Lung Cancer Data Prediction (synopsis)

Rishik Pathak (CS22BCAGN003)
Priyanku Gogoi (CS22BCAGN028)
Amit Karmakar (CS22BCAGN006)
Pranjeet Gogoi (CS22BCAGN013)
Minmoy Khound (CS22BCAGN007)

Introduction

Lung Cancer is one of the deadliest diseases, and early detection can significantly improve survival rates. This project focuses on lung cancer data and to identify patterns and predict the likelihood of lung cancer diagnosis using machine learning models.

Objectives

- To analyze key risk factors influencing lung cancer diagnosis.
- To implement machine learning models for predicting lung cancer based on patient data.
- To compare model performances and identify the most effective approach.
- To provide data-driven insights into environmental, lifestyle, and genetic contributors to lung cancer.

Methodology

Data Collection & Preprocessing:

- The dataset undergoes thorough cleaning by handling missing values, removing irrelevant columns, and encoding categorical variables.
- Features such as age, smoking habits, pollution exposure, and occupational hazards are considered.

Exploratory Data Analysis (EDA)

Before applying machine learning, the dataset undergoes thorough cleaning, preprocessing, and analysis to gain meaningful insights. The dataset is processed by removing irrelevant columns, handling missing values, and encoding categorical variables. Key trends and patterns are visualized through various charts:

1. **Line Chart:** Shows how annual lung cancer deaths vary across different age groups, helping identify high-risk age ranges.
2. **Bar Chart:** Highlights the top 10 countries with the highest number of lung cancer cases, offering a geographical perspective on disease prevalence.
3. **Pie Chart:** Illustrates the proportion of lung cancer diagnoses attributed to air pollution exposure, revealing environmental risk factors.
4. **Histogram:** Analyzes the distribution of smokers and non-smokers in lung cancer diagnosis, giving insights into the overall frequency of cases.

These visualizations help identify key contributors to lung cancer and guide the selection of features for machine learning models.

Machine Learning Models for Prediction

To predict lung cancer diagnoses, **multiple classification models** are trained and evaluated. Each model has unique strengths and helps compare different approaches to identifying at-risk patients:

1. **Logistic Regression –** A simple, interpretable baseline model that provides probability-based predictions.
2. **K-Nearest Neighbors (KNN) –** Classifies patients by comparing them to similar cases in the dataset.
3. **Support Vector Machine (SVM) –** Finds the optimal boundary between lung cancer-positive and negative cases using mathematical techniques.
4. **Decision Tree –** Uses a tree-like structure to split the dataset into meaningful decision rules.
5. **Random Forest –** An ensemble model that combines multiple decision trees to improve accuracy and robustness.
6. **Naive Bayes** - A probabilistic classifier based on Bayes' theorem, assuming feature independence.

Model Evaluation

Each model is tested using **accuracy scores**, **confusion matrices**, and **classification reports** to assess performance. The Logistic Regression and Support Vector Machine achieved the highest accuracy.

Results

The analysis reveals key factors influencing lung cancer diagnoses, such as **smoking habits**, **air pollution exposure**, **occupational hazards**, and **age**.

The models show that incorporating multiple risk factors significantly improves predictive accuracy.

The findings highlight the importance of environmental awareness, smoking cessation programs, and early screening.

Challenges & Limitations

- Data imbalance may affect model performance, requiring techniques such as resampling.
- The dataset does not fully capture real-time genetic predispositions or lifestyle changes.
- Some models require hyperparameter tuning for optimal performance.

Conclusion

This project provides a data-driven approach to understanding lung cancer risks and improving early detection. By leveraging machine learning, we can enhance diagnostic accuracy, support healthcare professionals, and contribute to preventive strategies. With further improvements, such as larger datasets and deep learning techniques, this approach could become even more effective in real-world medical applications.