Lung Cancer Prediction (lung2.ipynb)

- **Problem Statement**: The project aims to predict whether a patient has lung cancer (LUNG_CANCER = YES/NO). This is a binary classification problem.
- **Dataset Used**
  :
  - The dataset was loaded from survey lung cancer.csv.
  - The target variable is 'LUNG_CANCER'.
  - Features include 'GENDER', 'AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY', 'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING', 'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH', 'SWALLOWING DIFFICULTY', 'CHEST PAIN'.

- **Methodology and Approach**
  :
  - **Data Preprocessing**
    :
    - Categorical features ('GENDER', 'LUNG_CANCER', and all other symptom-based features which are initially 1/2 for NO/YES) were encoded: 'GENDER' to 0/1 and other features (including 'LUNG_CANCER') to 0/1 (originally 1 mapped to 0 for NO, 2 mapped to 1 for YES).
    - Numerical features were scaled using StandardScaler.

  - **Model Training**
    :
    - The data was split into training (70%) and testing (30%) sets.
    - The following classification models were implemented:
      - Logistic Regression (LogisticRegression)
      - K-Nearest Neighbors (KNeighborsClassifier)
      - Support Vector Classifier (SVC)
      - Decision Tree Classifier (DecisionTreeClassifier with criterion='entropy')
      - Random Forest Classifier (RandomForestClassifier with criterion='entropy')

- **Model Evaluation**
  :
  - Models were evaluated using accuracy scores and confusion matrices.

- **Results and Conclusion**
  :
  - Logistic Regression: Accuracy 93.55%.
  - K-Nearest Neighbors (k=1): Accuracy 91.40%.
  - Support Vector Classifier: Accuracy 92.47%.
  - Decision Tree: Accuracy 91.40%.
  - Random Forest: Accuracy 91.40%.
  - Logistic Regression provided the highest accuracy.
  - The Logistic Regression model was saved to lung_cancer_model.pkl.