PROJECT: 4

Lung Cancer Prediction

Name: Mani P

Email: maniprabu991@gmail.com

Program: Unified Mentor ML Internship Program

1. Problem Statement

Lung cancer is one of the most fatal cancers worldwide, and early prediction of patient survival plays a critical role in treatment planning and decision-making. Traditional clinical analysis methods are time-consuming and depend heavily on manual interpretation. This project addresses the need for an automated, data-driven system that predicts lung cancer survival using patient diagnostic, lifestyle, and treatment data.

2. Objective

The primary objective of this project is to build a machine learning model that can predict whether a lung cancer patient is likely to survive, based on clinical, demographic, and lifestyle features.

3. Tech Stack Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- **Machine Learning Algorithm:** Random Forest Classifier
- **Model Optimization:** GridSearchCV
- **Domain:** Healthcare / Medical Analytics

4. Project Architecture / Workflow

1. Data collection
2. Exploratory Data Analysis (EDA)

3. Data preprocessing and cleaning
4. Feature encoding and preparation
5. Model training using Random Forest
6. Hyperparameter tuning
7. Model evaluation
8. Lung cancer survival prediction

5. Implementation Details

- Dataset includes patient details such as age, gender, cancer stage, smoking status, BMI, cholesterol level, comorbidities, and treatment type
- Missing values handled appropriately
- Categorical features encoded for model compatibility
- Random Forest Classifier selected for its robustness and interpretability
- Hyperparameter tuning performed to optimize recall and accuracy
- Feature importance analyzed to understand key survival factors

6. Output / Results

Model Evaluation Metrics:

- **Accuracy:** 92.33%
- **Precision:** 93.07%
- **Recall:** 93.00%
- **F1-Score:** 93.00%

Sample Prediction:
Patient Details:

- Age: 55
- Cancer Stage: II
- Smoking Status: Smoker
- Treatment Type: Non-surgical

Prediction: Lung Cancer (High Risk)
Confidence: 92.33%

Top Important Features:

1. Age
2. Cancer Stage
3. Smoking Status
4. Surgical Treatment

7. Challenges Faced

- Handling mixed data types (numerical and categorical)
- Selecting appropriate metrics for medical prediction

- Ensuring minimal false negatives in survival prediction

8. Future Enhancements

- Integration with clinical decision support systems
- Applying advanced models like XGBoost or LightGBM
- Using SHAP for better interpretability
- Deploying the model as a web application using Streamlit or Flask

9. Conclusion

This project demonstrates an end-to-end machine learning approach to lung cancer survival prediction. By leveraging Random Forest and feature importance analysis, the model achieved high accuracy and recall, making it suitable as a decision-support tool in healthcare environments.