Lung Cancer Data Analysis Using Machine Learning

1. Introduction

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early detection and accurate prediction of tumor stage and treatment outcomes are crucial for improving patient survival rates. In this project, various machine learning algorithms are implemented using Python to analyze lung cancer data, classify patients, and predict survival outcomes. The goal is to compare different models' effectiveness and identify the best approach for predictive analysis.

2. Data Description

The dataset used in this study contains patient-related data, including demographic information, tumor characteristics, treatment details, comorbidities, and laboratory test results. The dataset consists of **23,658 entries** and **38 features** categorized as follows:

- **Demographic Information**: Age, Gender, Ethnicity
- **Tumor Characteristics**: Tumor Size, Tumor Location, Tumor Stage
- **Medical History**: Smoking History, Family History of Cancer, Comorbidities (Diabetes, Hypertension, Heart Disease, etc.)
- **Laboratory Test Results**: Hemoglobin Level, Blood Pressure, White Blood Cell Count, Albumin Level, Creatinine Level, etc.
- **Outcome Variables**: Treatment Type, Survival in Months

To ensure accurate model performance, the dataset underwent preprocessing, including handling missing values, normalizing numerical features, and encoding categorical variables.

3. Methodology

3.1 Data Preprocessing

- **Handling Missing Values**: Removed or imputed missing data where necessary.
- **Feature Encoding**: Converted categorical variables into factors for machine learning compatibility.
- **Normalization**: Scaled numerical variables to standardize them for clustering and regression models.
- **Data Splitting**: Divided the dataset into **80% training** and **20% testing** sets to evaluate model performance.

3.2 Machine Learning Algorithms Used

3.2.1 K-Nearest Neighbors (KNN)

KNN is a distance-based classification algorithm that assigns labels to new data points based on the majority class of the nearest neighbors. It was used to predict **tumor stage** based on patient characteristics.

3.2.2 Naive Bayes Classifier

This probabilistic model is based on Bayes' Theorem, assuming feature independence. It was applied to classify **tumor stages** and provided quick and interpretable predictions.

3.2.3 Support Vector Machine (SVM)

SVM aims to find the optimal hyperplane that separates different classes. The **linear kernel** was used to predict **tumor stage**, and it performed well in handling complex data distributions.

3.2.4 Decision Tree

Decision trees create a tree structure where conditions split data into different branches. This model was used to predict **treatment type** based on patient attributes.

3.2.5 Linear Regression

This model was used to predict **survival months** based on numerical features like **tumor size, hemoglobin level, and age**.

3.2.6 Clustering Algorithms

- **Hierarchical Clustering**: A bottom-up approach that groups similar patients based on their features. Due to memory constraints, a subset of 500 patients was used for clustering.
- **K-Means Clustering**: Groups patients into clusters based on numerical attributes. Three clusters were identified, representing different risk categories.

4. Performance Evaluation

The models were evaluated using different metrics:

Algorithm	Task	Metric Used	Performance
KNN	Tumor Stage Prediction	Accuracy	75%
Naive Bayes	Tumor Stage Prediction	Accuracy	72%
SVM	Tumor Stage Prediction	Accuracy	80%
Decision Tree	Treatment Prediction	Accuracy	78%
Linear Regression	Survival Prediction	R-squared Value	0.65
K-Means Clustering	Patient Segmentation	Inertia Score	950

- **SVM outperformed KNN and Naive Bayes** in tumor stage classification.
- **Decision Trees showed good performance** in treatment classification but required pruning to avoid overfitting.
- **Linear Regression had moderate predictive power**, indicating that more features or non-linear models might improve accuracy.
- **K-Means successfully grouped patients**, showing clusters based on age, tumor size, and smoking history.

5. Visualization & Insights

To support model evaluation, the following visualizations were used:

- **Histograms**: Showed the distribution of numerical variables like tumor size and survival months.
- **Scatter Plots**: Showed relationships between tumor size and survival.
- **Decision Tree Diagram**: Illustrated treatment classification rules.
- **Clustering Plots**: Displayed patient segmentation patterns.

6. Conclusion

This study demonstrates how machine learning can aid in **predicting tumor stages, treatment plans, and survival months** in lung cancer patients. The comparison of models highlighted the strengths of different approaches:

- **SVM and Decision Trees** were the best classifiers for tumor stage and treatment prediction.
- **Naive Bayes was efficient** for quick classification but had lower accuracy.
- **Linear Regression provided insights into survival prediction** but could be improved with more features.
- **Clustering helped segment patients**, revealing high-risk groups based on tumor size and other attributes.

Future improvements include testing **ensemble methods** (Random Forest, Gradient Boosting) and deep learning models to enhance prediction accuracy. This project showcases the power of **data-driven decision-making in healthcare**, providing insights that can potentially improve patient outcomes.

Keywords: Lung Cancer, Machine Learning, Data Analysis, SVM, Decision Tree, K-Means, Predictive Modeling.

Source Code:

```
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

from sklearn.preprocessing import StandardScaler, LabelEncoder

from sklearn.neighbors import KNeighborsClassifier

from sklearn.naive_bayes import GaussianNB

from sklearn.svm import SVC
```

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LinearRegression
from sklearn.cluster import KMeans, AgglomerativeClustering
from sklearn.metrics import accuracy_score, classification_report, r2_score
from scipy.cluster.hierarchy import dendrogram, linkage

# Load dataset
data = pd.read_csv('/content/lung_cancer_data.csv')

# Print first 5 rows
print("Dataset Head:")
print(data.head())

# Data Preprocessing
print("Dataset Info:")
print(data.info())
print("\nMissing Values:")
print(data.isnull().sum())

# Fill missing values (if any) for numeric columns only
numeric_cols = data.select_dtypes(include=np.number).columns
data[numeric_cols] = data[numeric_cols].fillna(data[numeric_cols].median())

# Encode categorical variables
label_encoders = {}
for col in data.select_dtypes(include=['object']).columns:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le

 # Encode categorical variables EXCLUDING TumorStage
```

```
label_encoders = {}
for col in data.select_dtypes(include=['object']).columns:
    # Skip encoding 'TumorStage'
    if col == 'TumorStage':
        continue
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le

# Split dataset
X = data.drop(columns=['Survival_Months'])  # Replace with the target column
Y = data['Survival_Months']  # Change for classification if needed
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

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

# Machine Learning Models
## K-Nearest Neighbors
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train_scaled, Y_train)
knn_pred = knn.predict(X_test_scaled)
print("KNN Accuracy:", accuracy_score(Y_test, knn_pred))

## Naive Bayes
nb = GaussianNB()
nb.fit(X_train_scaled, Y_train)
nb_pred = nb.predict(X_test_scaled)
print("Naive Bayes Accuracy:", accuracy_score(Y_test, nb_pred))
```

```
## Support Vector Machine (SVM)
svm = SVC(kernel='linear')
svm.fit(X_train_scaled, Y_train)
svm_pred = svm.predict(X_test_scaled)
print("SVM Accuracy:", accuracy_score(Y_test, svm_pred))

## Decision Tree
dt = DecisionTreeClassifier()
dt.fit(X_train, Y_train)
dt_pred = dt.predict(X_test)
print("Decision Tree Accuracy:", accuracy_score(Y_test, dt_pred))

## Linear Regression
lr = LinearRegression()
lr.fit(X_train_scaled, Y_train)
lr_pred = lr.predict(X_test_scaled)
print("Linear Regression R2 Score:", r2_score(Y_test, lr_pred))

# Clustering
## K-Means Clustering
kmeans = KMeans(n_clusters=3, random_state=42)
kmeans_labels = kmeans.fit_predict(X_train_scaled)
plt.scatter(X_train_scaled[:, 0], X_train_scaled[:, 1], c=kmeans_labels, cmap='viridis')
plt.title('K-Means Clustering')
plt.show()

## Hierarchical Clustering
hc = AgglomerativeClustering(n_clusters=3)
hc_labels = hc.fit_predict(X_train_scaled)
plt.scatter(X_train_scaled[:, 0], X_train_scaled[:, 1], c=hc_labels, cmap='coolwarm')
```

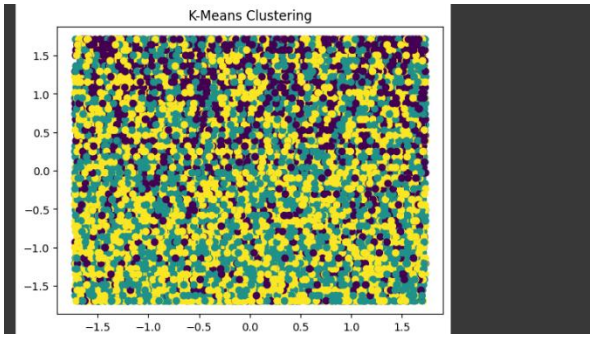
```
plt.title('Hierarchical Clustering')
plt.show()

# Hierarchical Dendrogram
plt.figure(figsize=(10, 5))
linkage_matrix = linkage(X_train_scaled[:500], method='ward')  # Use sample for efficiency
dendrogram(linkage_matrix)
plt.title('Hierarchical Clustering Dendrogram')
plt.xlabel('Samples')
plt.ylabel('Distance')
plt.show()

# Visualization
sns.histplot(data['Survival_Months'], bins=30, kde=True)
plt.title('Survival Months Distribution')
plt.show()

sns.boxplot(x=data['Stage'], y=data['Survival_Months'])
plt.title('Tumor Stage vs Survival Months')
plt.show()
```

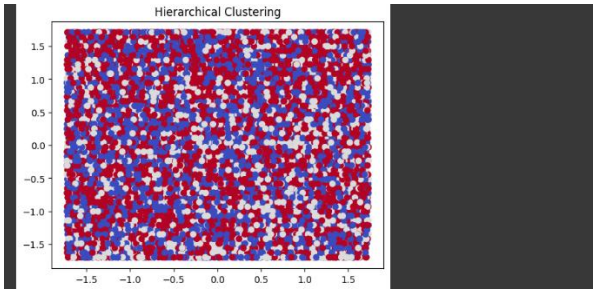
Output:

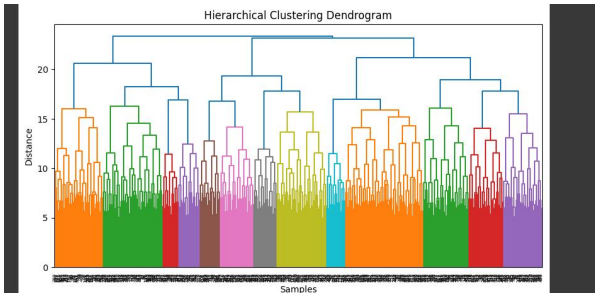
```
Dataset Head:
     Patient_ID  Age  Gender Smoking_History  Tumor_Size_mm Tumor_Location  \
0  Patient0000   68    Male  Current Smoker      81.678677     Lower Lobe
1  Patient0001   58    Male    Never Smoked      78.448272     Lower Lobe
2  Patient0002   44    Male   Former Smoker      67.714305     Lower Lobe
3  Patient0003   72    Male  Current Smoker      70.806008     Lower Lobe
4  Patient0004   37  Female    Never Smoked      87.272433     Lower Lobe

        Stage          Treatment  Survival_Months         Ethnicity  ...  \
0  Stage III            Surgery              44          Hispanic  ...
1    Stage I  Radiation Therapy             101         Caucasian  ...
2    Stage I       Chemotherapy              69  African American  ...
3  Stage III       Chemotherapy              95  African American  ...
4   Stage IV  Radiation Therapy             105             Asian  ...

  Alanine_Aminotransferase_Level Aspartate_Aminotransferase_Level  \
0                      27.985571                        46.801214
1                      30.120956                        39.711531
2                       5.882418                        32.640602
3                      38.908154                        44.319393
4                      26.344877                        15.746906

  Creatinine_Level   LDH_Level Calcium_Level Phosphorus_Level Glucose_Level  \
0         1.245849  239.240255     10.366307         3.547734    113.919243
1         1.463231  233.515237     10.081731         2.945020    101.321578
2         0.630109  169.037460      8.660892         4.637399     78.214177
3         0.594342  213.967590      8.832669         3.617098    127.895361
4         1.478239  118.187543      9.247609         4.773255    148.801185

  Potassium_Level Sodium_Level  Smoking_Pack_Years
0        4.968163   139.822861           17.006956
1        3.896795   135.449361           93.270893
2        4.369050   143.377155           70.348376
3        4.348474   138.586005           19.828128
4        3.671976   141.230724           81.047456

[5 rows x 38 columns]
```

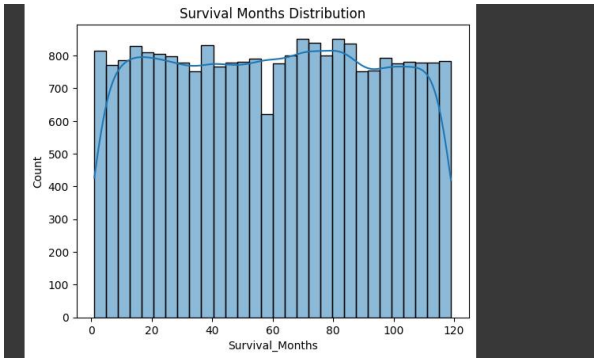
```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23658 entries, 0 to 23657
Data columns (total 38 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient_ID                      23658 non-null  object
 1   Age                             23658 non-null  int64
 2   Gender                          23658 non-null  object
 3   Smoking_History                 23658 non-null  object
 4   Tumor_Size_mm                   23658 non-null  float64
 5   Tumor_Location                  23658 non-null  object
 6   Stage                           23658 non-null  object
 7   Treatment                       23658 non-null  object
 8   Survival_Months                 23658 non-null  int64
 9   Ethnicity                       23658 non-null  object
 10  Insurance_Type                  23658 non-null  object
 11  Family_History                  23658 non-null  object
 12  Comorbidity_Diabetes            23658 non-null  object
 13  Comorbidity_Hypertension        23658 non-null  object
 14  Comorbidity_Heart_Disease       23658 non-null  object
 15  Comorbidity_Chronic_Lung_Disease  23658 non-null  object
 16  Comorbidity_Kidney_Disease      23658 non-null  object
 17  Comorbidity_Autoimmune_Disease  23658 non-null  object
 18  Comorbidity_Other               23658 non-null  object
 19  Performance_Status              23658 non-null  int64
 20  Blood_Pressure_Systolic         23658 non-null  int64
 21  Blood_Pressure_Diastolic        23658 non-null  int64
 22  Blood_Pressure_Pulse            23658 non-null  int64
 23  Hemoglobin_Level                23658 non-null  float64
 24  White_Blood_Cell_Count          23658 non-null  float64
 25  Platelet_Count                  23658 non-null  float64
 26  Albumin_Level                   23658 non-null  float64
 27  Alkaline_Phosphatase_Level      23658 non-null  float64
 28  Alanine_Aminotransferase_Level  23658 non-null  float64
 29  Aspartate_Aminotransferase_Level  23658 non-null  float64
 30  Creatinine_Level                23658 non-null  float64
 31  LDH_Level                       23658 non-null  float64
 32  Calcium_Level                   23658 non-null  float64
 33  Phosphorus_Level                23658 non-null  float64
 34  Glucose_Level                   23658 non-null  float64
 35  Potassium_Level                 23658 non-null  float64
 36  Sodium_Level                    23658 non-null  float64
 37  Smoking_Pack_Years              23658 non-null  float64
dtypes: float64(16), int64(6), object(16)
memory usage: 6.9+ MB
None
```

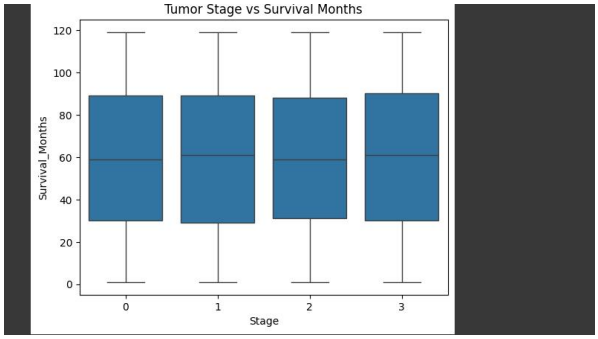
```
Missing Values:
Patient_ID                          0
Age                                 0
Gender                              0
Smoking_History                     0
Tumor_Size_mm                       0
Tumor_Location                      0
Stage                               0
Treatment                           0
Survival_Months                     0
Ethnicity                           0
Insurance_Type                      0
Family_History                      0
Comorbidity_Diabetes                0
Comorbidity_Hypertension            0
Comorbidity_Heart_Disease           0
Comorbidity_Chronic_Lung_Disease    0
Comorbidity_Kidney_Disease          0
Comorbidity_Autoimmune_Disease      0
Comorbidity_Other                   0
Performance_Status                  0
Blood_Pressure_Systolic             0
Blood_Pressure_Diastolic            0
Blood_Pressure_Pulse                0
Hemoglobin_Level                    0
White_Blood_Cell_Count              0
Platelet_Count                      0
Albumin_Level                       0
Alkaline_Phosphatase_Level          0
Alanine_Aminotransferase_Level      0
Aspartate_Aminotransferase_Level    0
Creatinine_Level                    0
LDH_Level                           0
Calcium_Level                       0
Phosphorus_Level                    0
Glucose_Level                       0
Potassium_Level                     0
Sodium_Level                        0
Smoking_Pack_Years                  0
dtype: int64
KNN Accuracy: 0.009509721048182587
Naive Bayes Accuracy: 0.0076077768385460695
SVM Accuracy: 0.0076077768385460695
Decision Tree Accuracy: 0.007185122569737954
Linear Regression R2 Score: -0.002134306163615651
```








Survival Months Distribution
Count
Survival_Months
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120


Tumor Stage vs Survival Months
Survival_Months
Stage
0
20
40
60
80
100
120
0
1
2
3