Lung Cancer Prediction Report

This report outlines the workflow and results of a Lung Cancer Prediction model using Logistic Regression. The goal of this project is to predict whether a person has lung cancer based on various health-related features.

Workflow Overview 🔄

1. **Data Collection and Preprocessing** 📂
2. **Exploratory Data Analysis (EDA)** 🔍
3. **Data Standardization** 📏
4. **Splitting Data into Train and Test Sets** 🧩
5. **Model Training (Logistic Regression)** 🏋️
6. **Model Evaluation** 📈
7. **Building a Predictive System** 🤖

---

1. Data Collection and Preprocessing 📂

Dataset Overview

- The dataset used is **survey lung cancer.csv**, which contains 309 rows and 16 columns.
- Features include:
  - **GENDER**: Male (1) or Female (0)
  - **AGE**: Age of the person
  - **SMOKING**: Smoking habit (1: Yes, 2: No)
  - **YELLOW_FINGERS**: Presence of yellow fingers (1: Yes, 2: No)
  - **ANXIETY**: Anxiety level (1: Yes, 2: No)
  - **PEER_PRESSURE**: Peer pressure (1: Yes, 2: No)
  - **CHRONIC DISEASE**: Chronic disease history (1: Yes, 2: No)
  - **FATIGUE**: Fatigue level (1: Yes, 2: No)
  - **ALLERGY**: Allergy history (1: Yes, 2: No)
  - **WHEEZING**: Wheezing (1: Yes, 2: No)
  - **ALCOHOL CONSUMING**: Alcohol consumption (1: Yes, 2: No)
  - **COUGHING**: Coughing (1: Yes, 2: No)

  - **SHORTNESS OF BREATH**: Shortness of breath (1: Yes, 2: No)
  - **SWALLOWING DIFFICULTY**: Difficulty in swallowing (1: Yes, 2: No)
  - **CHEST PAIN**: Chest pain (1: Yes, 2: No)
  - **LUNG_CANCER**: Target variable (1: Yes, 0: No)

Preprocessing Steps

- The target variable **LUNG_CANCER** was converted from categorical ("YES"/"NO") to numerical (1/0).
- The **GENDER** column was also converted from categorical ("M"/"F") to numerical (1/0).
- No missing values were found in the dataset.

---

2. Exploratory Data Analysis (EDA) 🔍

Dataset Statistics

- The dataset contains **309 entries** with **16 features**.
- The mean age of the individuals is **62.67 years**.
- The dataset is imbalanced, with **270 cases of lung cancer (87.4%)** and **39 cases without lung cancer (12.6%)**.

Grouped Analysis

- The mean values of features were grouped by the target variable **LUNG_CANCER** to observe differences between individuals with and without lung cancer.
  - For example, individuals with lung cancer tend to have higher values for features like **SMOKING**, **YELLOW_FINGERS**, and **FATIGUE**.

---

3. Data Standardization 📏

- The dataset was standardized using **StandardScaler** from **sklearn.preprocessing**.
- Standardization ensures that all features have a mean of 0 and a standard deviation of 1, which is crucial for Logistic Regression.

---

4. Splitting Data into Train and Test Sets 🧩

- The dataset was split into **training (80%)** and **testing (20%)** sets using **train_test_split**.
- The split was stratified to maintain the same proportion of the target variable in both sets.

---

5. Model Training (Logistic Regression) 🏋️

- A **Logistic Regression** model was trained on the standardized training data.
- The model was trained using the **LogisticRegression** class from **sklearn.linear_model**.

---

6. Model Evaluation 📈

Training Data Accuracy

- The model achieved an accuracy of **93.52%** on the training data.

Test Data Accuracy

- The model achieved an accuracy of **91.94%** on the test data.

---

7. Building a Predictive System 🤖

Input Data

- A random input data point was selected to test the predictive system:

python

Copy

```
input_data = (0, 48, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 1)
```

  - This represents a **48-year-old female** with various symptoms.

Prediction

- The input data was standardized and passed to the trained model.
- The model predicted that the person **has lung cancer**.

Conclusion 🎯

- The Logistic Regression model performed well, achieving **91.94% accuracy** on the test data.
- The predictive system can be used to predict lung cancer based on health-related features with high accuracy.
- Future improvements could include handling the class imbalance and exploring other machine learning models.