking | Number of years individual smoked | Numerical (INT) |
| 8 | Cigarettes\_per\_Day | Number of cigarettes smoked per day | Numerical (INT) |
| 9 | Passive\_Smoker | Determination if individual is passive smoker or not | Categorical |
| 10 | Family\_History | Family history of lung cancer | Categorical |
| 11 | Lung\_Cancer\_Diagnosis | Diagnosis of lung cancer (Yes / No) | Categorical |
| 12 | Cancer\_Stage | Stage of cancer from stage 1 – stage 4 | Categorical |
| 13 | Air\_Pollution\_Exposure | Air pollution rating of the country (Low – High) | Categorical |
| 14 | Occupational\_Exposure | Occupational exposure contributing to lung cancer | Categorical |
| 15 | Indoor\_Pollution | Indoor pollution rating (Yes / No) | Categorical |
| 16 | Healthcare\_Access | Healthcare access for the country (Poor / Good) | Categorical |
| 17 | Annual\_Lung\_Cancer\_Deaths | Annual lung cancer deaths by country | Numerical (INT) |

**Methodology**

* Data Preprocessing: Handle missing values, encode categorical variables, and normalize numerical values.
* Exploratory Data Analysis: Identify patterns, correlations, and key features influencing diabetes.
* Model Selection: Because of their predictive abilities, I will evaluate logistic regression and classification to find the most suitable model. If their performance is not remarkable, other machine-learning algorithms will be investigated.
* Model Training and Testing: Data will be split into training and testing sets to train the model and evaluate its performance.
* Validation: Cross-validation techniques to ensure the model’s generalizability.

**Expected Outcomes**

* Insights into the key factors influencing lung cancer.
* Insights into variables that correlate or contribute to lung cancer, enabling better planning and lifestyle habits to reduce risk.

**References**

Kaggle Dataset Website: [Lung Cancer Risk & Trends Across 25 Countries](https://www.kaggle.com/datasets/ankushpanday1/lung-cancer-risk-and-trends-across-25-countries/data)

GitHub Dataset: [Bonertz\_Brian\_Portfolio/Project\_10 Diabetes Prediction at main · bonertzb/Bonertz\_Brian\_Portfolio](https://github.com/bonertzb/Bonertz_Brian_Portfolio/tree/main/Project_10%20Diabetes%20Prediction)

Note: Other datasets may be used in the analysis to help shape the project and support objectives.