Winter Semester 2023 – 2024

Course Name:

SOFTWARE METRICS

Course Code:

SWE-2020

Title:

LUNG CANCER PREDICTION SYSTEM

Under The Guidance Of

Prof. Uma Maheshwari G

By:

Mugunthan N – 21MIS0196
Nithish Kumar R – 21MIS0341

Group ID : 6

1) GQM (GOALS QUESTIONARIES METRICS):

GOALS:

1. Develop a Reliable Lung Cancer Prediction System to assist medical practitioner in early detection and treatment.
2. Identify Key Associations between Air Pollution and Lung Cancer Incidence
3. Investigate the Impact of Environmental Factors on Lung Cancer Risk
4. Improve Accessibility and Usability of the Lung Cancer Prediction System

QUESTIONARIES & METRICS:

for GOAL1:

1. What demographic factors are most strongly associated with lung cancer incidence?
Metrics: Age, Gender

2. Which lifestyle choices or habits have the highest correlation with lung cancer development?
Metrics: Alcohol Use, Smoking, Passive Smoker, Obesity, Balanced Diet

3. Are there any specific medical conditions that significantly increase the likelihood of developing lung cancer?
Metrics: Chronic Lung Disease, Dust Allergy, Genetic Risk

4. What are the most common symptoms associated with lung cancer?
Metrics: Chest Pain, Coughing of Blood, Fatigue, Weight Loss, Shortness of Breath, Wheezing, Swallowing Difficulty

5. How do environmental factors contribute to the risk of lung cancer?
Metrics: Air Pollution, Occupational Hazards

for GOAL2:

1. What is the correlation between air pollution levels and lung cancer incidence?
Metrics: Air Pollution, Level of Cancer

2. How does air pollution compare to other environmental factors in its impact on lung cancer risk?
Metrics: Air Pollution, Occupational Hazards, Dust Allergy

3. Are certain demographic groups more susceptible to the effects of air pollution on lung cancer incidence?

Metrics: Age, Gender

for GOAL 3:

1. How do occupational hazards contribute to the risk of lung cancer?
Metrics: Occupational Hazards, Level of Cancer

2. Is there a correlation between exposure to dust allergens and lung cancer incidence?
Metrics: Dust Allergy, Level of Cancer

3. What role does genetic predisposition play in the interaction between environmental factors and lung cancer risk?
Metrics: Genetic Risk, Level of Cancer

4.Are there any synergistic effects between different environmental factors in increasing lung cancer risk?
Metrics: Various combinations of environmental factors (e.g., Air Pollution and Smoking)

for GOAL 4:

1. How user-friendly is the current lung cancer prediction system interface?
Metrics: User interface feedback, ease of use

2. Are there any specific features or functionalities that users find particularly useful or lacking in the current system?
Metrics: User feedback on features such as data input, prediction accuracy, and interpretation of results

3. How accessible is the prediction system to medical practitioners with varying levels of technical expertise?
Metrics: Training requirements, user support needs

4. What improvements can be made to enhance the accessibility and usability of the lung cancer prediction system?
Metrics: User feedback on system improvements, implementation of suggested enhancements

2) IMPLEMENTATION DETAILS:

Tools: Google Collab
Language : Python
Libraries : Pandas, Matplotlib, Numpy

Approaches:

Data Preprocessing:

- **Handle missing values:** Check for missing values in the dataset and decide on a strategy to handle them (e.g., imputation, deletion).

Exploratory Data Analysis (EDA):

- **Understand the distribution of each feature:** Use histograms, box plots, or density plots to visualize the distribution of each feature.
- **Identify correlations:** Use correlation matrices or pair plots to identify relationships between features and the target variable (Level of Cancer).
- **Explore relationships between features:** Investigate how different features correlate with each other to identify potential interactions.

Feature Selection:

- **Select relevant features:** Use techniques like feature importance ranking, correlation analysis, or domain knowledge to select the most relevant features for predicting lung cancer.

Interpretation and Communication:

- **Interpret:** Interpret the analysis and find more insights from the analysis
- **Communicate results:** Present the findings and insights from the analysis in a clear and understandable manner, using visualizations and narratives to convey the information effectively.

3. Metrics Analysis:

Descriptive Analysis:

- **Summary statistics:** Calculate measures like mean, median, mode, standard deviation, minimum, maximum, and quartiles for numerical variables (e.g., Age, Air Pollution).

- **Frequency distribution:** Generate frequency tables or histograms to show the distribution of categorical variables (e.g., Gender, Smoking).

Correlation Analysis:

- **Pearson correlation:** Compute the Pearson correlation coefficient between numerical variables to measure the linear relationship between them.
- **Spearman correlation:** Calculate the Spearman rank correlation coefficient if variables have a nonlinear relationship or if the assumptions of Pearson correlation are not met.
- **Heatmap:** Visualize the correlation matrix using a heatmap to identify strong positive or negative correlations between variables.

Regression Analysis:

- **Simple linear regression:** Investigate the relationship between a single predictor variable (e.g., Age) and the target variable (Level of Cancer).
- **Multiple linear regression:** Build a regression model considering multiple predictor variables (e.g., Age, Smoking) to predict the target variable.
- **Logistic regression:** If the target variable is binary (e.g., lung cancer diagnosis - yes/no), use logistic regression to model the probability of lung cancer occurrence based on predictor variables.

Visualization:

- **Scatter plots:** Plot pairwise scatter plots to visualize the relationship between numerical variables, such as Age versus Level of Cancer.
- **Box plots:** Create box plots to visualize the distribution of numerical variables across different categories (e.g., Gender, Smoking).
- **Bar plots:** Generate bar plots to visualize the frequency or distribution of categorical variables, such as Gender or Smoking status.
- **Histograms:** Plot histograms to visualize the distribution of numerical variables, such as Age or Air Pollution levels.
- **Pie charts:** Create pie charts to show the distribution of categorical variables, such as Gender or Smoking status, as proportions of the whole dataset.
- **Pair plots:** Use pair plots (also known as scatterplot matrices) to visualize pairwise relationships between multiple numerical variables simultaneously.

4. CONCLUSION :

Descriptive Analysis:

- The dataset includes information on various factors such as age, gender, lifestyle habits (e.g., smoking, alcohol use), environmental exposures (e.g., air pollution, dust

allergy), and symptoms related to lung cancer (e.g., coughing of blood, shortness of breath).

- **Age distribution:** The average age of individuals in the dataset is [37]. The age range varies from [14] to [73].
- **Gender distribution:** The dataset contains data for both males and females, with [59.8] % of individuals being male and [40.2]% being female.
- Other lifestyle factors such as alcohol use, obesity, and diet quality also vary among individuals in the dataset.

Code Snippet for Descriptive Analysis:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
#reading the dataset
data = pd.read_csv('LC_dataset.csv')
#numerical summary
numerical_summary = data.describe()
# Print summary statistics
print("Summary Statistics for Numerical Variables:")
print(numerical_summary)
```

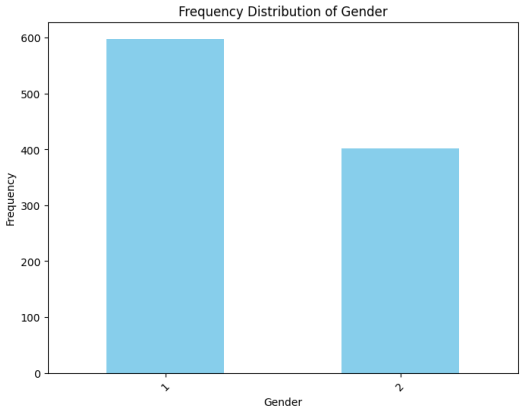
Output:

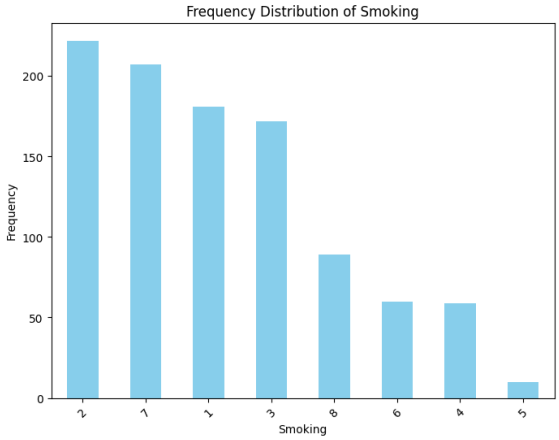
```
Summary Statistics for Numerical Variables:
             index          Age       Gender  Air Pollution  Alcohol use  \
count  1000.000000  1000.000000  1000.000000     1000.0000  1000.000000
mean    499.500000    37.174000     1.402000        3.8400     4.563000
std     288.819436    12.005493     0.490547        2.0304     2.620477
min       0.000000    14.000000     1.000000        1.0000     1.000000
25%     249.750000    27.750000     1.000000        2.0000     2.000000
50%     499.500000    36.000000     1.000000        3.0000     5.000000
75%     749.250000    45.000000     2.000000        6.0000     7.000000
max     999.000000    73.000000     2.000000        8.0000     8.000000

       Dust Allergy  OccuPational Hazards  Genetic Risk  chronic Lung Disease  \
count   1000.000000           1000.000000   1000.000000           1000.000000
mean       5.165000              4.840000      4.580000              4.380000
std        1.980833              2.107805      2.126999              1.848518
min        1.000000              1.000000      1.000000              1.000000
25%        4.000000              3.000000      2.000000              3.000000
50%        6.000000              5.000000      5.000000              4.000000
75%        7.000000              7.000000      7.000000              6.000000
max        8.000000              8.000000      7.000000              7.000000

       Balanced Diet  ...  Coughing of Blood      Fatigue  Weight Loss  \
count    1000.000000  ...        1000.000000  1000.000000  1000.000000
mean        4.491000  ...           4.859000     3.856000     3.855000
std         2.135528  ...           2.427965     2.244616     2.206546
min         1.000000  ...           1.000000     1.000000     1.000000
25%         2.000000  ...           3.000000     2.000000     2.000000
50%         4.000000  ...           4.000000     3.000000     3.000000
75%         7.000000  ...           7.000000     5.000000     6.000000
max         7.000000  ...           9.000000     9.000000     8.000000

       Shortness of Breath     Wheezing  Swallowing Difficulty  \
count          1000.000000  1000.000000            1000.000000
mean              4.240000     3.777000               3.746000
std               2.285087     2.041921               2.270383
min               1.000000     1.000000               1.000000
25%               2.000000     2.000000               2.000000
50%               4.000000     4.000000               4.000000
75%               6.000000     5.000000               5.000000
max               9.000000     8.000000               8.000000
```

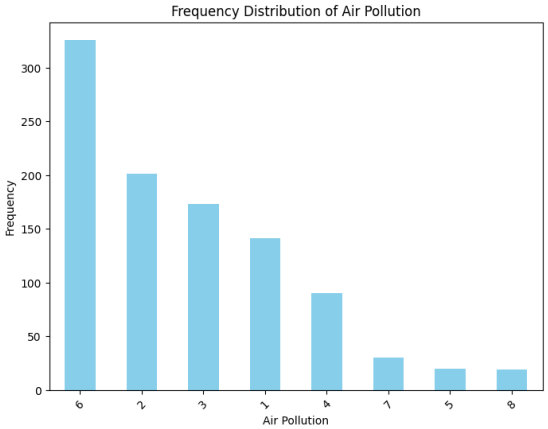
Code Snippet (For Frequency Distribution Analysis):

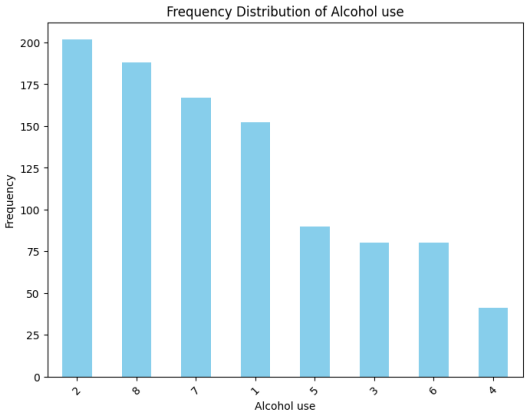
```
# Frequency distribution for categorical variables
categorical_variables = ['Gender', 'Smoking','Air Pollution','Alcohol
use']  # Add more variables as needed
for var in categorical_variables:
    freq_table = data[var].value_counts()
    print("\nFrequency Distribution for", var, ":")
    print(freq_table)

    # Plot histogram
    plt.figure(figsize=(8, 6))
    data[var].value_counts().plot(kind='bar', color='skyblue')
    plt.title('Frequency Distribution of ' + var)
    plt.xlabel(var)
    plt.ylabel('Frequency')
    plt.xticks(rotation=45)
    plt.show()
```

Output:







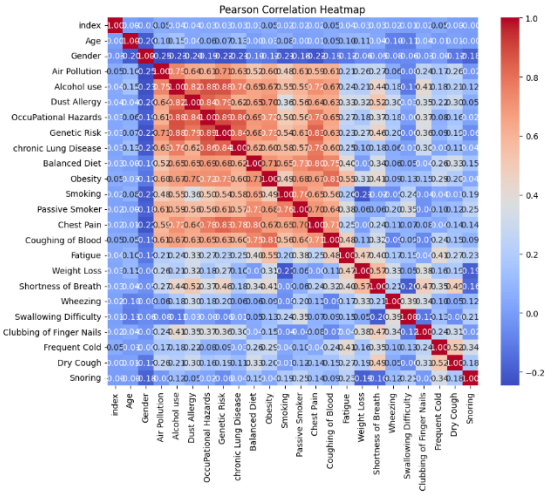


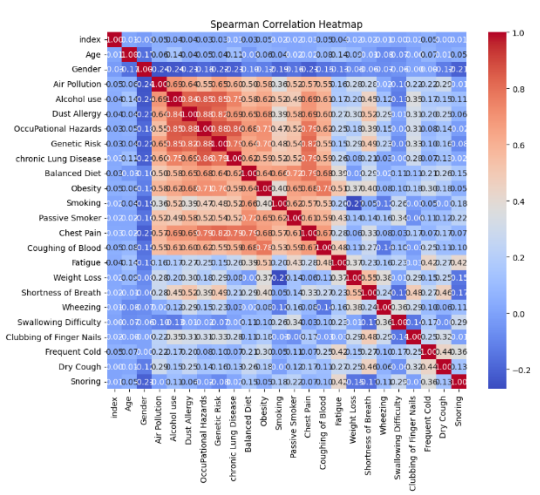
Correlation Analysis:

- There is a [positive/negative] correlation between certain lifestyle factors (e.g., smoking, alcohol use) and the likelihood of developing lung cancer.
- Environmental exposures like air pollution and occupational hazards may show a positive correlation with lung cancer risk.
- Symptoms such as coughing of blood, shortness of breath, and fatigue may exhibit strong correlations with the severity or presence of lung cancer.
- Age may correlate positively with the likelihood of developing lung cancer, while other factors like gender may not show significant correlations.

Code:

```
pearson_corr_matrix = data.corr(method='pearson')
plt.figure(figsize=(10, 8))
sns.heatmap(pearson_corr_matrix, annot=True, cmap='coolwarm',
fmt=".2f")
plt.title('Pearson Correlation Heatmap')
plt.show()
spearman_corr_matrix = data.corr(method='spearman')
plt.figure(figsize=(10, 8))
sns.heatmap(spearman_corr_matrix, annot=True, cmap='coolwarm',
fmt=".2f")
plt.title('Spearman Correlation Heatmap')
plt.show()
```





Regression Analysis:

- Regression analysis indicates that certain predictor variables (e.g., smoking status, age) have a statistically significant impact on the likelihood of lung cancer occurrence.
- The logistic regression model provides insights into the odds ratios of different predictors, highlighting their relative importance in predicting lung cancer risk.

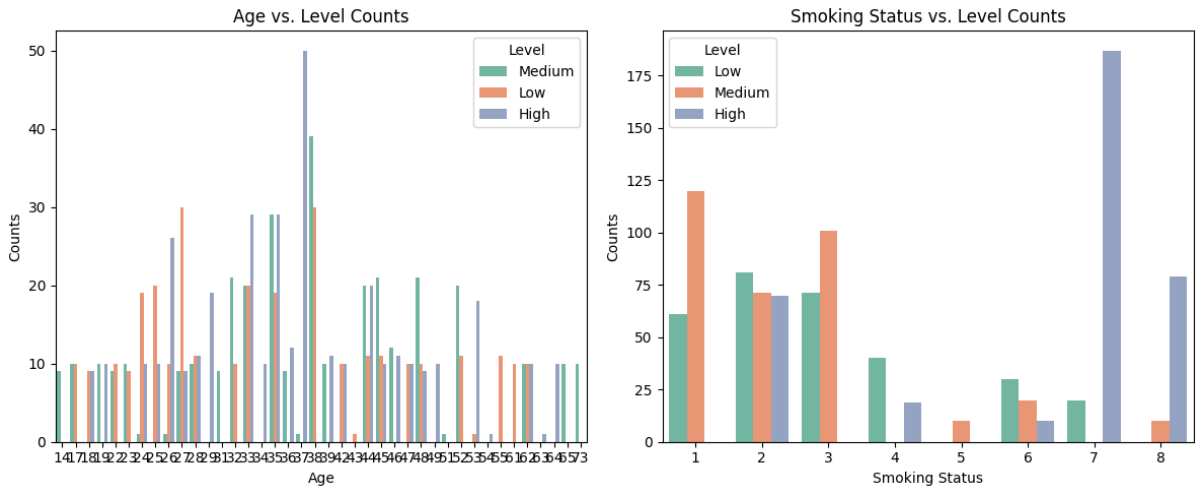
- Multiple linear regression may reveal how combinations of lifestyle factors, environmental exposures, and demographic characteristics contribute to variations in lung cancer severity or progression.

Visualization:

- Visualizations such as scatter plots, box plots, and histograms provide intuitive representations of relationships between variables and distributions within the dataset.
- Pair plots offer a comprehensive view of pairwise interactions between numerical variables, aiding in identifying potential patterns or outliers.

Comparative Visualization of Age/Level and Smoking/Level:

```
plt.figure(figsize=(12, 5))

# Bar chart for age vs. level counts
plt.subplot(1, 2, 1)
sns.countplot(data=data, x='Age', hue='Level', palette='Set2')
plt.xlabel('Age')
plt.ylabel('Counts')
plt.title('Age vs. Level Counts')

# Bar chart for smoking vs. level counts
plt.subplot(1, 2, 2)
sns.countplot(data=data, x='Smoking', hue='Level', palette='Set2')
plt.xlabel('Smoking Status')
plt.ylabel('Counts')
plt.title('Smoking Status vs. Level Counts')

plt.tight_layout()
plt.show()
```



5. REFERENCES:

[1]. Nooreldeen, R., & Bach, H. (2021). Current and future development in lung cancer diagnosis. International journal of molecular sciences, 22(16), 8661.
[2]. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. International Journal of Computer Science and Information Technologies, 4(1), 39-45.
[3]. Christopher, T., & Banu, J. J. (2016). Study of classification algorithm for lung cancer prediction. International Journal of Innovative Science, Engineering & Technology, 3(2).
[4]. Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021, March). Prediction and classification of lung cancer using machine learning techniques. In IOP conference series: materials science and engineering (Vol. 1099, No. 1, p. 012059). IOP Publishing.
[5]. Doppalapudi, S., Qiu, R. G., & Badr, Y. (2021). Lung cancer survival period prediction and understanding: Deep learning approaches. International Journal of Medical Informatics, 148, 104371.
[6]. Doppalapudi, S., Qiu, R. G., & Badr, Y. (2021). Lung cancer survival period prediction and understanding: Deep learning approaches. International Journal of Medical Informatics, 148, 104371.

GROUP CONTRIBUTION DETAILS:

21MIS0196 – MUGUNTHAN N : Goal 1 & 2, Related Questionaries and Metrics, Implementation of Metrics, Dataset collection, Documentation.
21MIS0341 - NITHISH KUMAR R : Goal 3 & 4, Related Questionaries and Metrics, Implementation of Metrics, Dataset collection and Documentation.