Cancer Prediction Using Machine Learning

End-to-End ML Project Report

1. Problem Statement

Early detection of lung cancer is crucial for patient survival. Manual diagnosis methods are often time-consuming and resource-intensive. The objective is to build a machine learning model that predicts lung cancer based on patient clinical and lifestyle data.

2. Objective

• Predict whether a patient is likely to have lung cancer. • Identify the most important features affecting survival. • Provide a reliable model that can be deployed for real-time predictions.

3. Solution & Exploratory Data Analysis (EDA)

The dataset contains features like age, cancer stage, smoking status, treatment type, and other comorbidities. EDA revealed that age, cancer stage, smoking, and surgery have the highest correlation with survival.

4. Algorithm Used

Random Forest Classifier was used due to its robustness, ability to handle non-linear relationships, and interpretability via feature importance. Hyperparameters were tuned for better recall and accuracy.

5. Model Evaluation

The model was evaluated using accuracy, precision, recall, and F1-score. Special focus was given to recall to minimize false negatives in cancer prediction.

Metric	Value
Accuracy	0.9233
Precision	0.9307
Recall	0.9300
F1-Score	0.9300

6. Sample Prediction

For a patient with the following attributes: • Age: 55 • Cancer Stage: 2 • Smoker: Yes • Surgical Treatment: No The model predicted: **Cancer** with a confidence of 92.33%

7. Feature Importance

The model identified the following features as most important: 1. Age 2. Cancer Stage 3. Smoking Status 4. Surgical Treatment

8. Conclusion

This project demonstrates an end-to-end ML pipeline for lung cancer prediction, from data preprocessing and exploratory analysis to model building, evaluation, and sample prediction. The high confidence of 92.33% indicates the model is reliable for decision-support in clinical scenarios.