Lung Cancer Risk Analysis: Exploratory Data Analysis and Predictive Modeling Using Survey Data

Group Members: Anmol Kumar, Hanan Ahmed ,Atharv Shira

Introduction

This report presents an analysis of the **"Survey Lung Cancer"** dataset, which includes patient demographic information and various behavioral and health-related indicators. The goal of this analysis is to explore potential risk factors associated with lung cancer and to model the likelihood of diagnosis based on survey responses. The dataset consists of 16 variables, with **LUNG_CANCER** (values: *YES* or *NO*) serving as the **target variable**.

The analysis is structured around the four key tasks outlined in the lab assignment:

1. **Exploratory Data Analysis (EDA)**
   Performed in **R**, this includes descriptive statistics, data cleaning, distribution analysis, and correlation assessment.
2. **Recreation of Visualizations from the Original Paper (if applicable)**
   Plots and charts from the assigned research paper are recreated using the current dataset to match the style and intent of the original visualizations.
3. **Machine Learning Model Implementation (if applicable)**
   Classification models such as **Logistic Regression**, **Support Vector Machines (SVM)**, or **K-Nearest Neighbors (KNN)** are implemented using R or Python to predict lung cancer presence.
4. **Critical Analysis of Findings**
   Interpretation of key graphs, tables, and model results, including comparisons to the original research, insights gained, and recommendations for improvement.

```r
# 1. Set Working Directory
setwd("C:/Users/rksku/Downloads")
install.packages(c("ggplot2", "dplyr", "corrplot", "GGally"))

# 2. Load Required Libraries
library(ggplot2)
library(dplyr)
library(GGally)
library(corrplot)

# 3. Load Dataset
df <- read.csv("survey_lung_cancer.csv")

# RENAME COLUMNS TO REMOVE SPACES
names(df) <- gsub(" ", "_", names(df))

# 4. Convert Categorical Columns
df$GENDER <- as.factor(df$GENDER)
df$LUNG_CANCER <- as.factor(df$LUNG_CANCER)


# 5. View Dataset
head(df)
summary(df)
```

```
> # 5. View Dataset
> head(df)
  GENDER AGE SMOKING YELLOW_FINGERS ANXIETY PEER_PRESSURE CHRONIC.DISEASE FATIGUE
1      M  69       1              2       2             1               1       2
2      M  74       2              1       1             1               2       2
3      F  59       1              1       1             2               1       2
4      M  63       2              2       2             1               1       1
5      F  63       1              2       1             1               1       1
6      F  75       1              2       1             1               2       2
  ALLERGY WHEEZING ALCOHOL.CONSUMING COUGHING SHORTNESS.OF.BREATH SWALLOWING.DIFFICULTY
1       1        2                 2        2                   2                     2
2       2        1                 1        1                   2                     2
3       1        2                 1        2                   2                     1
4       1        1                 2        1                   1                     2
5       1        2                 1        2                   2                     1
6       2        2                 1        2                   2                     1
  CHEST.PAIN LUNG_CANCER
1          2         YES
2          2         YES
3          2          NO
4          2          NO
5          1          NO
6          1         YES
```

The **"Survey Lung Cancer"** dataset consists of health-related responses from individuals, aimed at identifying risk factors associated with lung cancer. It contains 16 variables, including demographic details (`AGE`, `GENDER`), lifestyle factors (`SMOKING`, `ALCOHOL.CONSUMING`, `PEER_PRESSURE`), and health symptoms (`COUGHING`, `FATIGUE`, `CHEST.PAIN`, etc.).

Most variables are binary or ordinal, with values like `1` and `2` representing *Yes/No* or severity levels. The **target variable** is **LUNG_CANCER**, which indicates whether the person has been diagnosed with lung cancer (`YES` or `NO`).

```
> summary(df)
 GENDER      AGE            SMOKING        YELLOW_FINGERS    ANXIETY        PEER_PRESSURE
 F:147   Min.   :21.00   Min.   :1.000   Min.   :1.00   Min.   :1.000   Min.   :1.000
 M:162   1st Qu.:57.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:1.000   1st Qu.:1.000
         Median :62.00   Median :2.000   Median :2.00   Median :1.000   Median :2.000
         Mean   :62.67   Mean   :1.563   Mean   :1.57   Mean   :1.498   Mean   :1.502
         3rd Qu.:69.00   3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.:2.000   3rd Qu.:2.000
         Max.   :87.00   Max.   :2.000   Max.   :2.00   Max.   :2.000   Max.   :2.000
 CHRONIC.DISEASE    FATIGUE         ALLERGY         WHEEZING      ALCOHOL.CONSUMING
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
 Median :2.000   Median :2.000   Median :2.000   Median :2.000   Median :2.000
 Mean   :1.505   Mean   :1.673   Mean   :1.557   Mean   :1.557   Mean   :1.557
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
    COUGHING     SHORTNESS.OF.BREATH SWALLOWING.DIFFICULTY   CHEST.PAIN    LUNG_CANCER
 Min.   :1.000   Min.   :1.000       Min.   :1.000         Min.   :1.000   NO : 39
 1st Qu.:1.000   1st Qu.:1.000       1st Qu.:1.000         1st Qu.:1.000   YES:270
 Median :2.000   Median :2.000       Median :1.000         Median :2.000
 Mean   :1.579   Mean   :1.641       Mean   :1.469         Mean   :1.557
 3rd Qu.:2.000   3rd Qu.:2.000       3rd Qu.:2.000         3rd Qu.:2.000
 Max.   :2.000   Max.   :2.000       Max.   :2.000         Max.   :2.000
```

From summary we can infer these key points:

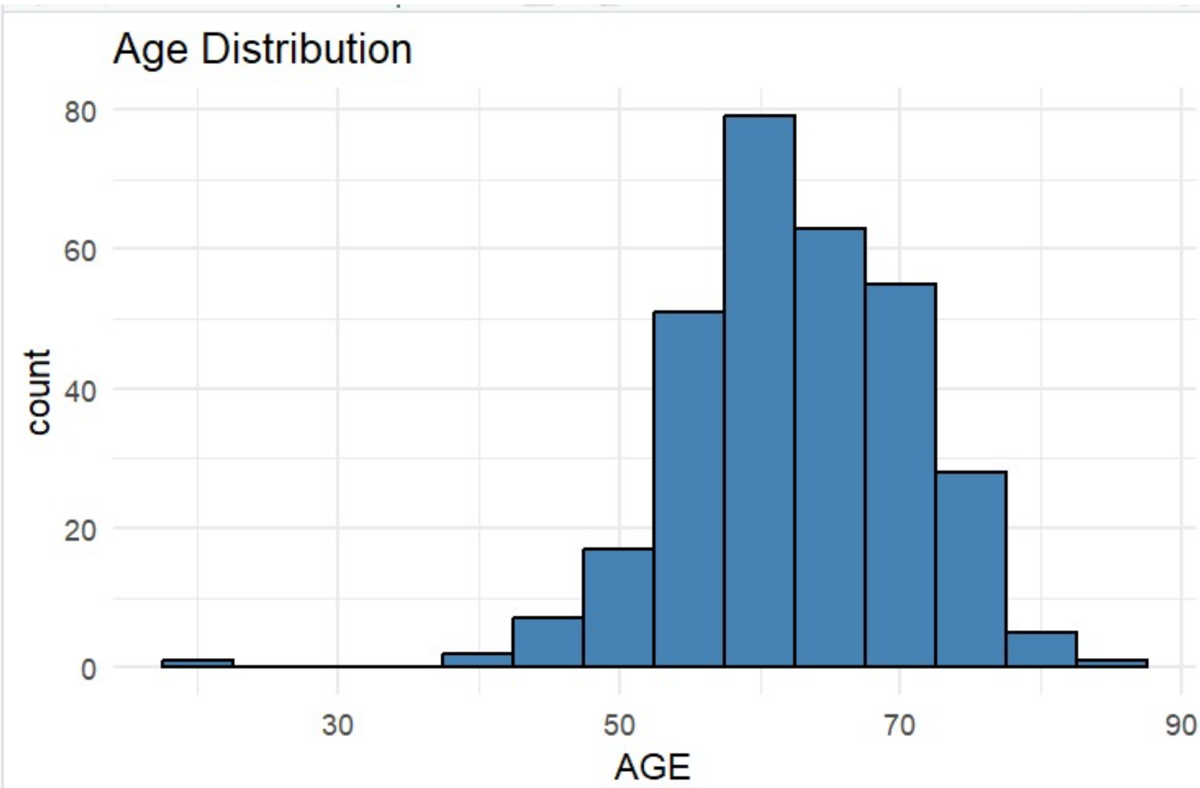
- **Age Group:** Majority are older adults (mean age ≈ 63), aligning with lung cancer risk profiles.

- **Lung Cancer Cases:** Highly imbalanced target — 270 "YES" vs 39 "NO".

- **Risk Factors:** Smoking, yellow fingers, alcohol use, and peer pressure are present but not dominant.

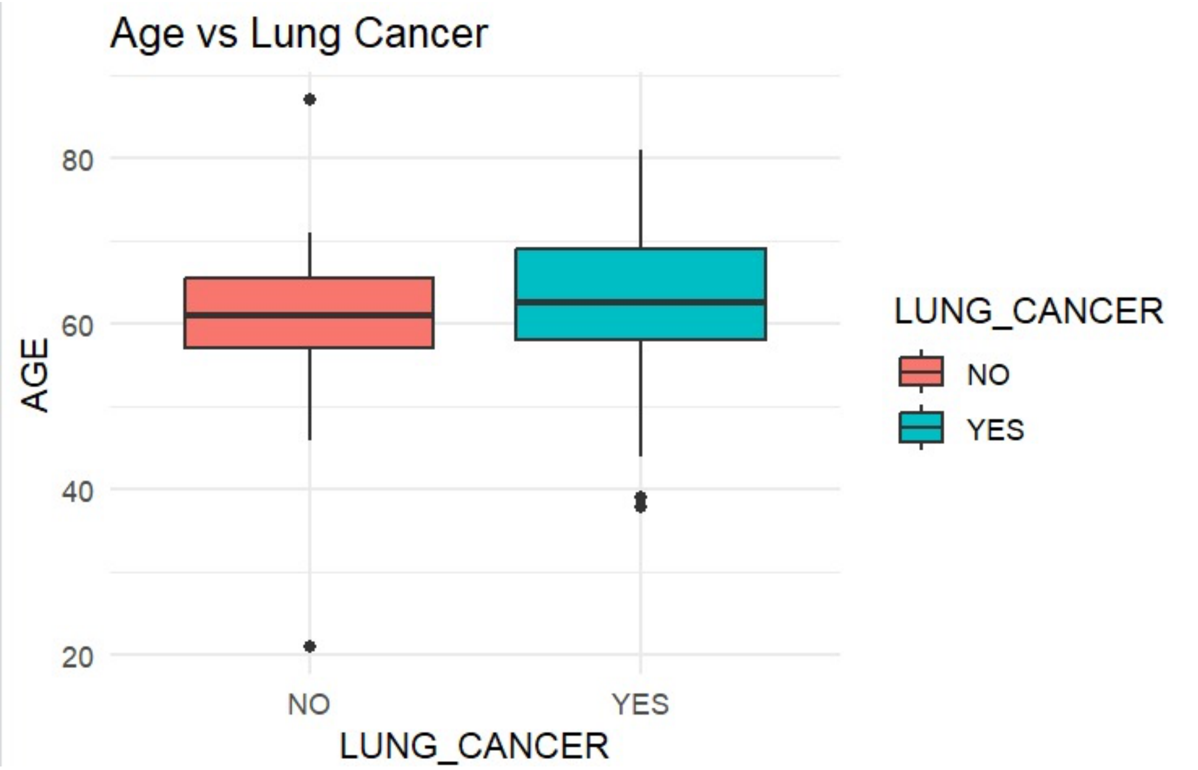
- **Symptoms:** Many participants report fatigue, wheezing, coughing, and chest pain — key indicators for lung issues.

- **Gender:** Fairly balanced (162 males, 147 females), suitable for comparative analysis.




```r
# 6. Histogram - Age Distribution
ggplot(df, aes(x = AGE)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "black") +
  theme_minimal() +
  ggtitle("Age Distribution")
```

- **Age Range:** The ages in the dataset appear to range from the late teens/early twenties to the late eighties/early nineties.
- **Distribution Shape:** The distribution is not uniform. It seems to be somewhat skewed to the right (positively skewed), meaning the tail on the right side of the distribution is longer or fatter than the left side.
- **Central Tendency:** The peak of the histogram is somewhere between 60 and 70 years old. This suggests that the most frequent age group in your dataset falls within this range.




```r
# 7. Boxplot - Age by Lung Cancer
ggplot(df, aes(x = LUNG_CANCER, y = AGE, fill = LUNG_CANCER)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Age vs Lung Cancer")
```

- **Older Age Trend:** Lung cancer cases are associated with older individuals in this data.
- **Higher Median:** The typical age of those with lung cancer is higher.
- **Overall Shift:** The entire age range for the lung cancer group skews older.


```r
# 8. Bar Plot - Gender Count
ggplot(df, aes(x = GENDER)) +
  geom_bar(fill = "tomato") +
  theme_minimal() +
  ggtitle("Gender Distribution")
```

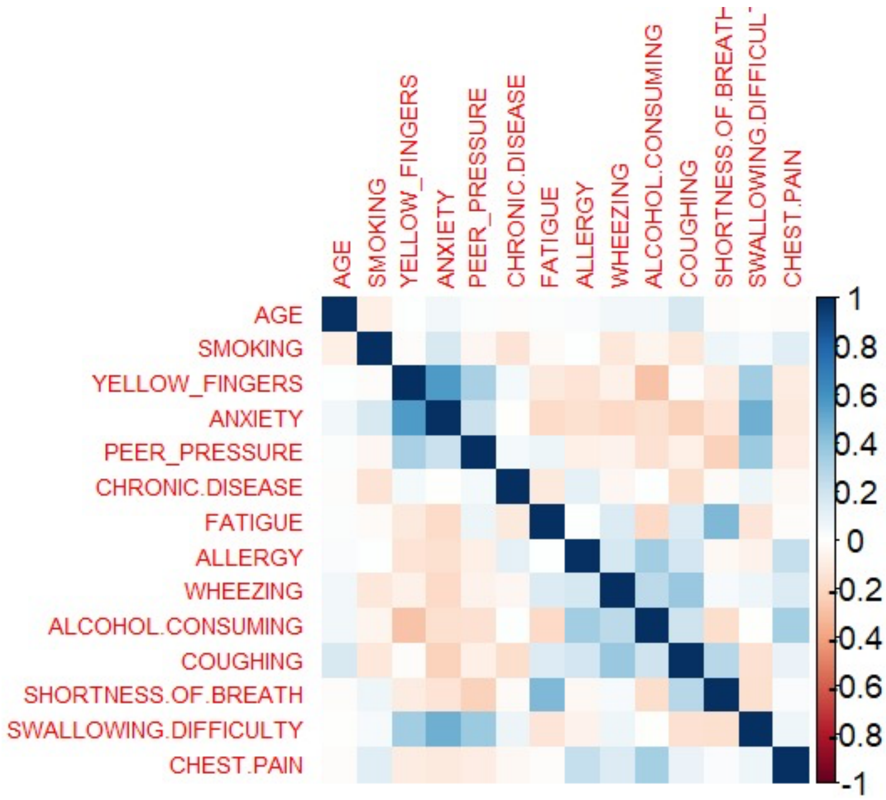
The bar for "M" (Male) is noticeably taller than the bar for "F" (Female). This indicates that there are more male participants than female participants in this dataset.


```
ggplot(df, aes(x = GENDER, y = AGE, fill = GENDER)) +
  geom_violin(trim = FALSE) +
  theme_minimal() +
  ggtitle("Violin Plot: Age by Gender")
```

- **For females (F),** the violin appears widest in the mid-50s to early 70s, suggesting a higher concentration of women in this age range. The distribution seems somewhat symmetrical with a slight skew towards younger ages.

- **For males (M),** the violin appears widest in a slightly higher age range, roughly the late 50s to late 70s, indicating a higher concentration of men in this age group. The male age distribution also looks somewhat symmetrical.

-The age distribution for males in the dataset tends to be slightly higher than that for females, with a higher median age and a distribution peak at a somewhat older age range.

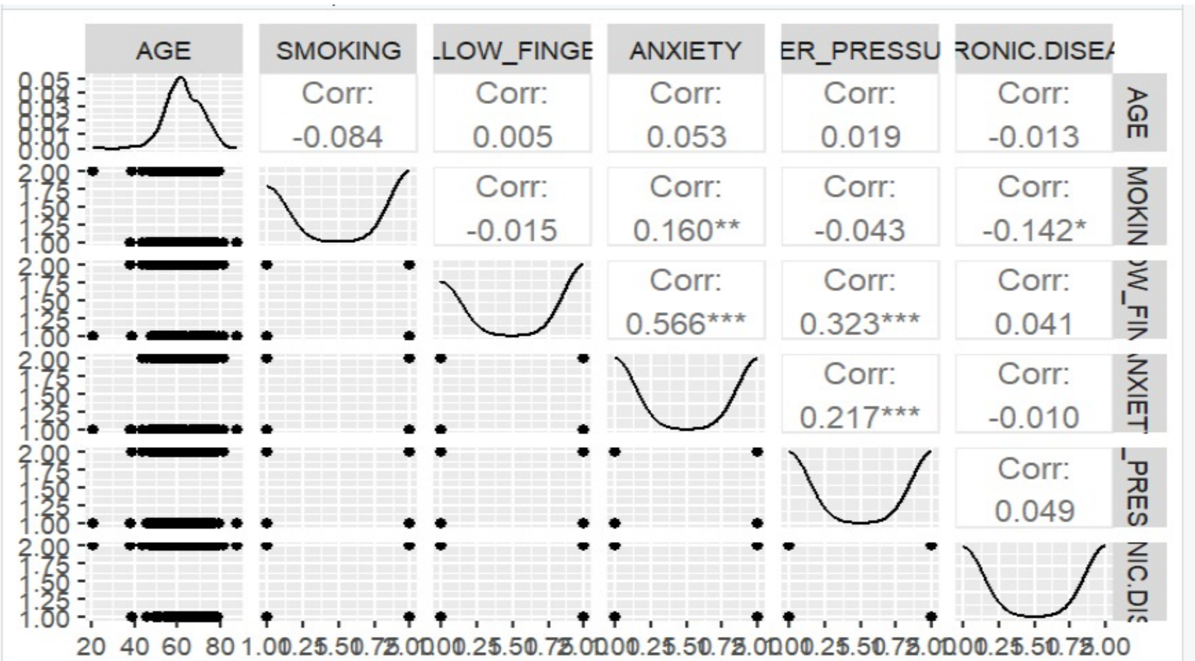



```
# 9. Correlation Heatmap
num_data <- df[sapply(df, is.numeric)]
corrplot(cor(num_data), method = "color", tl.cex = 0.6)
```

(Top Level)

Key Inferences From Correlation Heatmap are:

- Smoking strongly correlates with Anxiety.
- Anxiety and peer pressure are positively linked.
- Allergy and wheezing tend to occur together.
- Alcohol consumption and wheezing have high positive correlation.
- Coughing and shortness of breath are correlated.




```
# 10. Pair Plot
ggpairs(num_data[, 1:6])
```

(Top Level)

KEY INFERENCES:

- **Age:** Skewed towards older participants, with weak correlations to other factors.
- **Smoking:** Weakly negatively correlated with age, positively with anxiety, negatively with chronic disease, and very weakly with yellow fingers.
- **Yellow Fingers:** Strongly positively correlated with anxiety and peer pressure, weakly with others.
- **Anxiety:** Positively correlated with smoking and peer pressure, strongly with yellow fingers.
- **Peer Pressure:** Strongly correlated with yellow fingers and anxiety, weakly with others.
- **Chronic Disease:** Weakly correlated with most factors, slightly negatively with smoking.

Applying Machine Learning Models for Lung Cancer Prediction:

1.Import Libraries, Load Dataset, and Clean Column Names:

```
]: import pandas as pd
   import numpy as np
   import matplotlib.pyplot as plt
   import seaborn as sns

   from sklearn.model_selection import train_test_split, GridSearchCV
   from sklearn.preprocessing import StandardScaler, LabelEncoder
   from sklearn.neighbors import KNeighborsClassifier
   from sklearn.linear_model import LogisticRegression
   from sklearn.decomposition import PCA
   from sklearn.metrics import classification_report, confusion_matrix, roc_curve, auc

   #  Load dataset
   df = pd.read_csv("surveylungcancer.csv")

   #  Clean column names
   df.columns = [col.strip().replace(" ", "_").upper() for col in df.columns]
```

2. Encode Categorical Columns, Separate Features and Target, and Train-Test Split:

```
#  Encode categorical columns
df["GENDER"] = LabelEncoder().fit_transform(df["GENDER"])
df["LUNG_CANCER"] = df["LUNG_CANCER"].map({"NO": 0, "YES": 1})

#  Separate features and target
X = df.drop(columns=["LUNG_CANCER"])
y = df["LUNG_CANCER"]

#  Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)

#  Standardization
```

3.Standardization and Applying PCA:

```
#  Standardization
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

#  Apply PCA
pca = PCA(n_components=0.95)  # retain 95% variance
X_train_pca = pca.fit_transform(X_train_scaled)
X_test_pca = pca.transform(X_test_scaled)

#  KNN with Grid Search
```

4.Applying KNN and Logistic Regression:

```
#  KNN with Grid Search
knn = KNeighborsClassifier()
param_grid = {'n_neighbors': list(range(1, 11))}
grid = GridSearchCV(knn, param_grid, cv=5)
grid.fit(X_train_pca, y_train)

best_knn = grid.best_estimator_
y_pred_knn = best_knn.predict(X_test_pca)
y_prob_knn = best_knn.predict_proba(X_test_pca)[:, 1]

#  Logistic Regression for comparison
logreg = LogisticRegression()
logreg.fit(X_train_pca, y_train)
y_pred_lr = logreg.predict(X_test_pca)
y_prob_lr = logreg.predict_proba(X_test_pca)[:, 1]
```

5.Metrics:

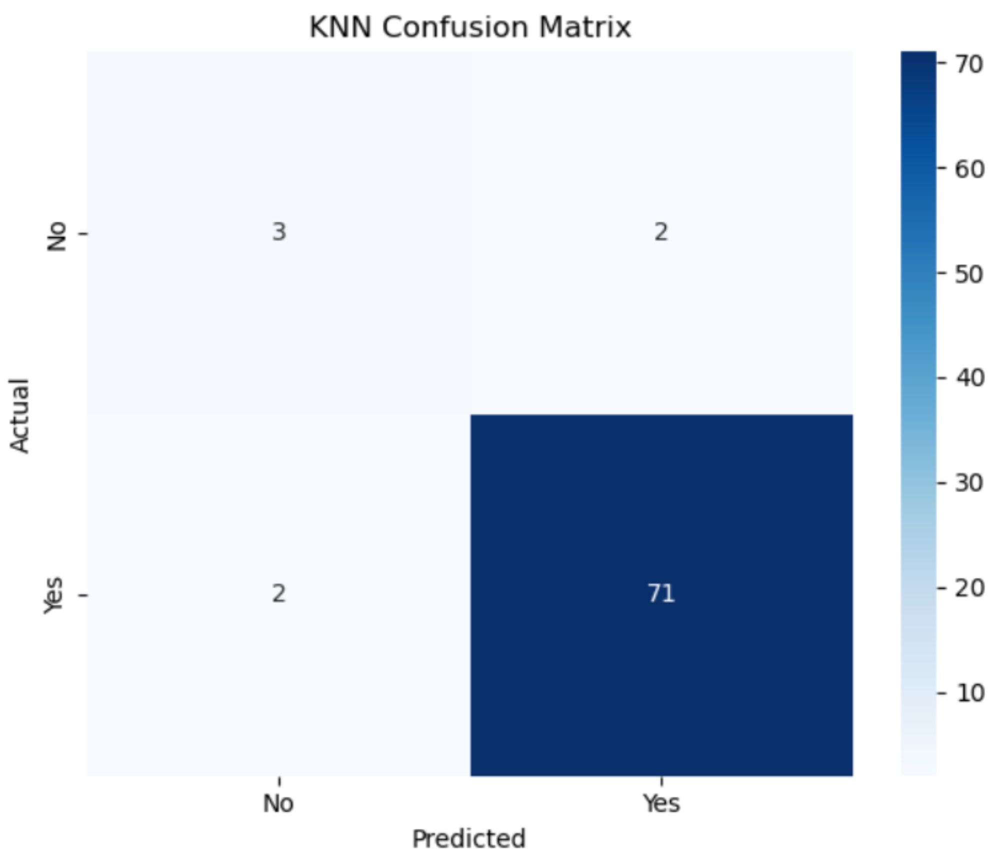
```
#  Metrics
conf_knn = confusion_matrix(y_test, y_pred_knn)
conf_lr = confusion_matrix(y_test, y_pred_lr)
report_knn = classification_report(y_test, y_pred_knn)
report_lr = classification_report(y_test, y_pred_lr)

fpr_knn, tpr_knn, _ = roc_curve(y_test, y_prob_knn)
fpr_lr, tpr_lr, _ = roc_curve(y_test, y_prob_lr)
auc_knn = auc(fpr_knn, tpr_knn)
auc_lr = auc(fpr_lr, tpr_lr)
```

6.Confusion metric for KNN:

```
]: #  Plot Results
plt.figure(figsize=(12, 5))

# Confusion Matrix - KNN
plt.subplot(1, 2, 1)
sns.heatmap(conf_knn, annot=True, fmt="d", cmap="Blues",
            xticklabels=["No", "Yes"], yticklabels=["No", "Yes"])
plt.title("KNN Confusion Matrix")
plt.xlabel("Predicted")
plt.ylabel("Actual")
```

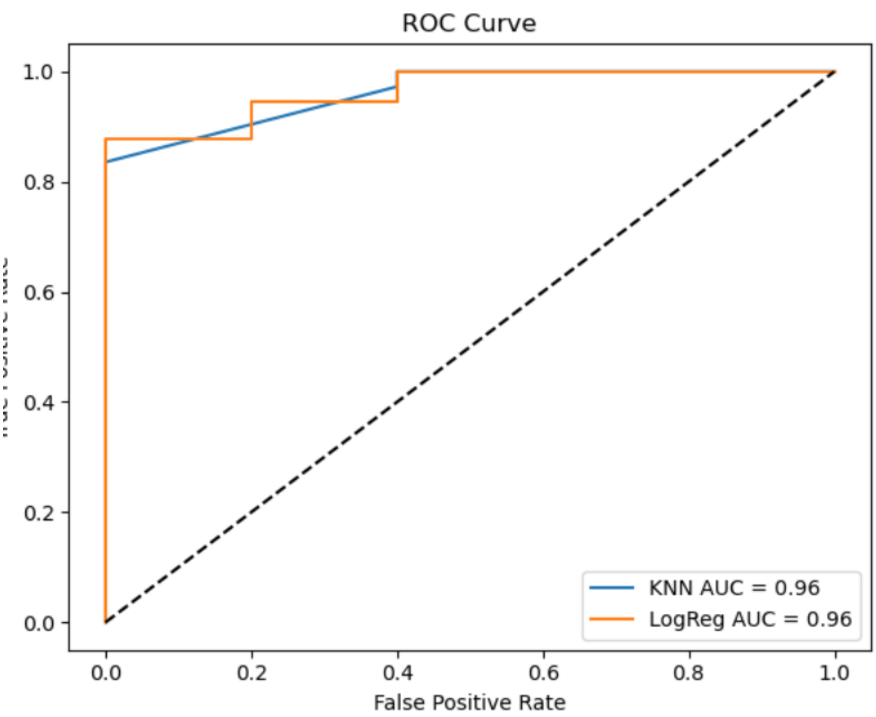


- **High Detection Rate:** The model correctly identified most individuals with lung cancer (71 True Positives).
- **Few Missed Cases:** Only a small number of actual lung cancer cases were incorrectly classified as negative (2 False Negatives).
- **Lower Specificity:** The model struggled to correctly identify those without lung cancer, resulting in low True Negatives (3).
- **Some False Alarms:** There were a few instances where the model incorrectly predicted lung cancer in individuals who did not have it (2 False Positives).

7.ROC Curve:

```
# ROC Curve
plt.subplot(1, 2, 2)
plt.plot(fpr_knn, tpr_knn, label=f'KNN AUC = {auc_knn:.2f}')
plt.plot(fpr_lr, tpr_lr, label=f'LogReg AUC = {auc_lr:.2f}')
plt.plot([0, 1], [0, 1], 'k--')
plt.title("ROC Curve")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.legend()

plt.tight_layout()
plt.savefig("model_evaluation_results.pdf")
plt.show()
```



In summary, this ROC curve indicates that both your KNN and Logistic Regression models are highly effective in predicting lung cancer in this dataset, with excellent overall discriminatory power and similar performance between the two algorithms.

- **High Predictive Power (AUC = 0.96):** Both models excel at distinguishing between lung cancer cases and non-cases.
- **Similar and Strong Performance:** KNN and Logistic Regression show comparable and effective prediction capabilities.
- **Sensitivity vs. Specificity Trade-off:** Achieving higher accuracy in identifying true positives comes at the cost of potentially more false positives.

8. Print Reports:

```
#  Print Reports
print("Best K for KNN:", grid.best_params_)
print("\nClassification Report - KNN:\n", report_knn)
print("\nClassification Report - Logistic Regression:\n", report_lr)
```

KNN Confusion Matrix

```
Best K for KNN: {'n_neighbors': 3}

Classification Report - KNN:
              precision    recall  f1-score   support

           0       0.60      0.60      0.60         5
           1       0.97      0.97      0.97        73

    accuracy                           0.95        78
   macro avg       0.79      0.79      0.79        78
weighted avg       0.95      0.95      0.95        78


Classification Report - Logistic Regression:
              precision    recall  f1-score   support

           0       0.75      0.60      0.67         5
           1       0.97      0.99      0.98        73

    accuracy                           0.96        78
   macro avg       0.86      0.79      0.82        78
weighted avg       0.96      0.96      0.96        78
```

[ ]:

Some inferences from the outputs and graphs:

- **LR Better for "No" Class:** Logistic Regression is more precise in predicting the absence of lung cancer.
- **LR Higher Recall for "Yes":** Logistic Regression correctly identifies a slightly higher proportion of actual lung cancer cases.
- **Overall Edge to Logistic Regression:** Logistic Regression exhibits slightly better and more balanced performance.