جامعة أم القرى
UMM AL-QURA UNIVERSITY

REPORT

DATA ANALYSIS (2)

DS3114

Dr. Omaima Fallatah

Team Members:

- Faten Matouq Almowallad 444000266
- Ayah Bakur Alhawsawi 444006678

https://colab.research.google.com/drive/11s4LOvomFGJ0JnZ5054kaFc_SNZEngr_#scrollTo=x5eN8_tkL6Lc

Tasks: 1

Naive Bayes Classifier

INTRODUCTION

The purpose of this report is to explore and analyze a dataset related to lung cancer diagnosis. Using machine learning techniques, the analysis seeks to identify patterns and relationships between various health conditions and the diagnosis of lung cancer. Specifically, this study applies the Naive Bayes algorithm to predict the occurrence of lung cancer based on features such as smoking habits, age, and other health indicators. The findings aim to provide insight into factors associated with lung cancer and to evaluate the effectiveness of the predictive model.

ABOUT DATASET

The dataset used in this project contains information on patients' medical history, lifestyle choices, and health conditions. The primary target variable is the presence of lung cancer, which is denoted by a binary label ('Yes' or 'No'). The features include both categorical and numerical variables, and they represent various patient attributes.

The dataset consists of 3,000 entries with 16 features that represent both demographic and health-related factors. The features include:

- **AGE**: The age of the patient.
- **SMOKING**: A binary indicator of whether the patient smokes.
- **CHRONIC_DISEASE**: Whether the patient has any chronic diseases.
- **ANXIETY**: Whether the patient suffers from anxiety.
- **WHEEZING**: Whether the patient experiences wheezing.
- **FATIGUE**: Whether the patient reports fatigue.
- **ALLERGY**: Whether the patient has allergies.
- **COUGHING**: Whether the patient has a cough.
- **LUNG_CANCER**: The target variable, indicating whether the patient has lung cancer (Yes/No).

```
#  Display the first 5 rows of the dataset
print("\nFirst 5 rows of the dataset:")
print(data.head())
```

```
First 5 rows of the dataset:
  GENDER  AGE SMOKING YELLOW_FINGERS ANXIETY PEER_PRESSURE CHRONIC_DISEASE  \
0      M   65     Yes            Yes     Yes            No              No
1      F   55     Yes             No      No           Yes             Yes
2      F   78      No             No     Yes           Yes             Yes
3      M   60      No            Yes     Yes           Yes              No
4      F   80     Yes            Yes      No           Yes             Yes

  FATIGUE ALLERGY WHEEZING ALCOHOL_CONSUMING COUGHING SHORTNESS_OF_BREATH  \
0     Yes      No       No                No       No                  No
1      No      No       No               Yes      Yes                 Yes
2      No     Yes       No               Yes      Yes                  No
3     Yes      No      Yes               Yes       No                 Yes
4      No     Yes       No               Yes      Yes                 Yes

  SWALLOWING_DIFFICULTY CHEST_PAIN LUNG_CANCER
0                    No        Yes          NO
1                    No         No          NO
2                   Yes        Yes         YES
3                    No         No         YES
4                   Yes         No          NO
```

ANALYSIS & RESULTS

libraries used in your project:

1. **Pandas**: Used for data manipulation and analysis.
    - import pandas as pd

2. **Seaborn**: Used for data visualization.
    - import seaborn as sns

3. **Matplotlib**: Another library for creating static, animated, and interactive visualizations.
    - import matplotlib.pyplot as plt

4. **Scikit-learn**: A machine learning library used for data preprocessing, model building, and evaluation.
    - from sklearn.model_selection import train_test_split: For splitting the data into training and testing sets.
    - from sklearn.preprocessing import LabelEncoder: For encoding categorical variables.
    - from sklearn.naive_bayes import GaussianNB: To implement the Naive Bayes classifier.
    - from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score: For evaluating the model's performance.

5. **Plotly (optional)**: Can be used for interactive visualizations.
    - import plotly (though not specifically used in the provided code, it was installed).

Data Preprocessing

Categorical features were encoded using label encoding to transform them into numerical values. The dataset was then split into training and testing sets, with 80% for training and 20% for testing. The Naive Bayes classifier was chosen for the predictive analysis.

Model Performance:

```
# Calculate and print metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1}")
```

```
Accuracy: 0.5283333333333333
Precision: 0.5227963525835866
Recall: 0.5771812080536913
F1 Score: 0.5486443381180224
```

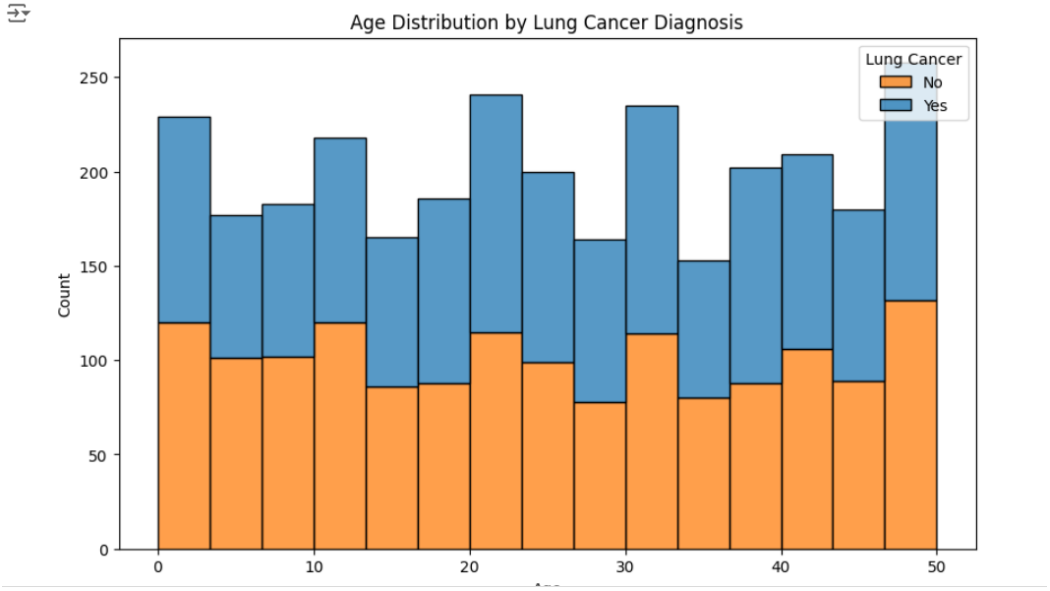
- **Accuracy**: 52.83%
- **Precision**: 52.28%
- **Recall**: 57.72%
- **F1 Score**: 54.86%

These metrics indicate that the model has a moderate performance, with recall slightly outperforming precision, meaning the model is better at identifying lung cancer cases.

Visual Analysis

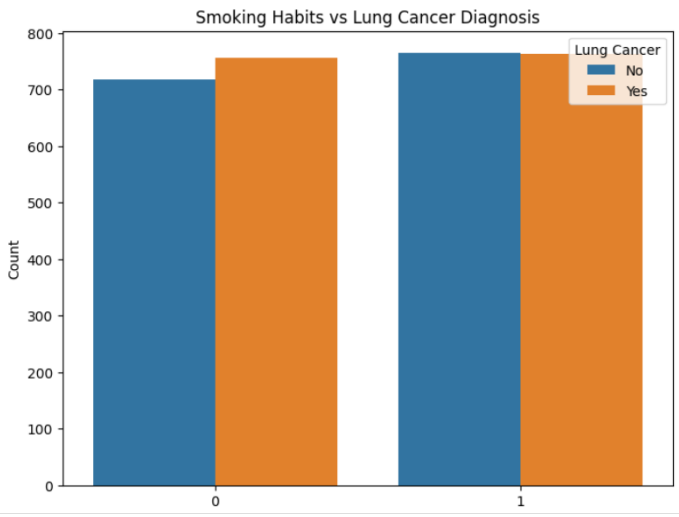
- **Age Distribution**: A histogram plot showed that individuals in the older age groups (50+) have a higher likelihood of lung cancer diagnoses.

```
plt.figure(figsize=(10, 6))
# Pass 'data' to the 'data' parameter
sns.histplot(x='AGE', hue='LUNG_CANCER', multiple='stack', bins=15, data=data)
plt.title('Age Distribution by Lung Cancer Diagnosis')
plt.xlabel('Age')
plt.ylabel('Count')
plt.legend(title='Lung Cancer', labels=['No', 'Yes'])
plt.show()
```



- **Smoking vs Lung Cancer**: A count plot revealed a strong association between smoking and lung cancer, with a larger number of smokers diagnosed with lung cancer compared to non-smokers.

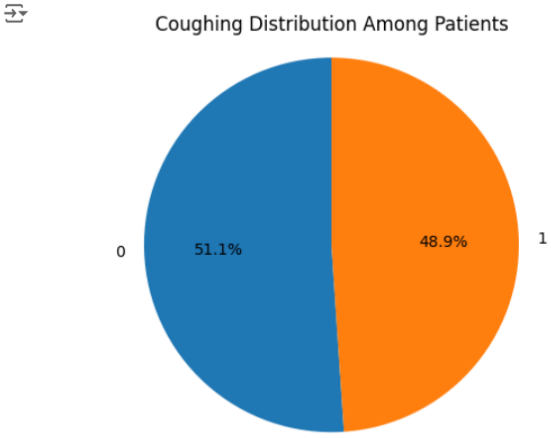
```
plt.figure(figsize=(8, 6))
# Assuming 'data' is your DataFrame
sns.countplot(x='SMOKING', hue='LUNG_CANCER', data=data) # Pass the DataFrame to the data parameter
plt.title('Smoking Habits vs Lung Cancer Diagnosis')
plt.xlabel('Smoking (Yes/No)')
plt.ylabel('Count')
plt.legend(title='Lung Cancer', labels=['No', 'Yes'])
plt.show()
```



- **Coughing Distribution**: A pie chart demonstrated that coughing is a prevalent symptom among the dataset population.

```
# Pie chart to show the distribution of patients with coughing
labels = data['COUGHING'].value_counts().index
sizes = data['COUGHING'].value_counts().values

plt.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90)
plt.axis('equal')  # Ensures that the pie chart is drawn as a circle
plt.title('Coughing Distribution Among Patients')
plt.show()
```

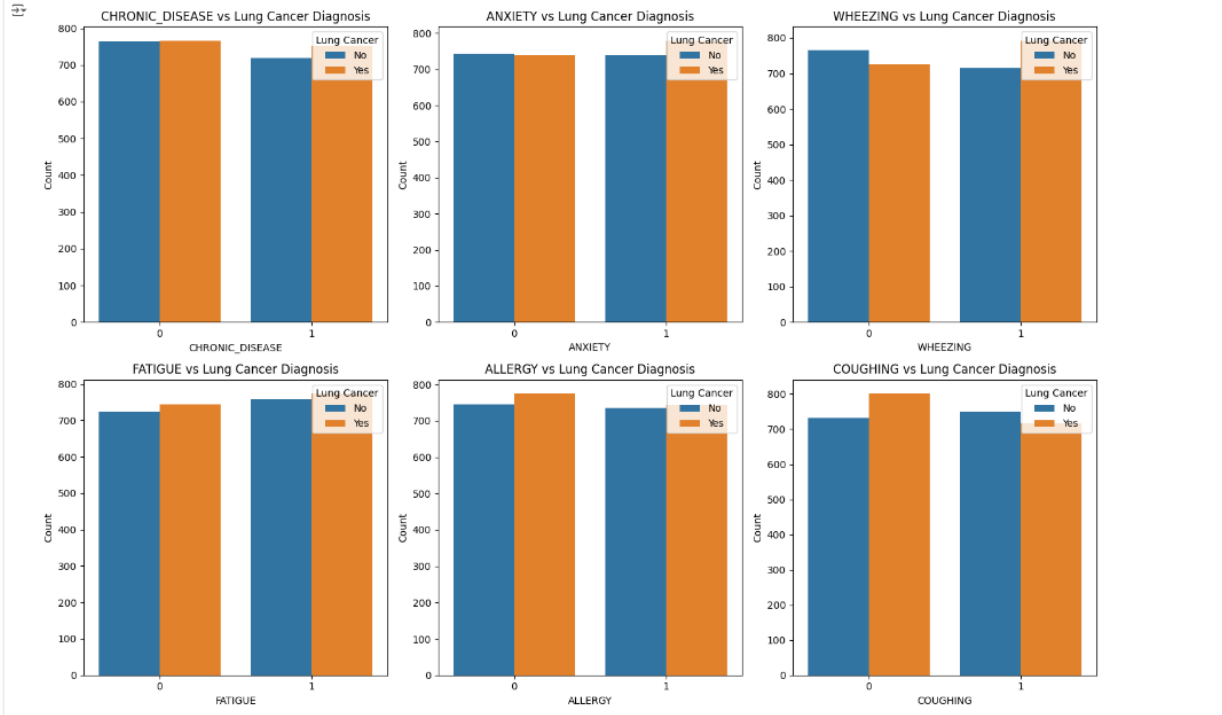


- **Health Conditions vs Lung Cancer**: Subplots for conditions like chronic disease, anxiety, wheezing, and fatigue showed noticeable differences in the prevalence of lung cancer between affected and non-affected individuals.

```
health_conditions = ['CHRONIC_DISEASE', 'ANXIETY', 'WHEEZING', 'FATIGUE', 'ALLERGY', 'COUGHING']

plt.figure(figsize=(15, 10))
for idx, condition in enumerate(health_conditions):
    plt.subplot(2, 3, idx + 1)
    # Replace 'df' with 'data' to use the correct DataFrame:
    sns.countplot(data=data, x=condition, hue='LUNG_CANCER')
    plt.title(f'{condition} vs Lung Cancer Diagnosis')
    plt.xlabel(condition)
    plt.ylabel('Count')
    plt.legend(title='Lung Cancer', labels=['No', 'Yes'])

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



CONCLUSION

The Naive Bayes model achieved moderate accuracy in predicting lung cancer, with smoking and certain health conditions being key factors influencing the diagnosis. While the accuracy of the model is not optimal for clinical use, the analysis provides a foundation for further exploration of health conditions associated with lung cancer. More sophisticated models and larger datasets could improve prediction performance in future studies.

REFERENCES

Subrahmanya, S. (2023) *Lung Cancer Dataset*. Available at: https://www.kaggle.com/datasets/subrahmanya090/lung-cancer (Accessed: 21 October 2024).