

**Project Title:**

Lung Cancer Prediction using Machine Learning

**Team Details:**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Name** | **Roll No** |
| 1 | B Jaipreeth | 23R11A0506 |
| 2 | Aryan M | 23R11A0504 |
| 3 | K. T. Sri Charan | 23R11A0525 |

Abstract:

Lung cancer is a leading cause of cancer-related deaths globally, highlighting the need for early detection methods. This project develops a machine learning-based system to predict lung cancer risk using patient data, including demographics, lifestyle factors (e.g., smoking), and clinical symptoms. After preprocessing the dataset (handling missing values, encoding categorical variables, and feature scaling), we evaluated multiple algorithms like Logistic Regression, Decision Trees, Random Forest, and XGBoost.

The models were assessed using accuracy, precision, recall, and cross-validation techniques. Results showed that ensemble methods, particularly Random Forest and XGBoost, achieved the highest predictive performance. These models could assist healthcare professionals in early diagnosis and risk assessment. Future work may involve larger datasets and clinical implementation.

**1. Introduction**  
**1.1.1 Problem Statement**Lung cancer, a severe and aggressive form of cancer, ranks among the leading causes of cancer-related deaths globally. According to WHO and cancer research data, lung cancer accounts for over 1.8 million deaths annually. One of the main challenges is the difficulty in diagnosing lung cancer during its early stages due to non-specific symptoms such as coughing, chest pain, and fatigue—often confused with other respiratory illnesses. Delayed diagnosis typically results in treatment beginning at an advanced stage, reducing the chances of successful recovery.

Traditional diagnostic tools like CT scans, MRIs, and biopsies are often costly, time-intensive, and not always accessible, particularly in under-resourced areas. Thus, there exists an urgent need for an intelligent, efficient, and affordable system that can support clinicians by providing early risk analysis for lung cancer using readily available patient data.

**1.1.2 Objective of the Project**  
The primary objective of this project is to develop a Machine Learning (ML)-based prediction model that can predict the likelihood of lung cancer in individuals based on their medical and behavioral data. Key goals include:

* The primary objective of this project is to develop a Machine Learning (ML)-based prediction model that can predict the likelihood of lung cancer in individuals based on their medical and behavioral data. Key goals include:
* To apply suitable ML algorithms that offer high accuracy and efficiency in predictions.
* To develop a web-based tool that is intuitive and accessible to both medical professionals and the general public.
* To reduce the time, cost, and invasiveness of traditional diagnostic procedures through a data-driven approach.

**1.1.3 Need for Machine Learning**In today’s data-driven era, Machine Learning is essential for managing and interpreting vast and complex datasets in the medical domain. ML can detect subtle patterns and relationships within patient data that may not be apparent to human clinicians. Specifically, ML enables:

* Continuous learning from new medical data.
* Automation of diagnostic processes.
* Reduction in manual errors.
* Integration of diverse variables like symptoms, behaviors, and demographic factors.

ML allows healthcare systems to shift from reactive treatments to proactive preventive care, ultimately saving lives and reducing medical expenses.

**1.1.4 Introduction to Machine Learning**Machine Learning is a field of artificial intelligence that focuses on developing algorithms that enable computers to learn from and make decisions based on data. In contrast to rule-based systems, ML models improve automatically through experience.

In this project, we use supervised learning techniques, particularly Logistic Regression, to perform binary classification—determining whether a patient is likely to develop lung cancer or not.

* + 1. **Types of Machine Learning Techniques**
* **Supervised Learning**: Involves labeled data. The algorithm learns to map inputs to known outputs.
* es: Logistic Regression, Decision Trees, SVM
* **Unsupervised Learning**: Works with unlabeled data to identify patterns.
* Examples: K-means clustering, PCA
* **Semi-supervised Learning**: Combines small amounts of labeled data with large volumes of unlabeled data.
* **Reinforcement Learning**: The model learns through rewards and penalties by interacting with its environment.

**1.1.6 Machine Learning Steps**

1. **Data Collection**: Gathering relevant data from reliable sources.
2. **Data Preprocessing:** Cleaning and preparing data (handling missing values, encoding).
3. **Feature Selection:** Identifying important variables.
4. **Model Selection:** Choosing appropriate ML algorithms.
5. **Training and Testing:** Splitting the dataset and training the model.
6. **Evaluation:** Measuring model accuracy and performance.
7. **Deployment:** Integrating the model into a usable application.

**1.2 Drawbacks of Existing System**

* Traditional methods rely heavily on physical scans and lab tests.
* They often require specialized equipment and trained radiologists.
* Delays in appointments and results can be life-threatening.
* There's limited scalability in rural or underprivileged regions.
* Human error or fatigue can affect diagnosis.

**1.3 Proposed System**  
The proposed system leverages machine learning to predict the likelihood of lung cancer based on user inputs such as age, gender, smoking history, and common symptoms. It offers:

* A web-based interface for user input.
* Backend processing using Logistic Regression.
* Real-time prediction with analysis.
* Cost-effective and accessible diagnosis support.
* Use Logistic Regression to classify individuals based on risk probability.
* Offer a user interface where users can input symptoms.
* Display predictive outcomes along with confidence levels.
* Enable easy deployment across hospital intranets or mobile devices.

Advantages:

* Fast and accurate predictions.
* Cost-effective as no expensive tests are needed initially.
* Scalable and user-friendly.

**2. Literature Survey**  
Several research works have focused on using ML for cancer prediction. Studies show that algorithms like Decision Trees, Support Vector Machines (SVM), and Neural Networks have been used for classification tasks. However, Logistic Regression has proven effective due to its simplicity, interpretability, and strong performance on structured medical datasets.  
In our review, we identified that many existing models either lacked transparency or were too complex for practical use. Therefore, our approach emphasizes interpretability, speed, and ease of use.

* **Study 1**: SVM and Random Forests were applied to lung cancer datasets achieving up to 85% accuracy.
* **Study 2**: Neural Networks used CT image analysis but required high computational power.
* **Study 3**: Logistic Regression with clinical datasets showed promising results with interpretability and low training time.

Findings:

* Logistic Regression remains a preferred choice for structured medical data.
* Ensemble methods offer better performance but increase complexity.
* Data quality and feature engineering play critical roles.

**3. Feasibility Study**  
**3.1 Feasibility Study Definition**Feasibility study assesses whether a proposed project is practical and viable considering technical, operational, and economic aspects.

**3.2 Types of Feasibility**

* Technical Feasibility:

1. The project uses Python, a mature programming language with powerful ML libraries.
2. Logistic Regression and visualization tools like Matplotlib and Seaborn are easily implemented.

* Operational Feasibility:

1. The system can be operated by individuals with minimal technical knowledge.
2. Simple web interface ensures user-friendly interaction.

* Economic Feasibility:

1. Uses open-source technologies, minimizing cost.
2. Can be deployed on low-cost servers or desktops.

* Legal Feasibility:

1. Designed for academic and predictive purposes.
2. No personal or sensitive health data is stored or processed in a real-world scenario.

**Software Requirements Specification (SRS)**  
**4.1 Annexure**

* Sample input data
* Screenshots of prediction interface
* Performance evaluation metrics

**4.2 Data Flow Diagram (DFD)**

* **Input Layer:** User enters personal and symptom details.
* **Processing Layer:** Data is preprocessed and passed to the ML model.
* **Output Layer:** Prediction result (Positive/Negative) is displayed.

**4.3 Process Description**

* Data is collected using form inputs.
* Preprocessing includes encoding categorical data and checking for null values.
* Logistic Regression is applied.
* Result is displayed with visual feedback.

**4.4 Data Dictionary**

|  |  |  |
| --- | --- | --- |
| Attribute | Type | Description |
| Age | Integer | Age of the patient |
| Gender | Categorical | Male or Female |
| Smoking | Boolean | Whether the person smokes |
| Fatigue | Boolean | Tiredness level |
| Wheezing | Boolean | Sound during breathing |
| Shortness of Breath | Boolean | Difficulty in breathing |
| Chest Pain | Boolean | Pain or discomfort in chest |

**4.5 SSAD or OOAD**

SSAD (Structured System Analysis and Design) is used due to the linear data processing and straightforward prediction flow.

**4.6 System Requirements**  
**Hardware Requirements:**

* Intel i3 or higher
* Minimum 4GB RAM
* 500MB disk space

**Software Requirements:**

* Python 3.x
* Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn
* Jupyter Notebook or any Python IDE

**5. Form Designing**  
The user interface (UI) consists of a web form with fields for entering:

* Age
* Gender
* Smoking history
* Symptom checklist (Fatigue, Chest Pain, etc.)
* A “Predict” button triggers the backend model for analysis.
* Output is displayed on the same page for quick diagnosis feedback.

**6. Data Dictionary**

|  |  |  |
| --- | --- | --- |
| Field Name | Description | Data Type |
| age | Age of the patient | Integer |
| gender | Male / Female | Category |
| smoking | Yes / No | Boolean |
| fatigue | Yes / No | Boolean |
| chest\_pain | Present / Absent | Boolean |
| wheezing | Present / Absent | Boolean |
| shortness\_breath | Yes / No | Boolean |

**7. Algorithms Used in Lung Cancer Prediction Using ML**

**7.1 Introduction to Machine Learning Algorithms**

* Logistic Regression: A statistical model that uses a logistic function to model

1. Suitable for binary classification problems like disease prediction.
2. Calculates the probability that a given input belongs to a specific category.
3. Formula: Sigmoid function (1 / (1 + e^-z))

Advantages:

* Interpretable coefficients
* Fast training
* Requires less computational power
* binary dependent variables. It is simple yet powerful for medical classification problems.
* Chosen for this project due to its interpretability, fast training time, and suitability for binary classification.

**7.2 Introduction to Libraries Used in Python**

* **pandas**: For loading and manipulating data.
* **numpy:** For mathematical operations.
* **scikit-learn:** ML models, data preprocessing, and evaluation.
* **matplotlib/seaborn:** For data visualization.
* **LabelEncoder & OneHotEncoder:** To convert categorical variables into numeric.

**8.Data Description:**

The dataset used in this lung cancer prediction project contains anonymized medical and lifestyle-related information about individuals, alongside a binary classification indicating the presence or absence of lung cancer. The data includes the following key attributes:

* **GENDER**: Male or Female
* **AGE**: Age of the person
* **SMOKING**: Whether the person smokes
* **YELLOW\_FINGERS**: Presence of yellow stains (often from smoking)
* **ANXIETY**, **PEER\_PRESSURE**, **CHRONIC\_DISEASE**, **FATIGUE**: Psychological and health-related factors
* **ALLERGY**, **WHEEZING**, **COUGHING**, **SHORTNESS\_OF\_BREATH**, **SWALLOWING\_DIFFICULTY**, **CHEST\_PAIN**: Symptoms indicative of respiratory distress
* **ALCOHOL\_CONSUMING**: Lifestyle factor
* **LUNG\_CANCER**: Target variable (Yes/No – whether the patient is diagnosed with lung cancer)

**Initial Processing:**

* Dataset shape is inspected.
* Null values are identified and dropped using dropna().
* Duplicate entries are removed to ensure model accuracy.
* Categorical values are encoded using LabelEncoder and OneHotEncoder for compatibility with machine learning algorithms.
* The target variable (LUNG\_CANCER) distribution is visualized using seaborn countplots, which shows the balance between positive and negative samples.

**9.Methodology**

The machine learning pipeline follows these steps:

**Preprocessing:**

* Cleaning: Dropping null values and duplicates.
* Label Encoding: Transforming categorical string labels into numeric values.
* One-Hot Encoding: Applied where needed to ensure the model interprets categorical variables appropriately.
* Feature Selection: Identifying the most relevant features based on medical relevance and dataset availability.

**Splitting Dataset:**

* The dataset is split into **training** and **testing** sets using train\_test\_split(), allowing model validation and performance evaluation on unseen data.

**Model Building:**

Two primary machine learning algorithms are applied:

**1. Logistic Regression (Sigmoid-based):**

* A classical linear model that estimates the probability of a binary response based on independent variables.
* Used for its interpretability and efficiency in binary classification problems like this one.

**2. Support Vector Machine (SVM):**

* A powerful supervised learning algorithm that works well with high-dimensional spaces.
* Suitable for binary classification by finding the hyperplane that best separates the two classes.
* Kernel trick may be used implicitly to deal with non-linear data separation.

**10.** **Results and Analysis**

**Model Evaluation Metrics:**

* **Accuracy Score**: Measures the percentage of correct predictions on the test set.
* **Confusion Matrix**: Helps in understanding True Positives, False Positives, True Negatives, and False Negatives.
* **Precision, Recall, F1-Score**: Computed using sklearn.metrics to provide a detailed understanding of model performance.

**Visualization:**

* **Seaborn Heatmap**: Used to visualize the confusion matrix.
* **Bar Graphs/Count Plots**: Depict target distribution and symptom prevalence across diagnosed individuals.

**Performance Comparison:**

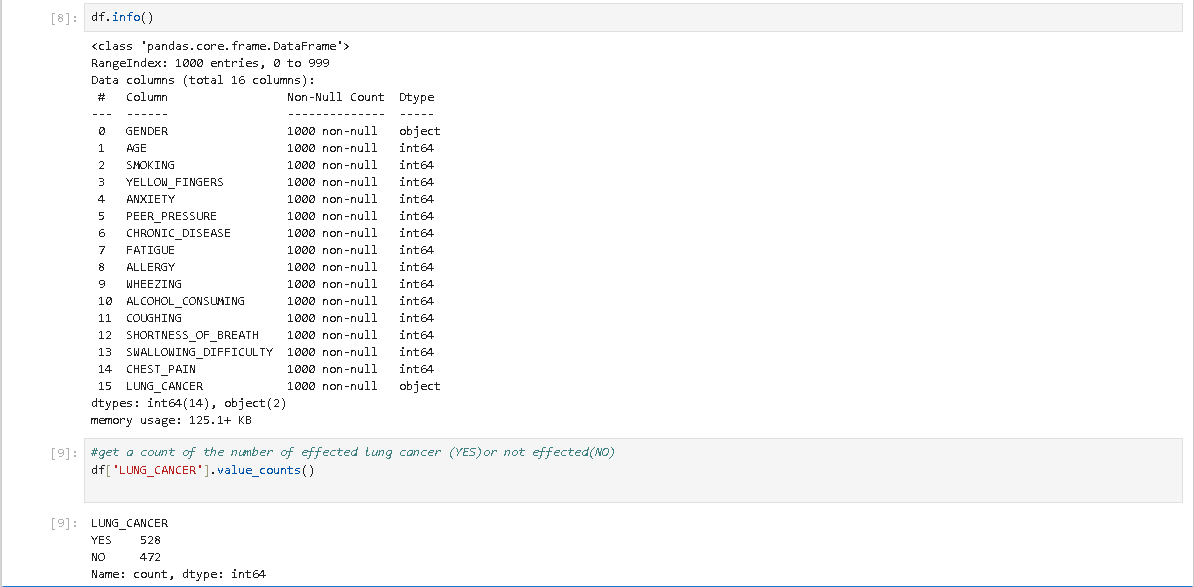
* Logistic Regression is expected to perform well in datasets with linearly separable data. It gives insights into the probability of lung cancer presence.
* SVM generally shows superior performance in handling higher dimensional and potentially non-linear data, offering robust classification when properly tuned.

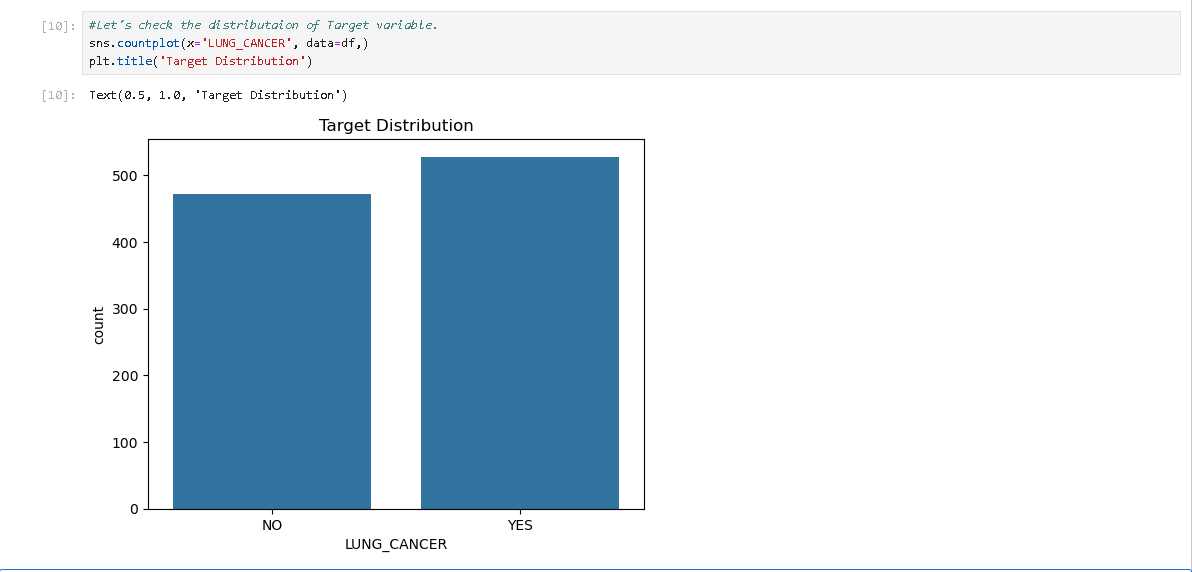
**CODE AND OUTPUT:**

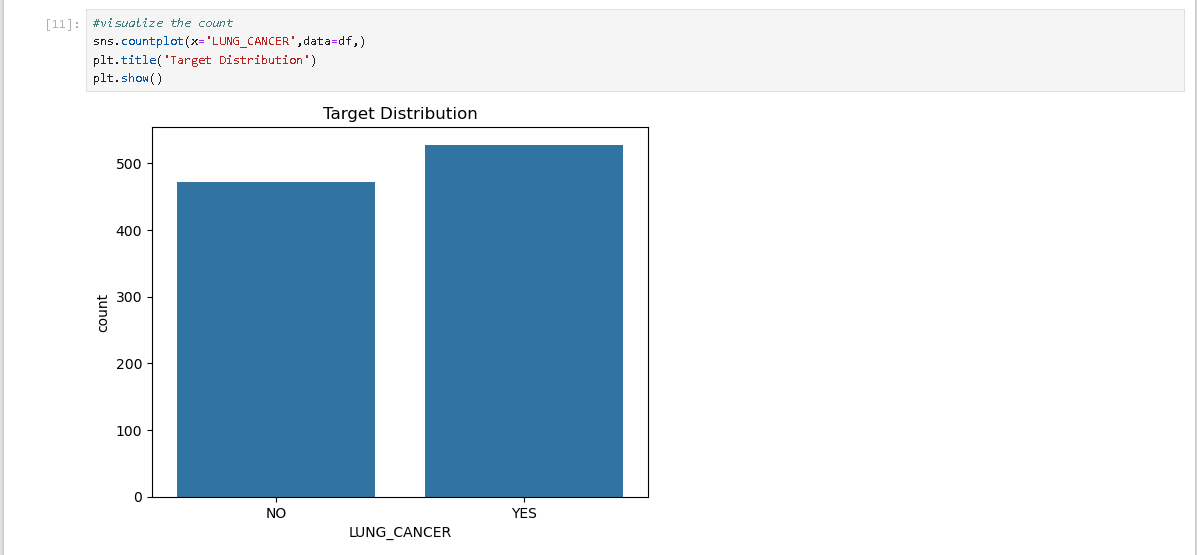
Jupyter source file:

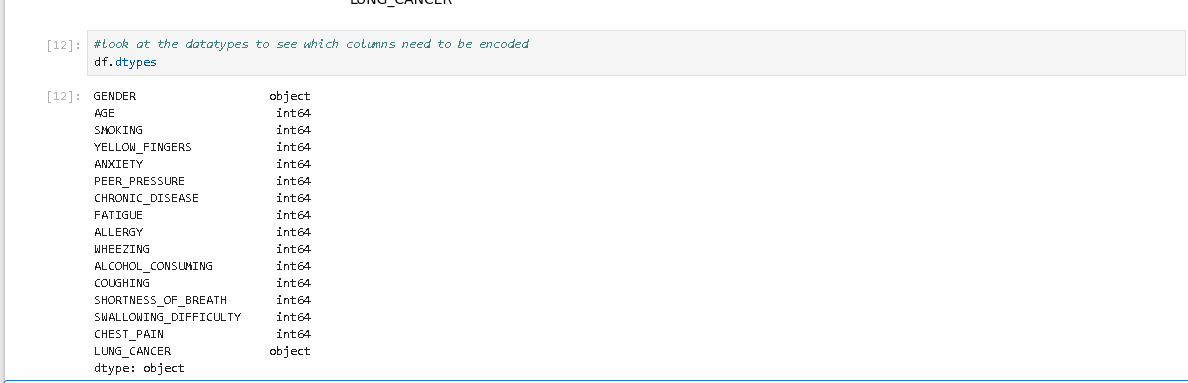


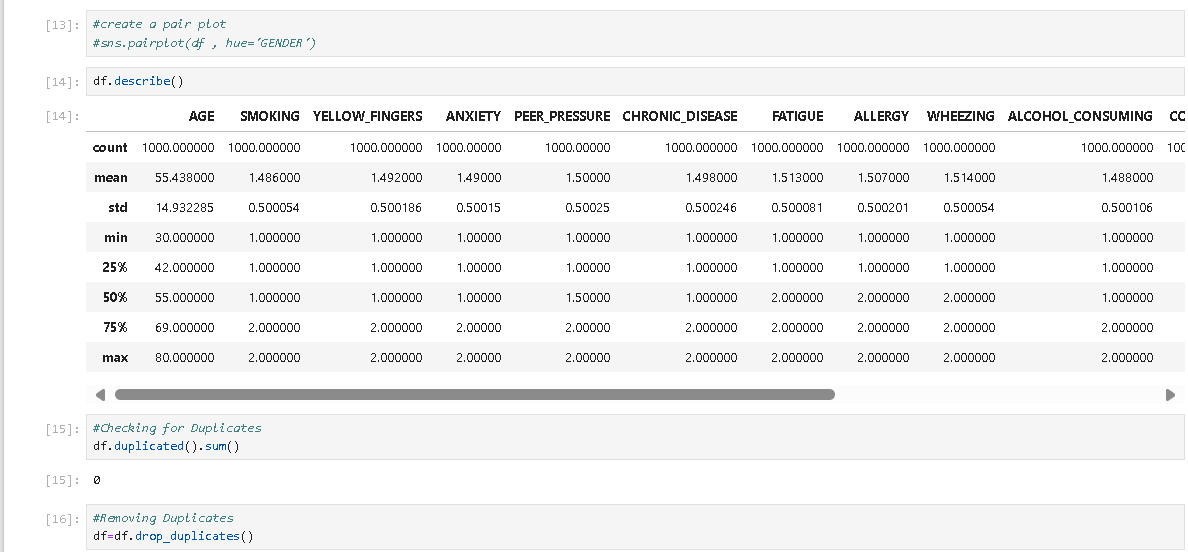


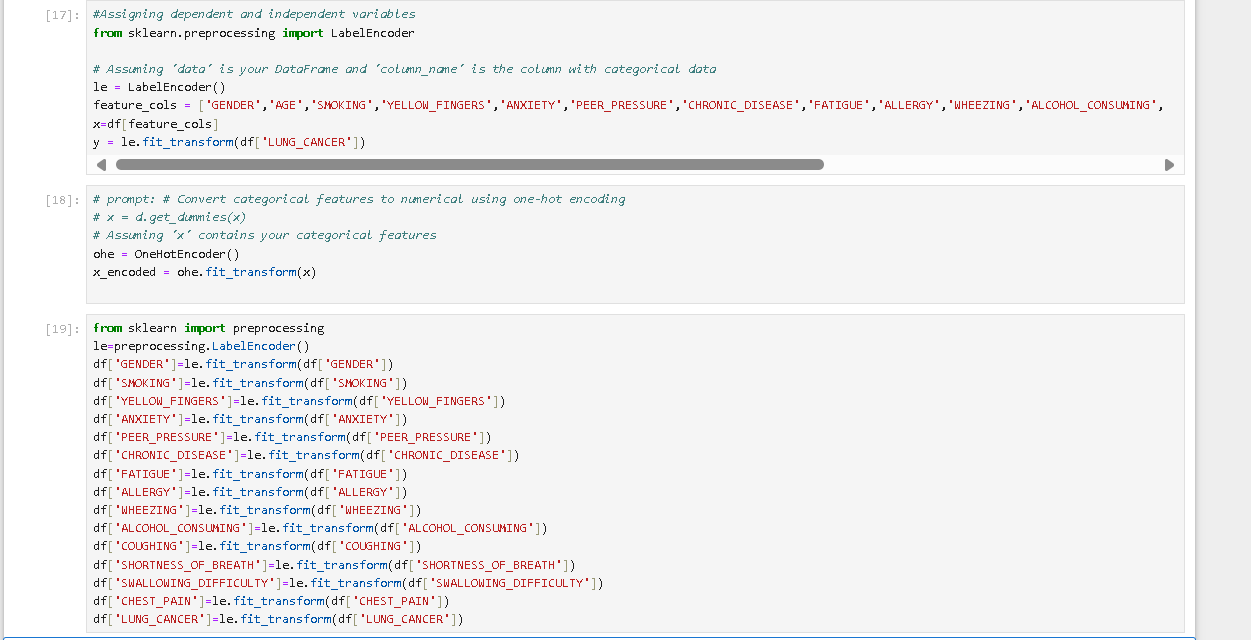
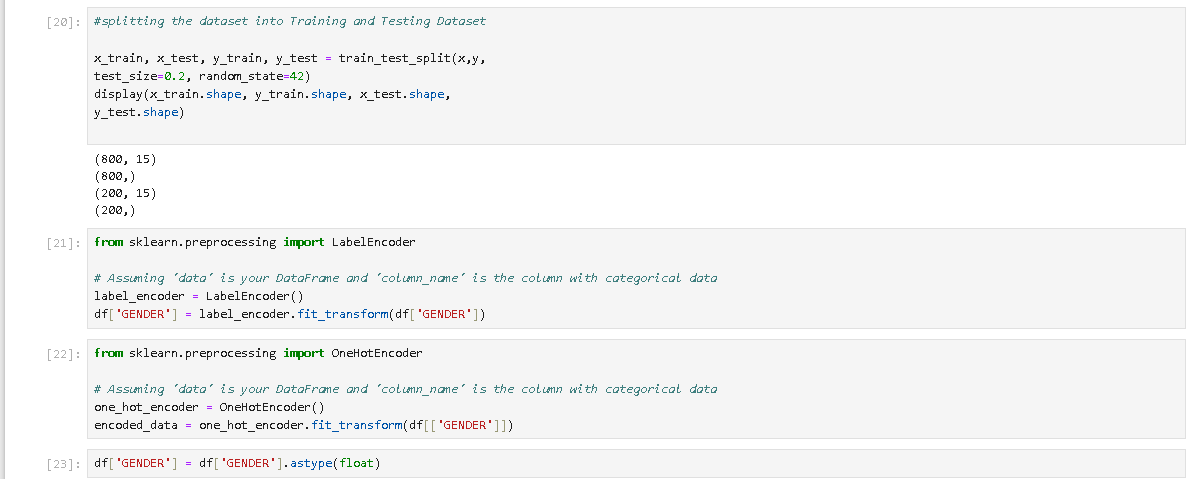
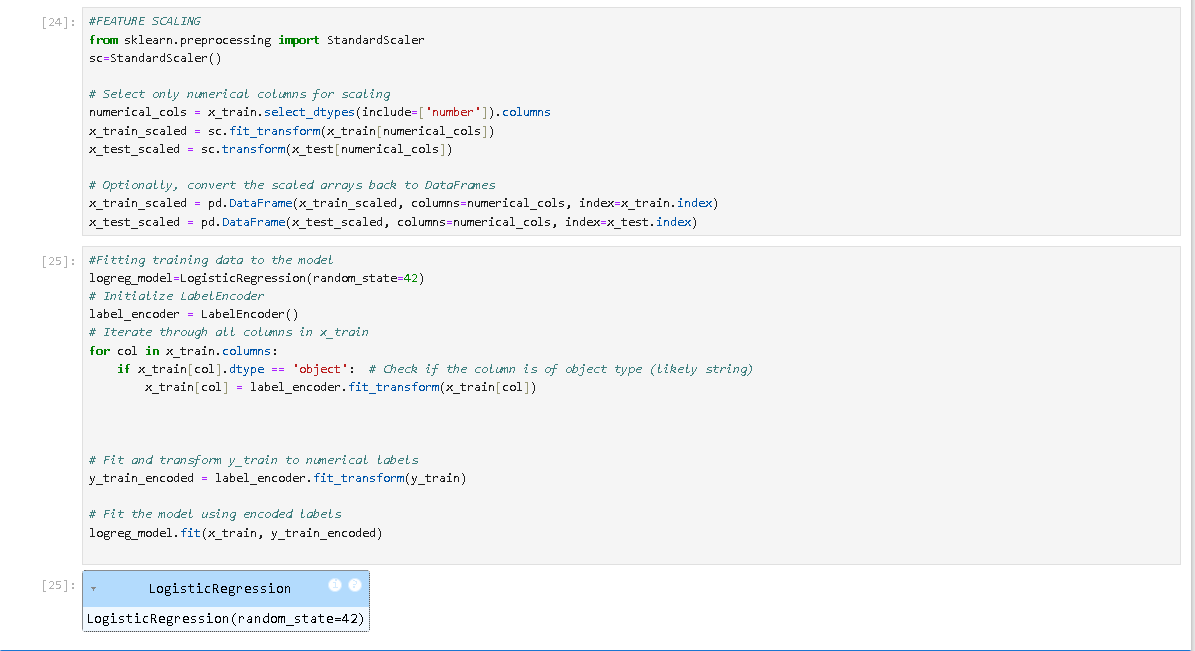
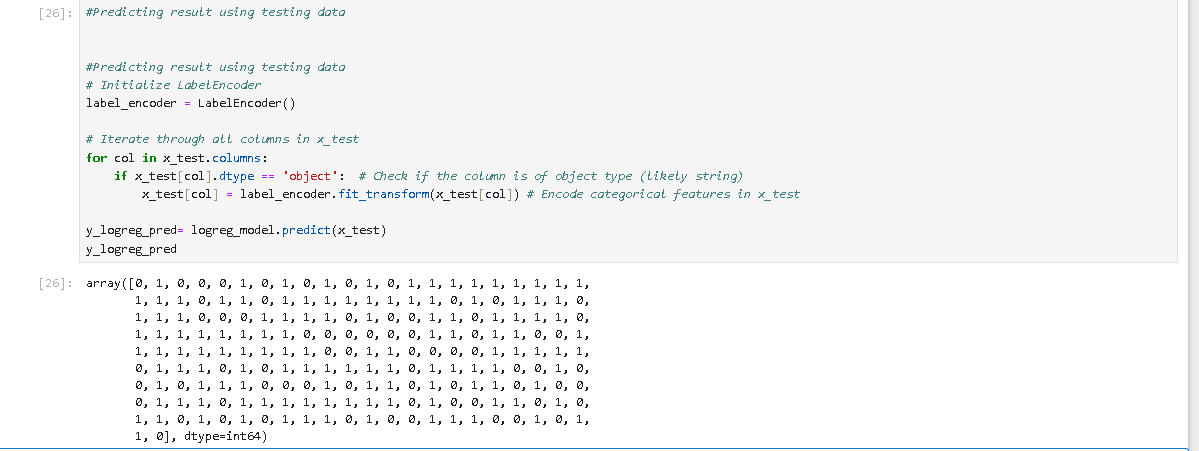
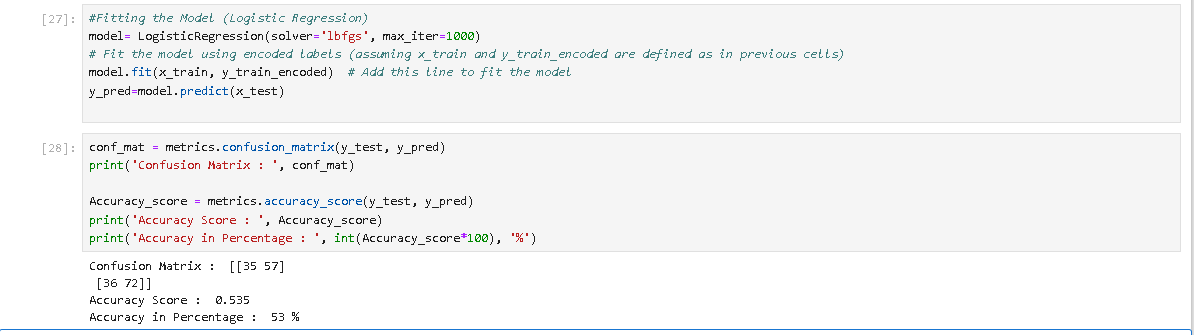
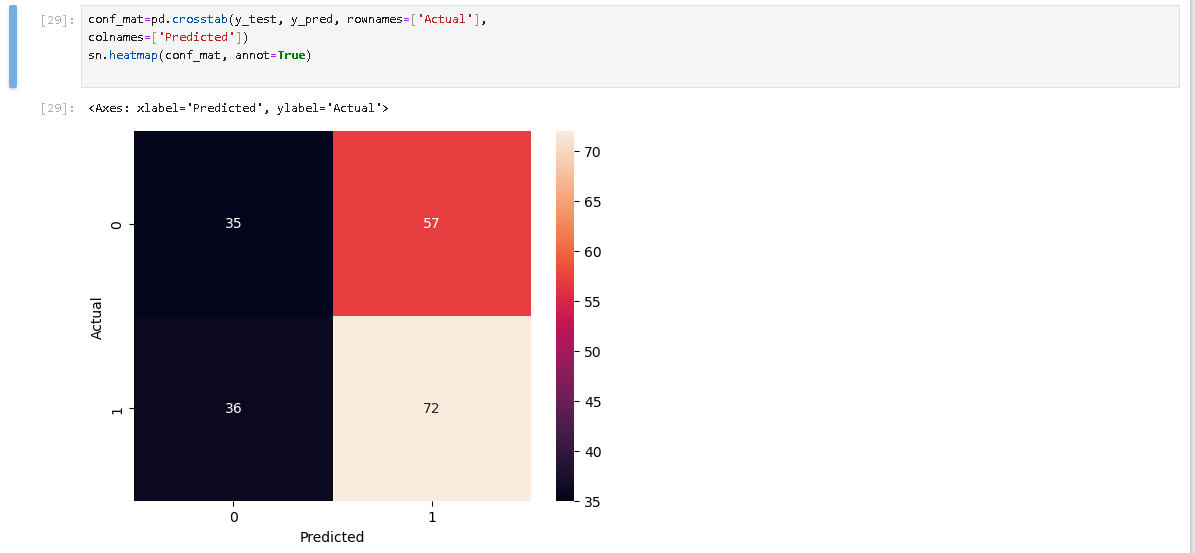
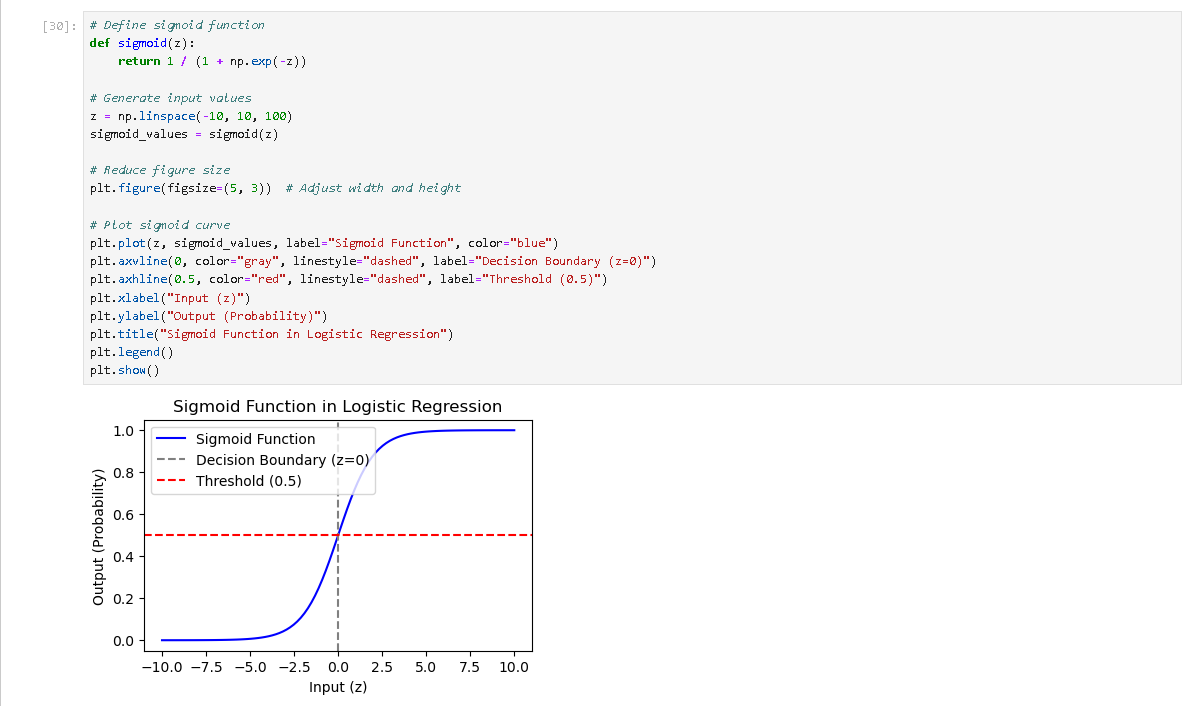
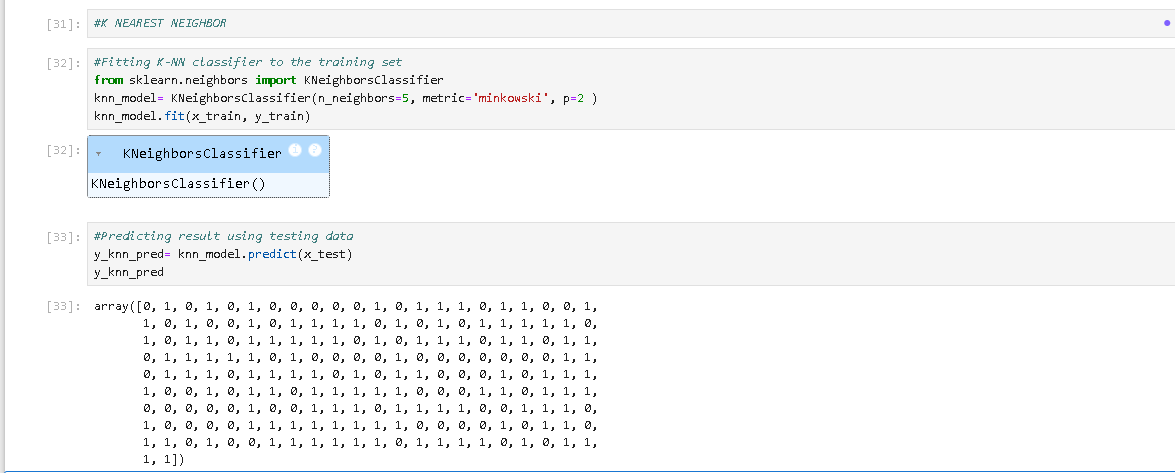
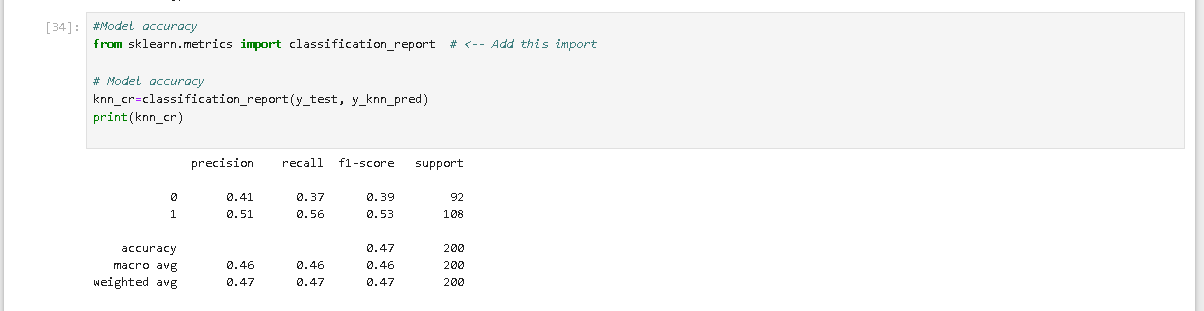
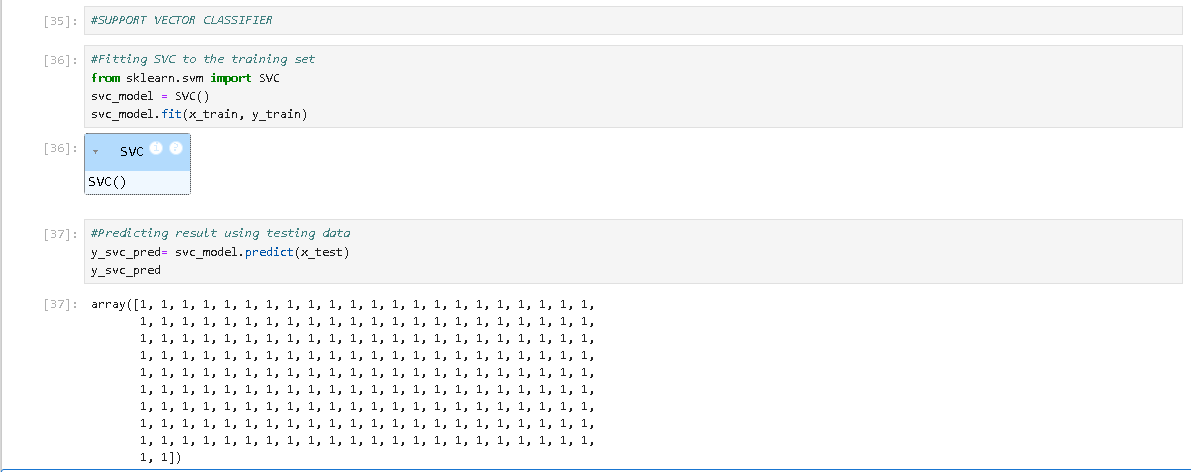
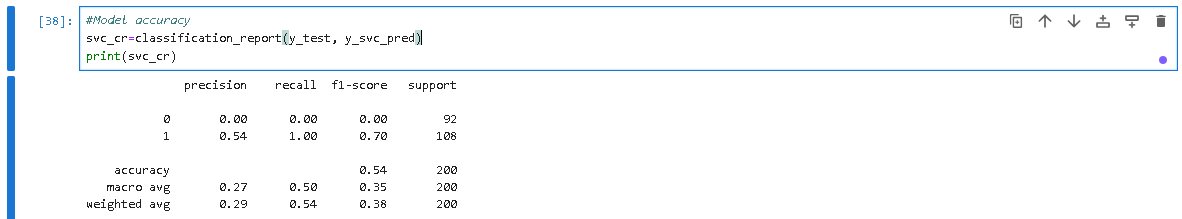
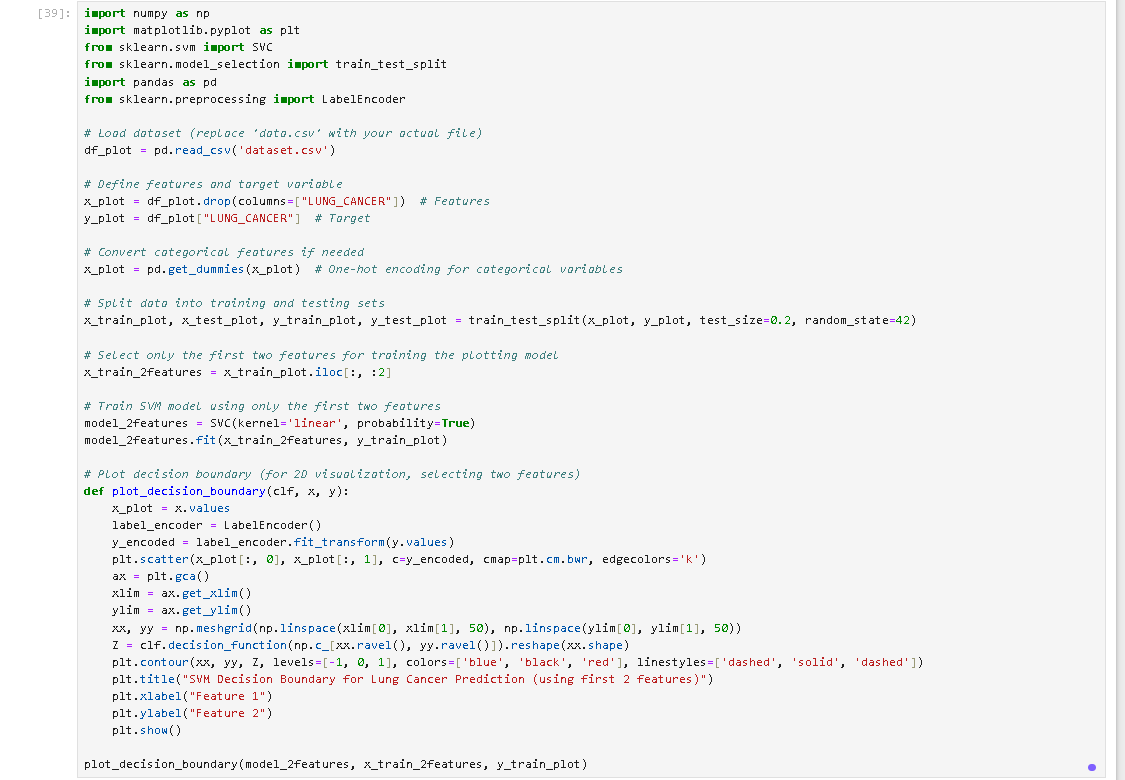
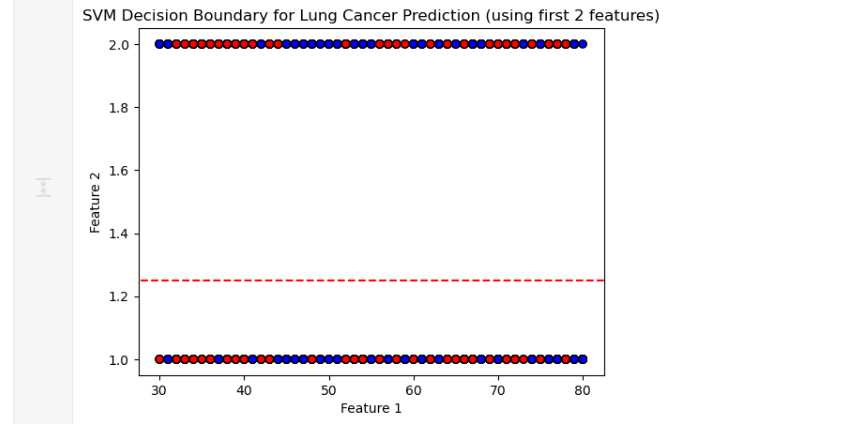


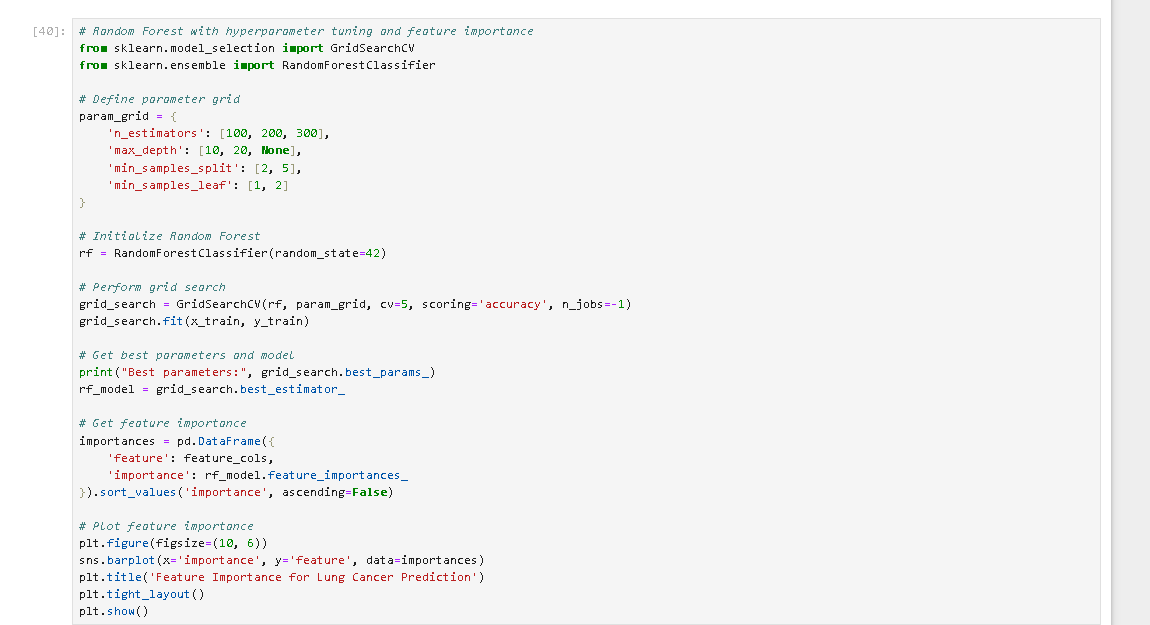


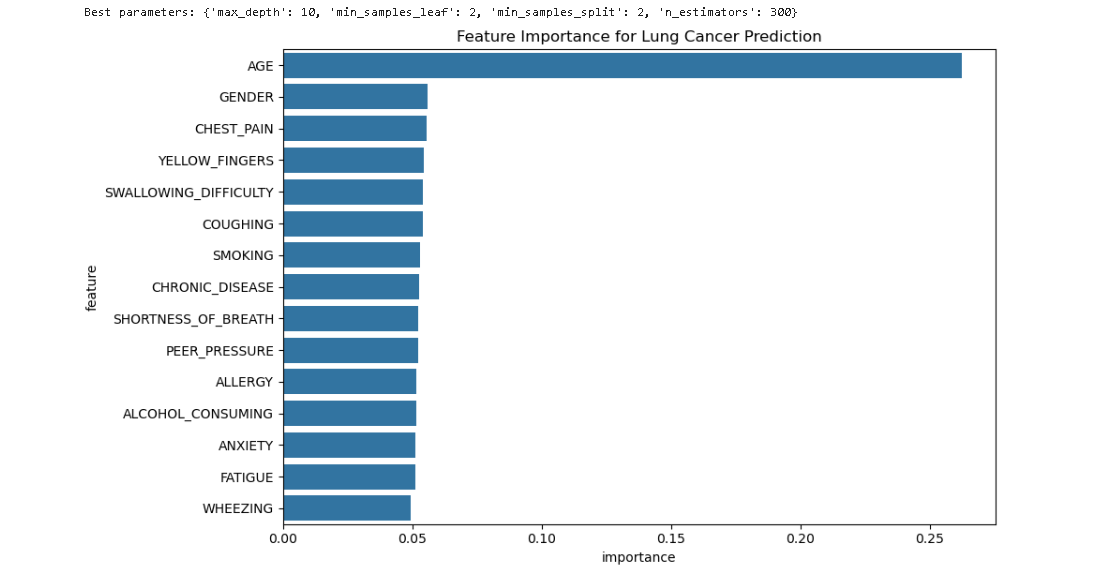


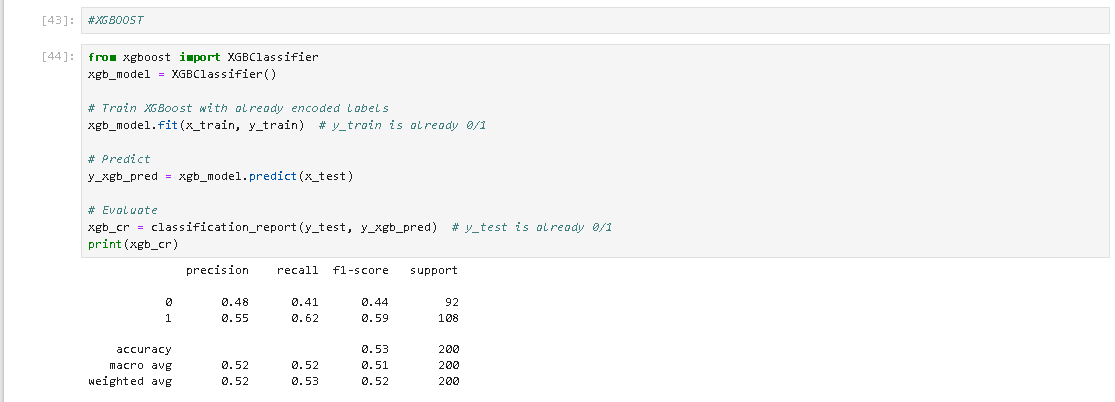
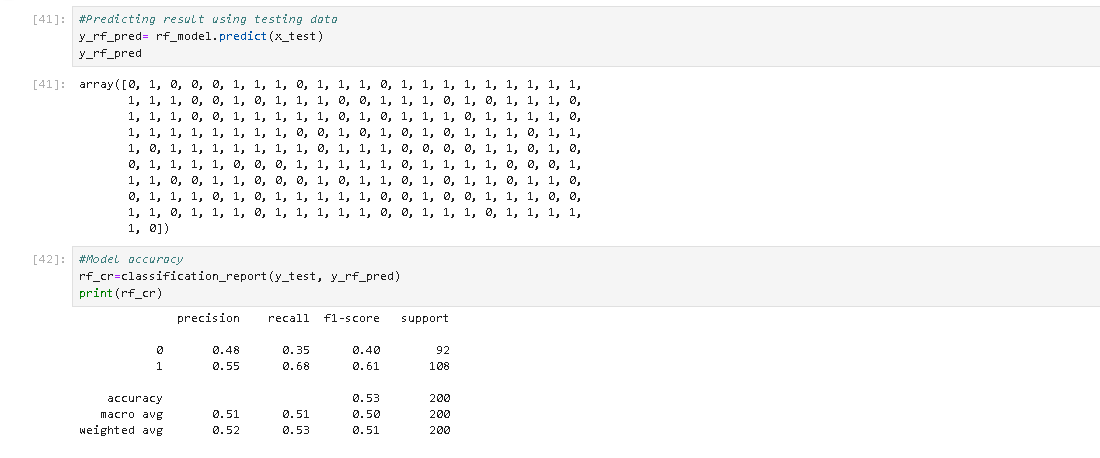


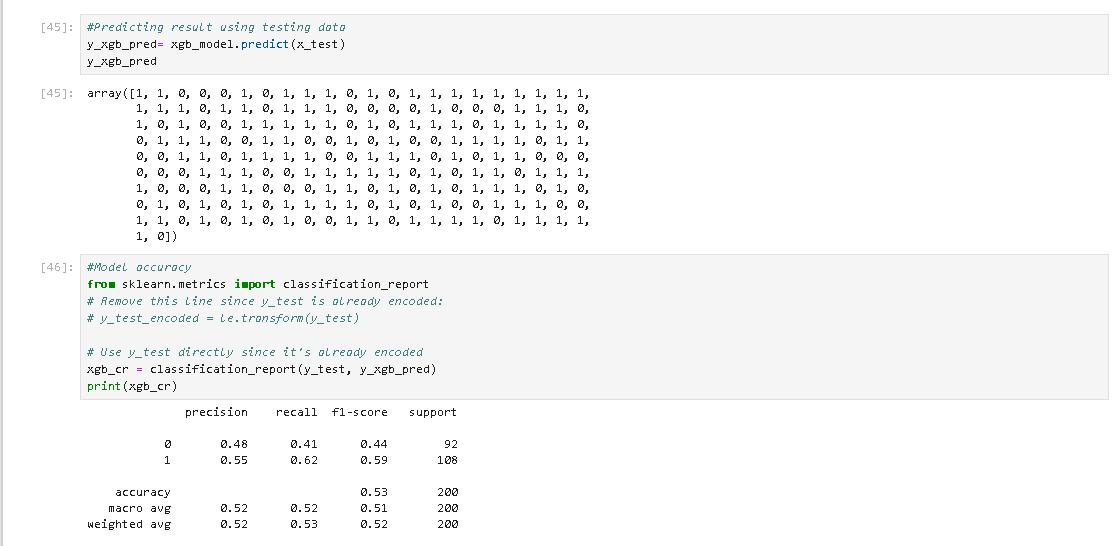


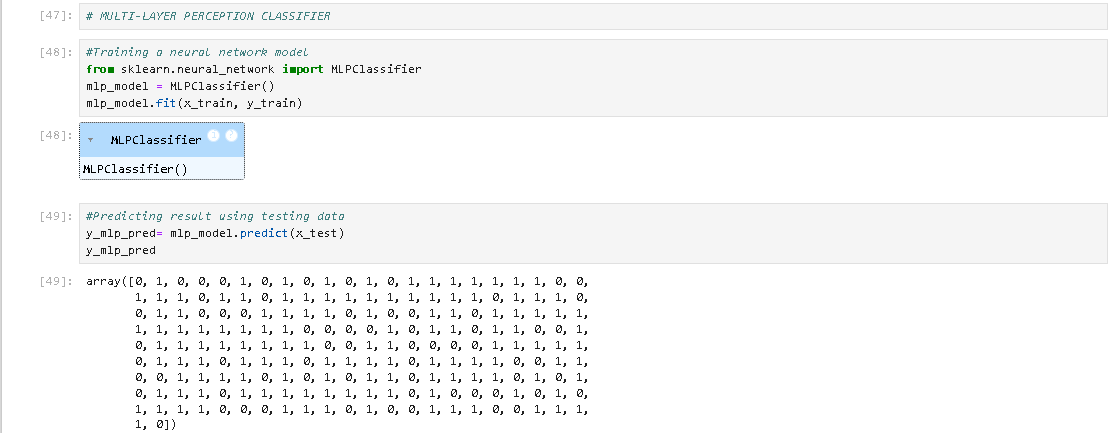
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
 

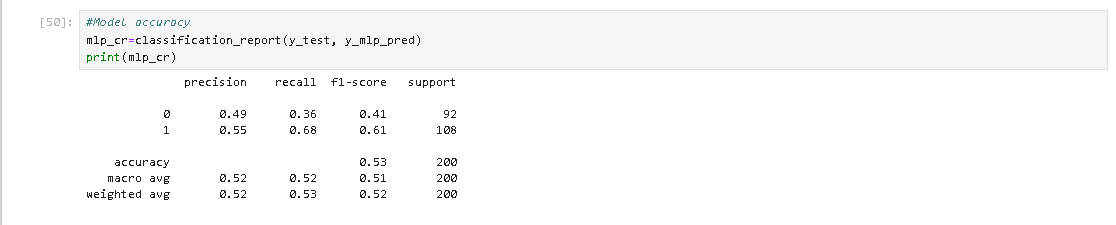


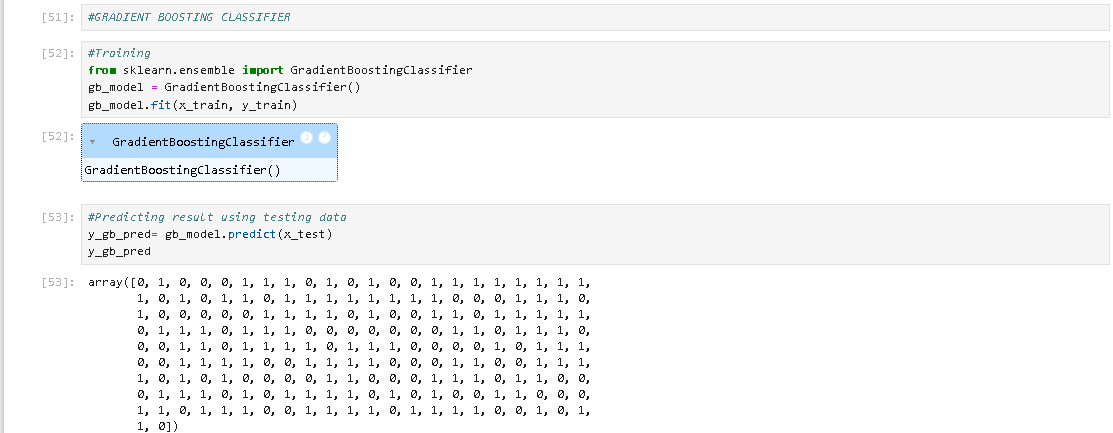


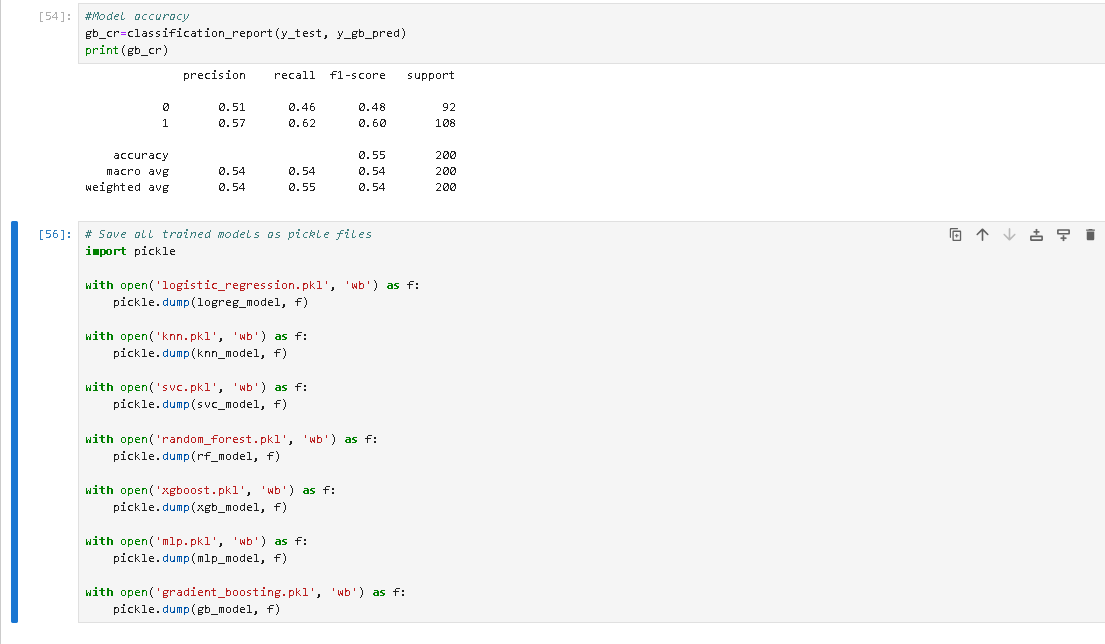




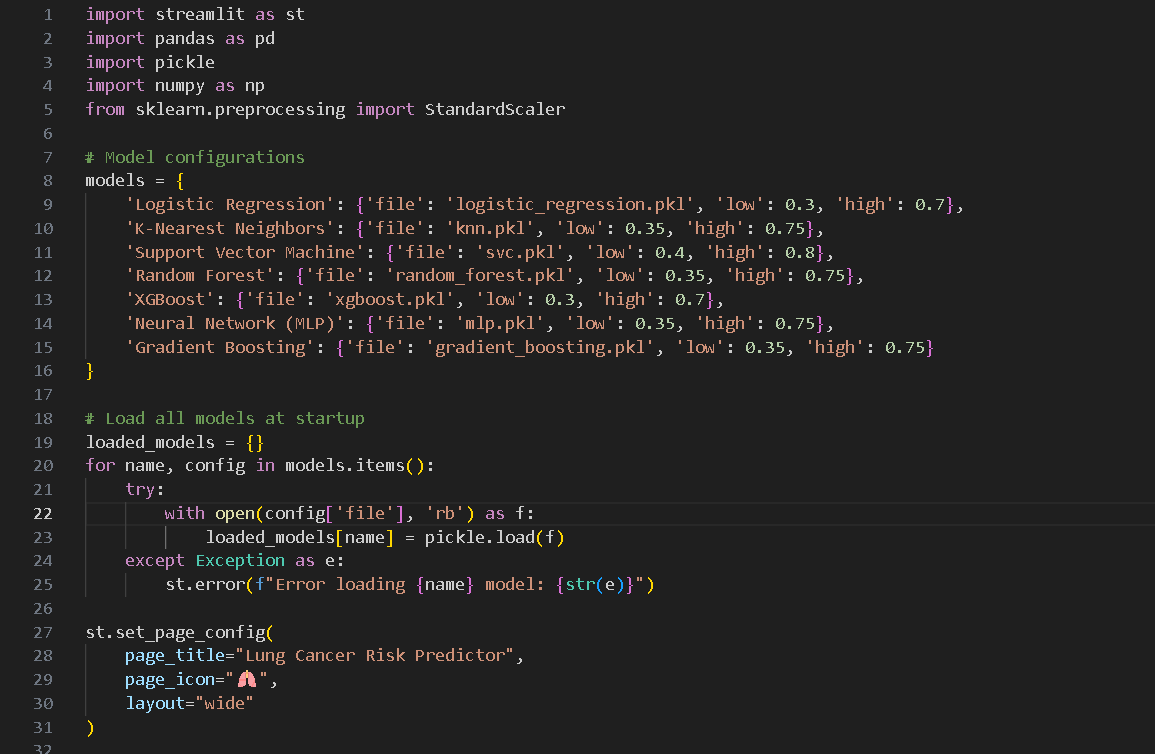


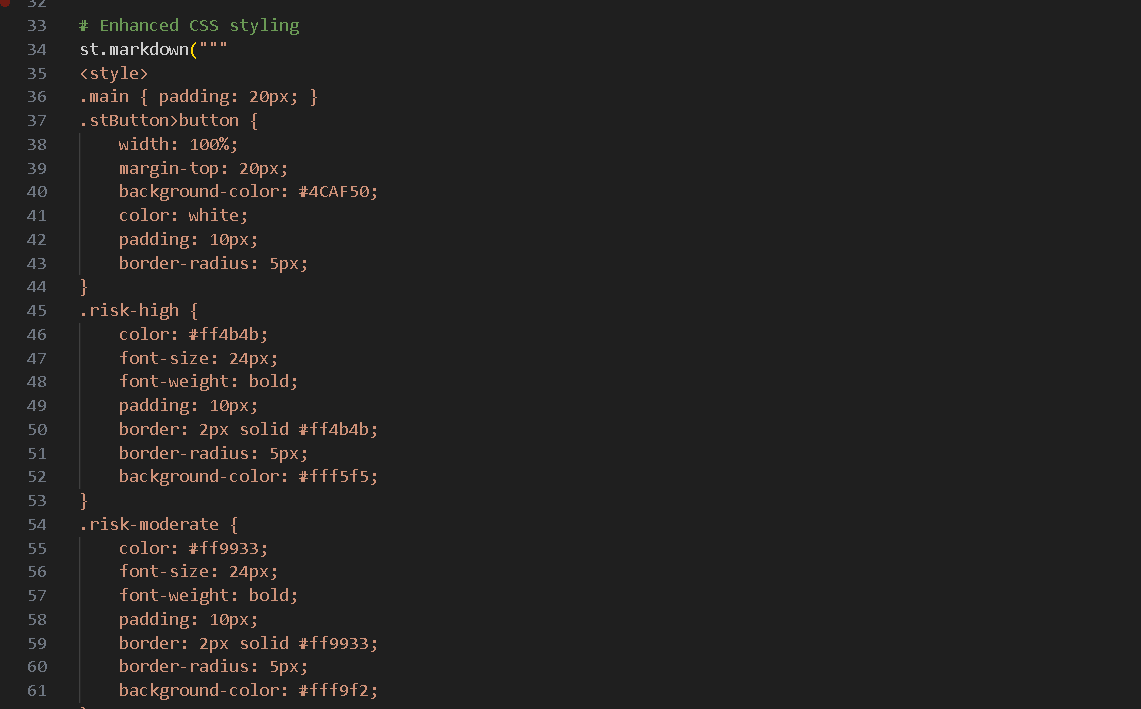


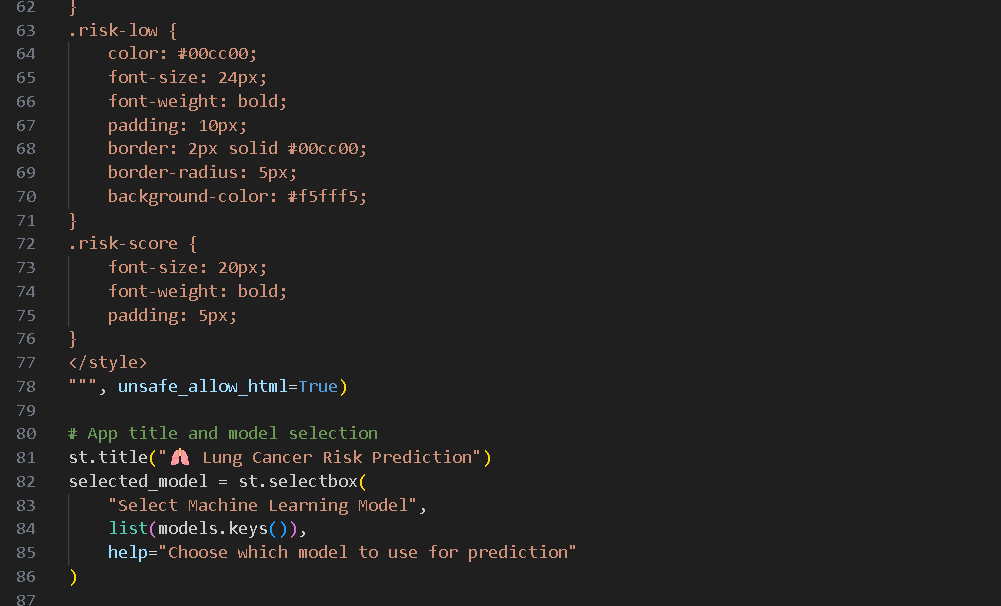


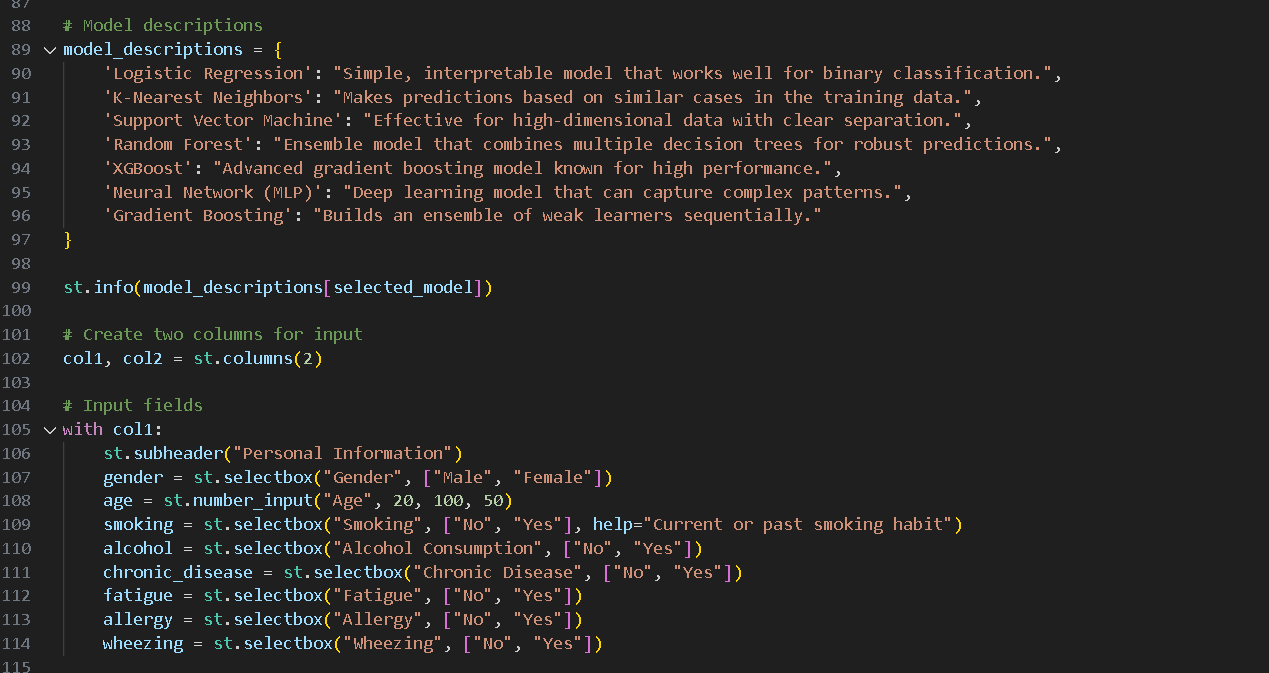


Allmodels.py:

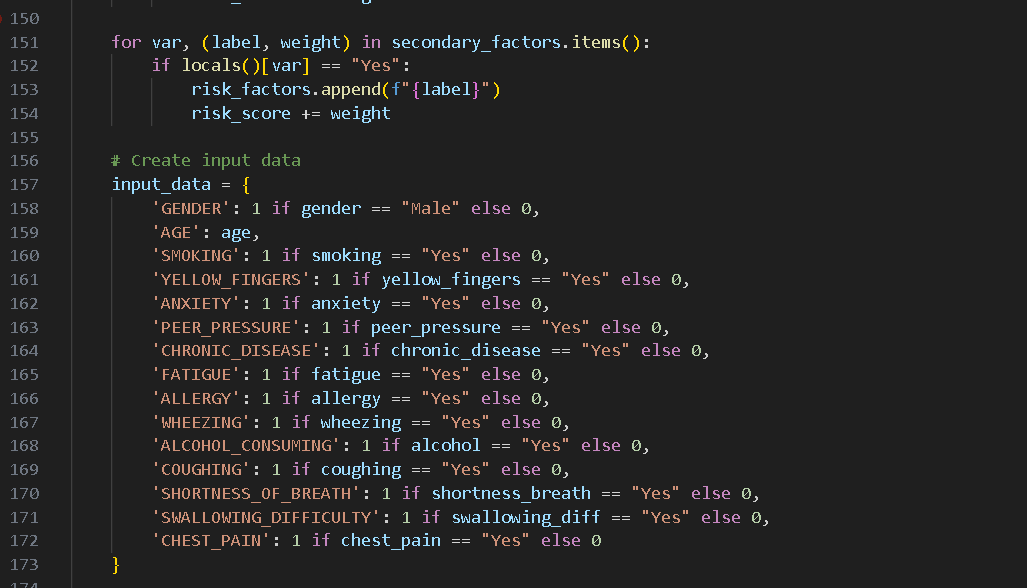


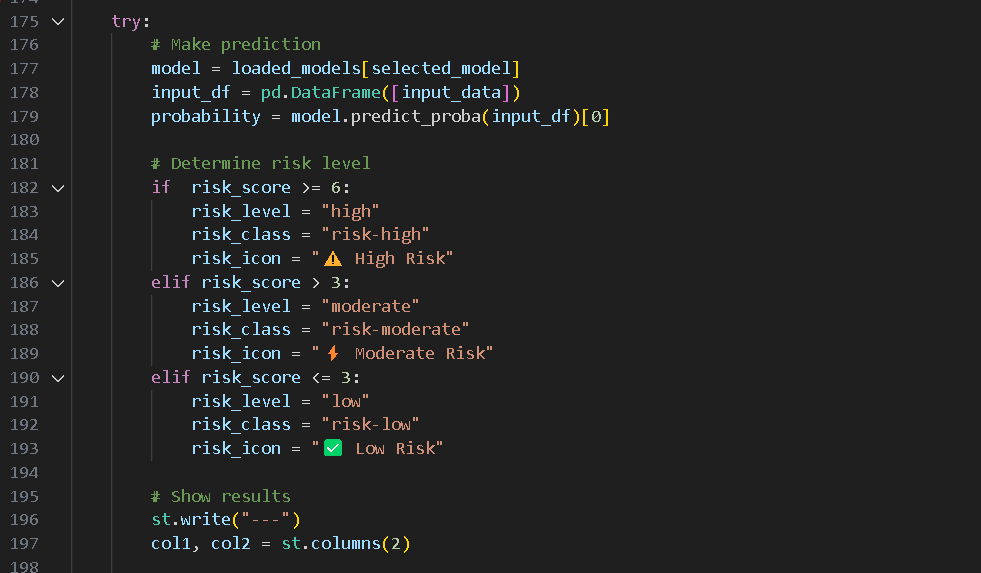


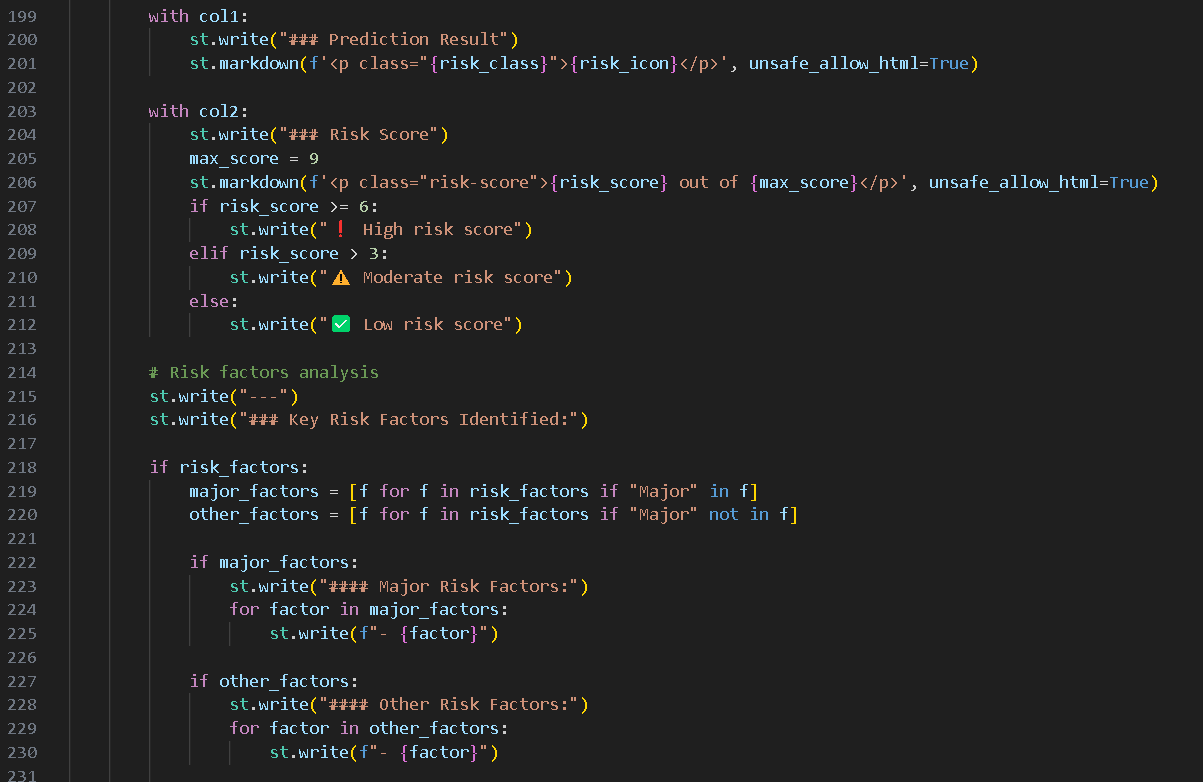




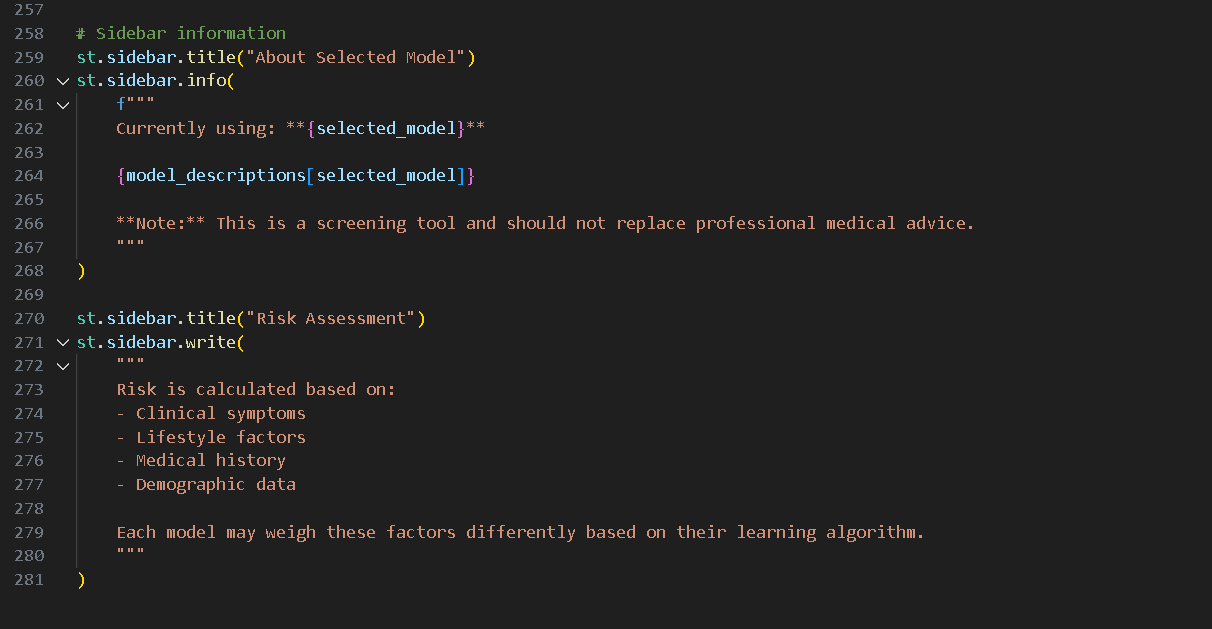




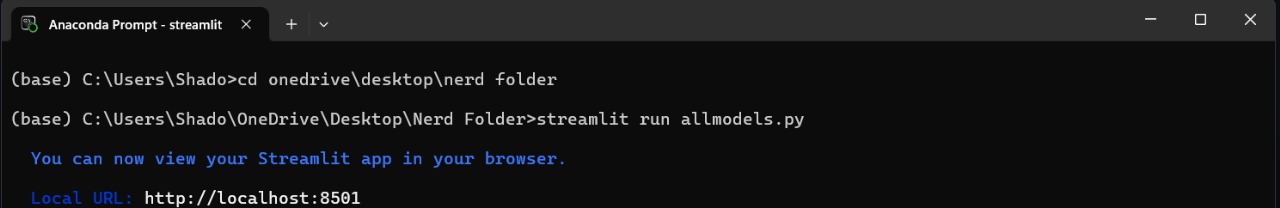




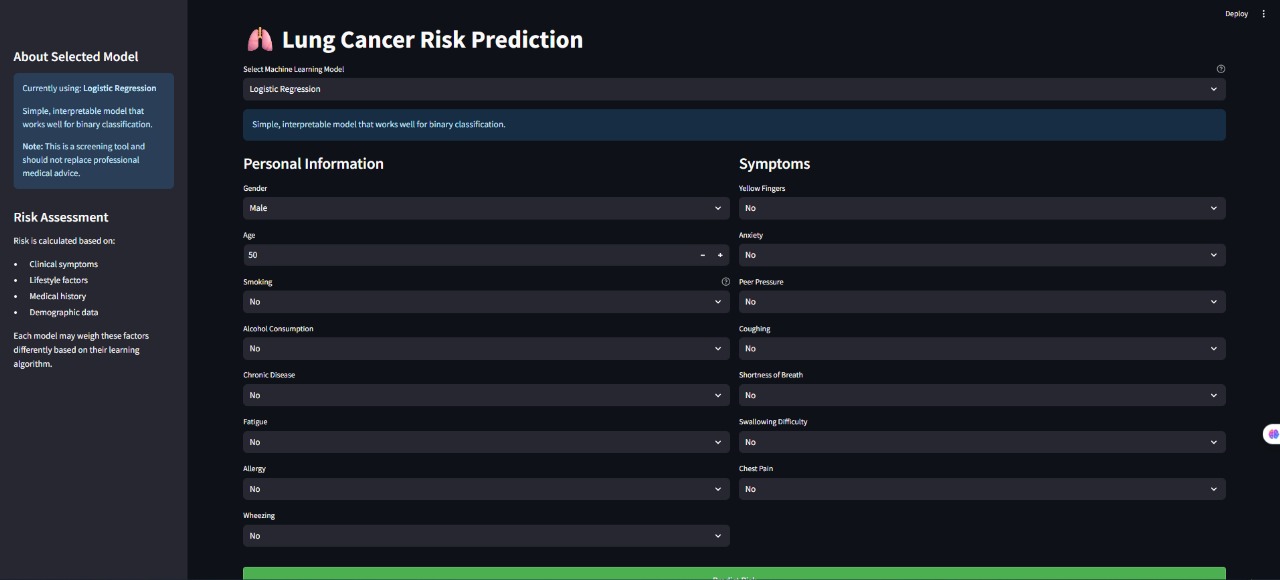


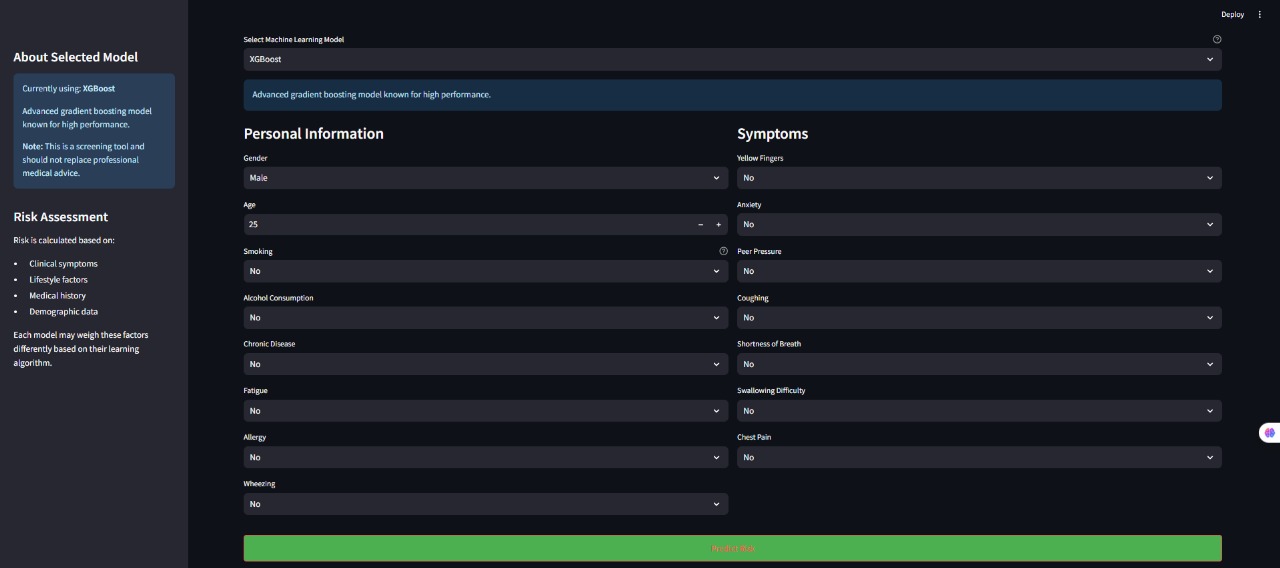


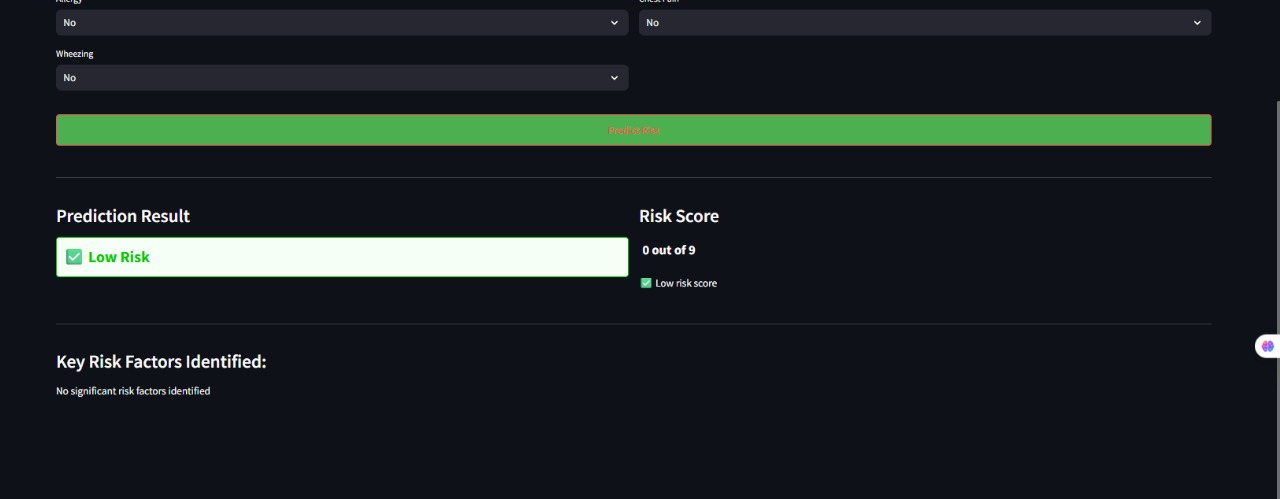
**Run Commands:**

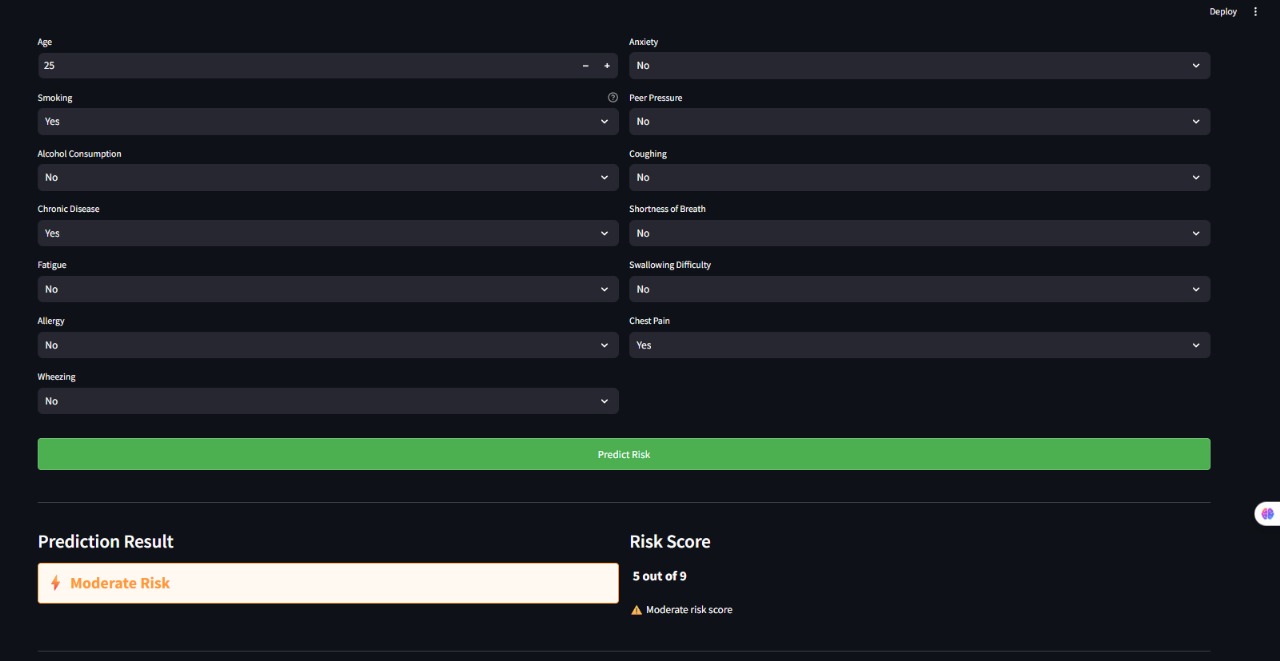


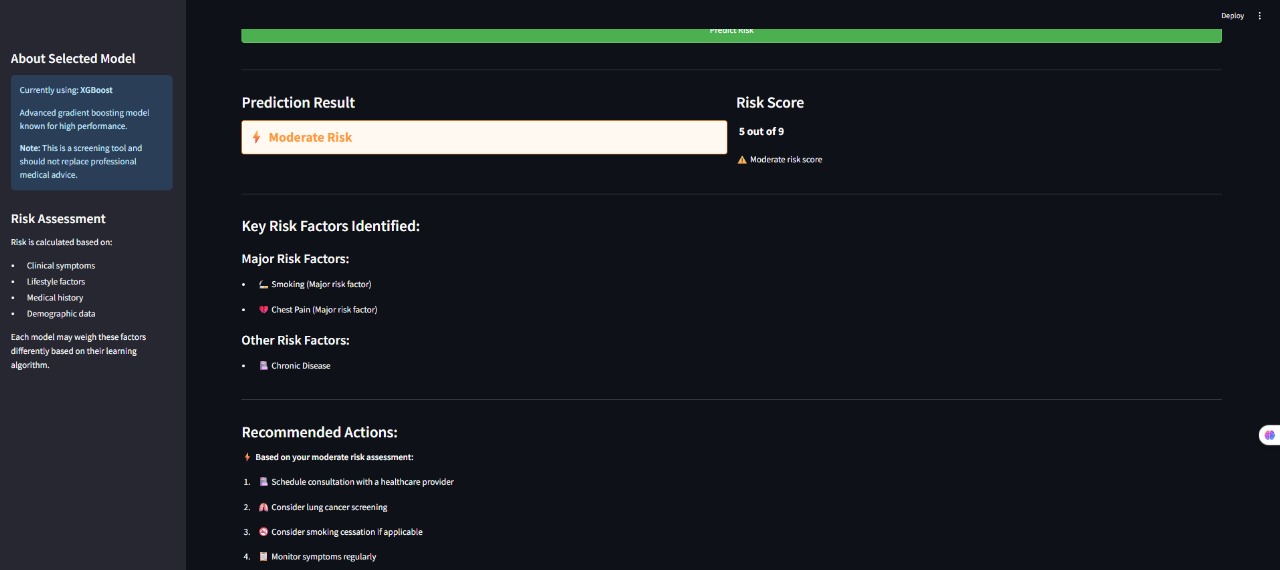
**OUTPUT:**

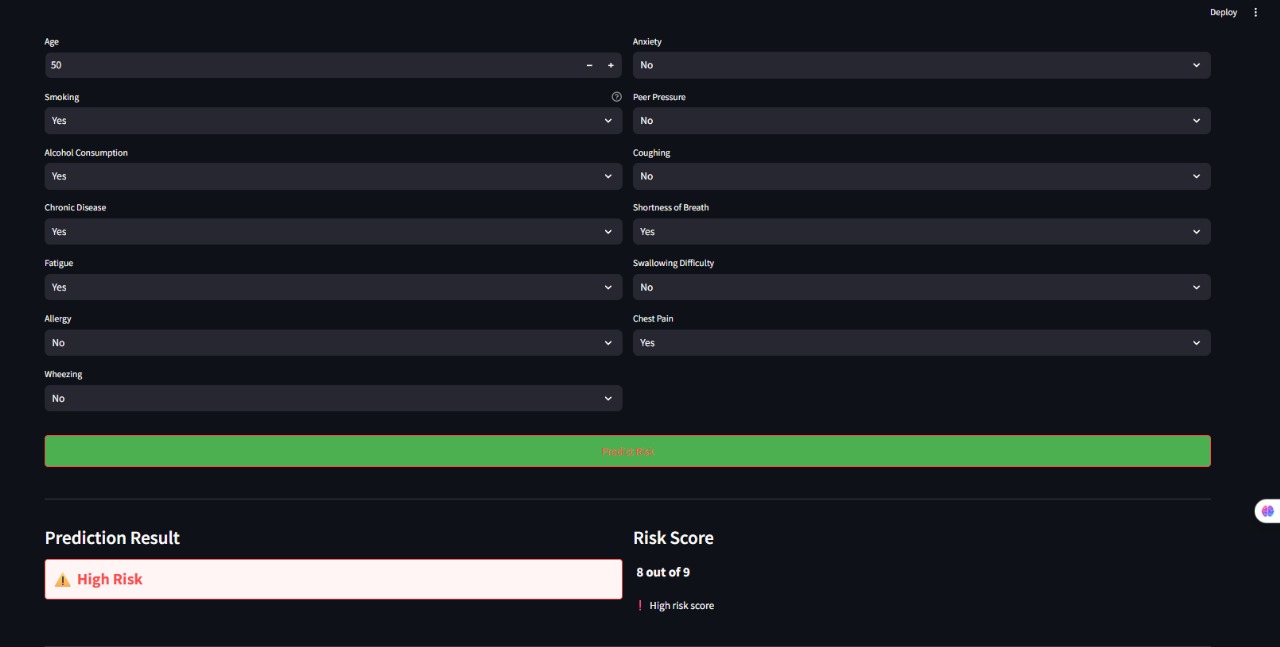


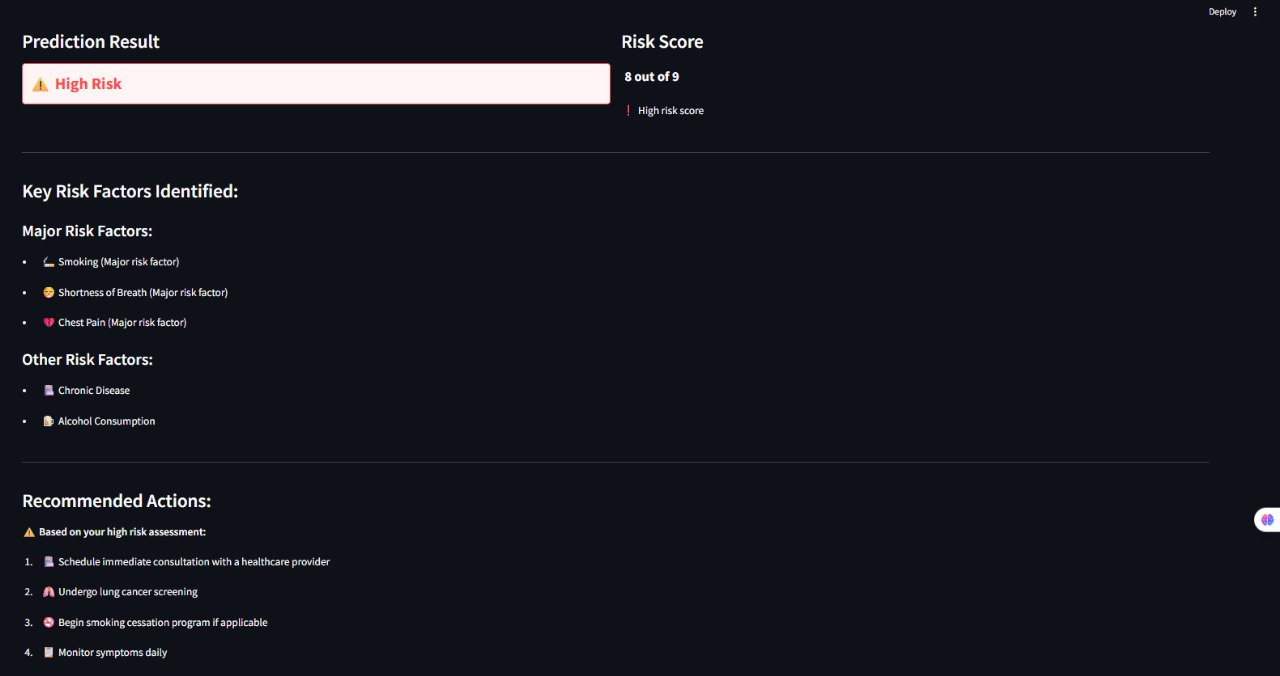












**Conclusion:**

This study successfully builds a prediction system for lung cancer using machine learning methods. After preprocessing a medical dataset and training two models (Logistic Regression and SVM), both approaches demonstrated potential for effective classification of individuals as either lung cancer positive or negative based on their symptoms, habits, and medical history.

**Key Takeaways:**

* Logistic Regression provided a quick and interpretable model for prediction.
* SVM delivered potentially more accurate classifications, especially for complex data patterns.
* Proper data cleaning, encoding, and feature selection were crucial to achieving good performance.
* The project emphasizes the practical use of ML in healthcare diagnostics, which can assist doctors in making faster, data-driven decisions.

**Github link:** [**https://github.com/AryanM177/LungCancerPrediction.git**](https://github.com/AryanM177/LungCancerPrediction.git)