Respiratory Cancer Recidivism Prediction

**Executive Summary:**

This capstone project aims to build a predictive model using machine learning to forecast 5-year survival among patients already diagnosed with respiratory cancer, based on demographic, clinical, and pathological data from the SEER dataset. The goal is to develop an efficient, interpretable, and reliable decision support system that can aid in early detection or prognosis prediction of respiratory cancers. The project involves data preprocessing, exploratory data analysis (EDA), feature engineering, model development, evaluation. Various supervised learning algorithms such as logistic regression, decision trees, random forests, and gradient boosting, will be tested. Model performance will be evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.

**Problem Background:**

Respiratory cancers, including lung and bronchial cancers, are among the leading causes of cancer-related deaths worldwide. Early detection and accurate prognosis are critical for improving patient outcomes. However, due to the complex nature of cancer progression and limitations in current diagnostic practices, predicting respiratory cancer outcomes remains a significant challenge. The integration of machine learning (ML) in medical diagnostics offers new potential for early and accurate prediction. By leveraging large-scale datasets such as SEER (Surveillance, Epidemiology, and End Results), ML models can identify hidden patterns and make reliable predictions, which can assist clinicians in decision-making.

**Objectives:**

**Primary Objective:**

To develop a machine learning model capable of predicting 5-year survival outcomes for patients diagnosed with respiratory cancer using the SEER dataset.

**Secondary Objectives:**

1. Perform data cleaning and preprocessing on SEER dataset entries relevant to respiratory cancers, ensuring that the data is ready for modelling.
2. Conduct exploratory data analysis (EDA) to uncover patterns, correlations, and potential risk factors for survival.
3. Evaluate multiple machine learning models, such as logistic regression, decision trees, random forests, and gradient boosting, to identify the most effective model for predicting 5-year survival.
4. Assess model performance using appropriate statistical metrics, including accuracy, precision, recall, F1-score, and AUC-ROC.
5. Develop a simple visualization or prediction dashboard for healthcare professionals to interpret model predictions and aid in clinical decision-making.

**Data Source: SEER Dataset**

The SEER (Surveillance, Epidemiology, and End Results) program of the National Cancer Institute provides comprehensive information on cancer statistics in the United States. For this project, data relevant to respiratory system cancers (e.g., lung and bronchial cancers) will be extracted. This includes patient demographics, tumour characteristics, survival time, treatment information, and vital status.

**Assumptions / Scope:**

This analysis focuses on patients diagnosed with respiratory cancers using data from a cancer registry. The dataset is assumed to accurately reflect patient characteristics at diagnosis and follow-up.

* Missing Data Handling: Missing values (~50%) will be carefully examined for each variable to understand the nature of the missingness. We will conduct missing data analysis to determine whether the data is missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). Based on this analysis, appropriate methods for imputation (e.g., mean, median, mode, or advanced techniques such as multiple imputation) will be applied.
* Scope of Study: The study includes only available variables. External factors, such as environment, lifestyle, or genetics, are not considered due to the absence of relevant data in the SEER dataset. Survival and recurrence predictions reflect historical patterns and do not account for future treatment or policy changes.

**Method:**  
This analysis predicts survival and recurrence of respiratory cancer using clinical, demographic, and treatment-related features. Following the CRISP-DM framework, it applies data preprocessing, feature engineering, and machine learning models for prediction.

**Data Cleaning and Preprocessing**

We address ~50% missing data using imputation within streamlined preprocessing pipelines (piping) in Python using Pandas and Scikit-learn. This ensures consistency and repeatability.

* **Numerical Variables**: Impute missing values using mean, median, or KNN-based methods.
* **Categorical Variables**: Impute using mode; apply Multiple Imputation (e.g., MICE, Iterative Imputer) for complex missing patterns.
* **Piping**: Chain imputation, encoding, and scaling tasks to maintain a clean and structured process.

**Dependent and Independent Variables**

* **Dependent Variables**:
  + *Survival Status*: Binary indicator of survival within a specified timeframe (e.g., 1 or 5 years).
  + *Recurrence*: Binary indicator of cancer recurrence after treatment.
* **Independent Variables**:
  + *Demographics*: Age, sex, race, origin.
  + *Tumor Characteristics*: Tumor size, grade, AJCC staging, histology codes.
  + *Treatment*: Surgery type, chemotherapy, radiation therapy.
  + *Follow-up*: Survival time, recurrence, status.
  + *Confounders*: Socioeconomic status, comorbidities.

**CRISP-DM Framework:**

**This analysis follows the CRISP-DM framework to guide predictive modeling.**

1. Business Understanding: We define the goal as predicting survival (5-year survival) in respiratory cancer patients.
2. Data Understanding: We begin by exploring variable types, distributions, and identifying patterns of missing data. Variables with more than 50% missingness will be carefully considered. If a variable is crucial for prediction, we may apply imputation techniques or retain it with reduced information. If a variable is deemed less critical, or if imputation is not feasible, we may choose to drop it from the dataset.
3. Data Preparation: We clean the data by addressing missing values, scaling numerical variables, and engineering relevant features, such as risk scores. We will use imputation techniques based on the nature of missingness (as determined by missing data analysis). Pipelines will be employed to streamline preprocessing tasks and ensure consistency.
4. Modelling: We will train multiple machines learning models, including Logistic Regression, Random Forest, Gradient Boosting Machines (GBM), and Support Vector Machines (SVM). To improve model robustness, we will use 5-fold cross-validation instead of a simple 70-30 split. This will allow for a more reliable evaluation of model performance by testing the models on different subsets of the data.
5. Evaluation: We will evaluate the models using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. Cross-validation results will be used to reduce overfitting and ensure that the models generalize well to unseen data.

**Constraints and Limitations**

* **Imbalanced Data**: Underrepresented outcomes may affect model performance; apply techniques like SMOTE.
* **Missing Data Bias**: Imputation may not fully eliminate bias if data is not MAR or MCAR.
* **Model Interpretability**: Complex models (e.g., GBM, RF) may lack transparency compared to Logistic Regression.

**Steps in the Analysis**

1. Clean and preprocess data using pipelines.
2. Engineer features including risk indicators.
3. Train multiple classification models.
4. Evaluate and compare model performance using standard metrics.

**Results**

We are in the process of developing predictive models—Logistic Regression, Random Forest, Gradient Boosting, and SVM—to estimate survival and recurrence in respiratory cancer patients. Once developed, each model will be evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. We plan to handle missing data through robust imputation to ensure reliable inputs for training.

The final output will include interactive visualizations like ROC curves, confusion matrices, and survival plots to aid in model interpretation and highlight key predictive factors. These models aim to provide data-driven insights that will assist healthcare professionals in identifying high-risk patients and informing treatment strategies more effectively.

**Recommendation**  
Healthcare providers could integrate these predictive models into clinical workflows to identify high-risk respiratory cancer patients early. By flagging individuals with a higher likelihood of recurrence or reduced survival, clinicians can prioritize follow-ups, personalize treatment plans, and allocate resources more effectively. The models also offer explainability for key risk factors, supporting informed, data-driven decisions.

**Risks**  
The analysis relies on historical data, which may not fully reflect future clinical practices or advancements in treatment. Missing data, even after imputation, may introduce biases. Additionally, imbalanced outcomes (e.g., low recurrence rates) can affect model accuracy and lead to over- or under-prediction for certain patient groups.

**Next Steps**  
Validate the models on a larger and more recent dataset to ensure generalizability. Explore integration with Electronic Health Record (EHR) systems for real-time predictions. To address class imbalance, consider advanced resampling techniques or cost-sensitive algorithms. Further, engage clinicians to interpret model outputs and refine the model features for improved practical relevance.