December 12

Predicting Lung Cancer Likelihood Using Machine Learning Models: A Comparative Analysis

2024

Authors:

Md. Shakil Talukdar [1] Musharaf Ahmed [2] Mahfuj Hussain Nadim [3]

Table of Contents

1.1 Abstract 1
1.2 Introduction 2
1.3 Literature Review 3
1.4 Dataset 4
1.5 Workflow-Diagram 7
1.6 Methods (Algorithm used) 7
1.7 Results 8
1.7.1 Train-Test Split Pie Chart: 8
1.7.2 Logistic Regression's Heatmap for Classification Report: 9
1.7.3 Bar Chart Comparing Model Accuracies: 9
1.8 Conclusion 10
1.9 Reference 10

1.1 Abstract

Lung cancer is one of the leading causes of cancer-related deaths globally, largely due to late diagnosis. Early detection plays a vital role in improving survival rates. This project focuses on predicting the likelihood of lung cancer using machine learning (ML) techniques to enhance early diagnosis accuracy. The dataset used for this study was sourced from Kaggle and contains features such as age, gender, smoking habits, and a binary target indicating the presence or absence of lung cancer.

Several machine learning algorithms, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, were implemented to classify patients based on their risk of lung cancer. The dataset underwent preprocessing, including encoding categorical variables, correlation analysis, and feature selection. The models were evaluated based on accuracy, precision, recall, and F1 score. Among the models, XGBoost achieved the highest

accuracy on both training (e.g., 98%) and test (e.g., 94%) datasets, demonstrating its robustness in handling tabular data.

This research highlights the potential of machine learning in automating lung cancer detection, enabling healthcare providers to intervene at earlier stages. The study underscores the importance of integrating AI technologies into clinical workflows to assist in effective and timely cancer diagnosis.

1.2 Introduction

Lung cancer is a leading cause of cancer-related mortality worldwide, accounting for over 1.7 million deaths annually. Despite advances in medical technology, its survival rate remains low, largely due to late diagnosis. Lung cancer often progresses silently, with symptoms only appearing in advanced stages, limiting the efficacy of therapeutic interventions. Early detection is critical for improving patient outcomes, yet existing diagnostic tools such as imaging and biopsies are resource-intensive, invasive, and heavily dependent on healthcare professionals' expertise.

Challenges in early detection lie in its reliance on observable symptoms, which are often nonspecific and overlap with other respiratory conditions. Furthermore, existing diagnostic approaches are prone to inaccuracies and require advanced infrastructure that may not be accessible in resource-limited settings. This gap underscores the need for alternative, efficient, and scalable diagnostic solutions.

Artificial Intelligence (AI) and machine learning (ML) offer significant promise in addressing these challenges. By analyzing large datasets of patient information, ML models can identify patterns and relationships that may not be immediately apparent to human clinicians. These models, when trained on relevant features such as smoking habits, age, and genetic predisposition, can predict lung cancer risk with remarkable accuracy. Such tools have the potential to complement existing diagnostic methods, providing non-invasive, cost-effective, and reliable early detection capabilities.

This project aims to explore the use of machine learning algorithms to predict lung cancer risk and compare their performance. Using a publicly available dataset, models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost were implemented and evaluated on their predictive accuracy and robustness. By identifying the most effective model for lung cancer prediction, this study contributes to the broader effort of integrating AI-driven solutions into healthcare, ultimately enhancing early detection and improving patient outcomes.

1.3 Literature Review

The literature on the application of machine learning in healthcare, particularly for lung cancer detection, has grown substantially in recent years. Here are insights from three relevant papers:

1. **"Machine learning for lung cancer prediction using clinical and imaging data"** (2022)
   - **Authors**: Yawei Li, Xin Wu, Ping Yang, Guoqian Jiang, Yuan Luo
   - **Summary**: This paper reviews the integration of clinical data and imaging in predicting lung cancer. It emphasizes the use of ensemble models like Random Forest and SVM, which have shown superior performance in classifying malignant and benign tumors. [1]
2. **"Deep learning approaches in lung cancer diagnosis: A review"** (2023)
   - **Authors:** Lal Hussain, Hadeel Alsolai, Mohammad K Nour
   - **Summary**: This paper explores deep learning models like CNNs and their applications in analyzing radiological images for lung cancer detection. The research highlights the potential of deep learning in achieving high sensitivity and specificity.

[2]

3. **"A comparative study of machine learning algorithms for lung cancer prediction"** (2023)
   - **Authors**: Radhika P.R.; Rakhi A.S. Nair; Veena G.
   - **Summary**: This study compares several ML algorithms, including SVM, XGBoost, and Logistic Regression, for lung cancer survival prediction. The authors found that ensemble methods like Random Forest outperformed traditional methods.

[3]

1.4 Dataset

- **Source**:

The dataset used for this project is sourced from Kaggle, specifically from the "Survey Lung Cancer" dataset. This dataset includes a variety of features related to the prediction of lung cancer.

- **Features**:

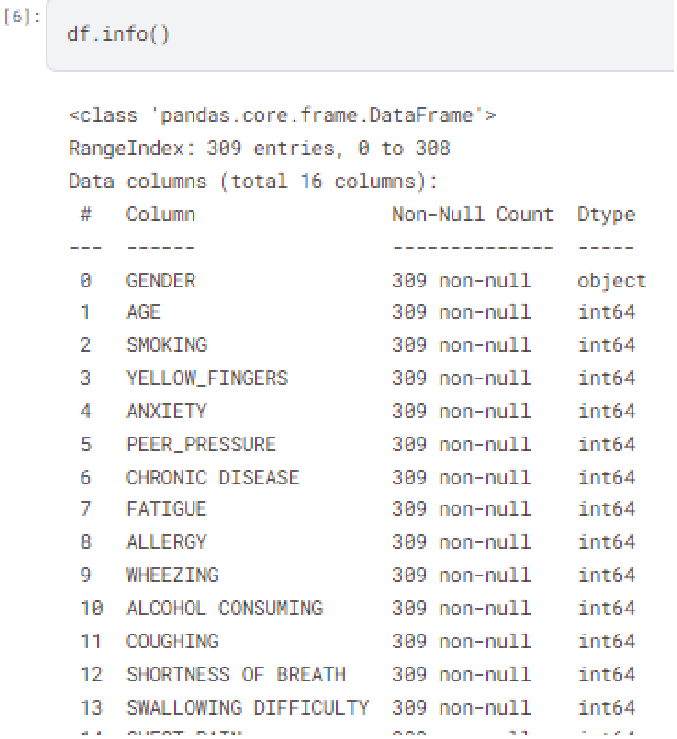
  - AGE: The age of the individual.
  - SMOKING: Smoking status (e.g., Yes/No).
  - GENDER: Gender of the individual (Male/Female).
  - LUNG_CANCER: The target variable indicating whether the individual has lung cancer (Yes/No).

- **Data Preprocessing**:

  - The LUNG_CANCER column is converted from categorical values ('YES', 'NO') to numeric values (1, 0) for model compatibility.
  - The GENDER column is encoded using Label Encoding to convert 'Male' and 'Female' into numeric values.
  - A correlation matrix is created to identify relationships between numeric features.

Descriptive Outputs:

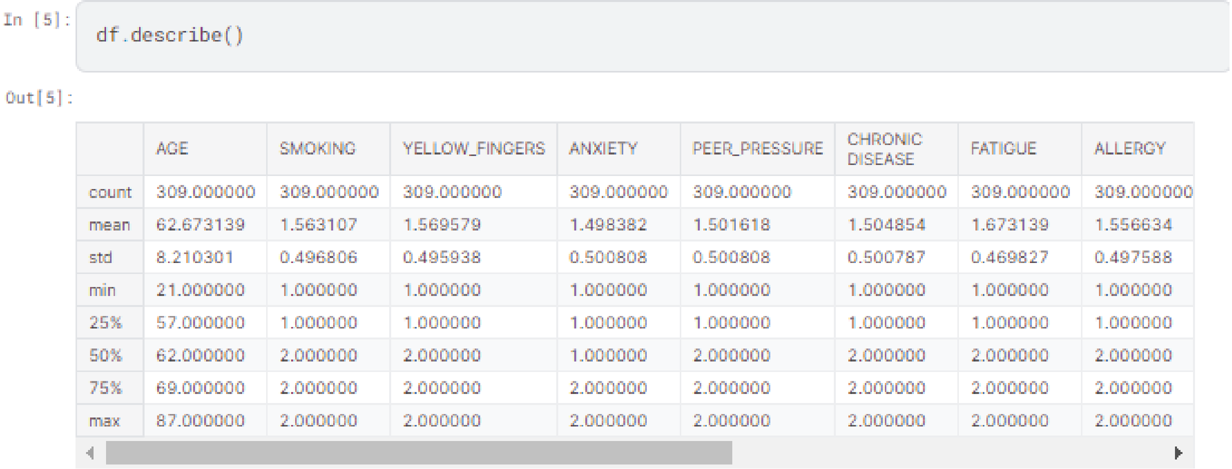
- **info()**: Provides details about the dataset including the number of entries and data types.



```
[6]:  df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype 
---  ------                 --------------  ----- 
 0   GENDER                 309 non-null    object
 1   AGE                    309 non-null    int64 
 2   SMOKING                309 non-null    int64 
 3   YELLOW_FINGERS         309 non-null    int64 
 4   ANXIETY                309 non-null    int64 
 5   PEER_PRESSURE          309 non-null    int64 
 6   CHRONIC DISEASE        309 non-null    int64 
 7   FATIGUE                309 non-null    int64 
 8   ALLERGY                309 non-null    int64 
 9   WHEEZING               309 non-null    int64 
 10  ALCOHOL CONSUMING      309 non-null    int64 
 11  COUGHING               309 non-null    int64 
 12  SHORTNESS OF BREATH    309 non-null    int64 
 13  SWALLOWING DIFFICULTY  309 non-null    int64 
```

- **describe()**: Displays basic descriptive statistics for the numeric columns (mean, standard deviation, min, max, etc.).

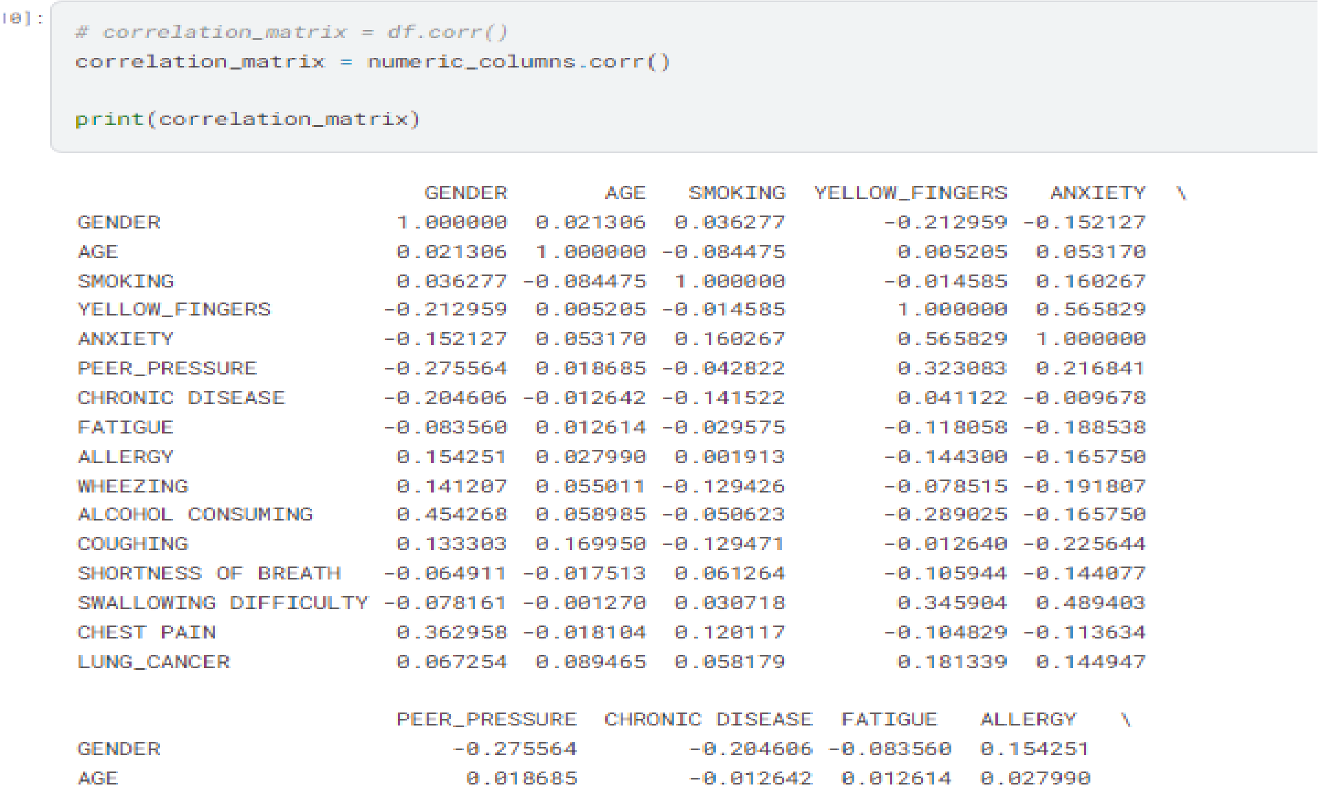


```
In [5]:  df.describe()
```

Out[5]:

	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY
count	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000
mean	62.673139	1.563107	1.569579	1.498382	1.501618	1.504854	1.673139	1.556634
std	8.210301	0.496806	0.495938	0.500808	0.500808	0.500787	0.469827	0.497588
min	21.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	57.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
50%	62.000000	2.000000	2.000000	1.000000	2.000000	2.000000	2.000000	2.000000
75%	69.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
max	87.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000

- **Correlation Matrix**: Shows the correlation between numerical columns to help understand their relationships



10]:

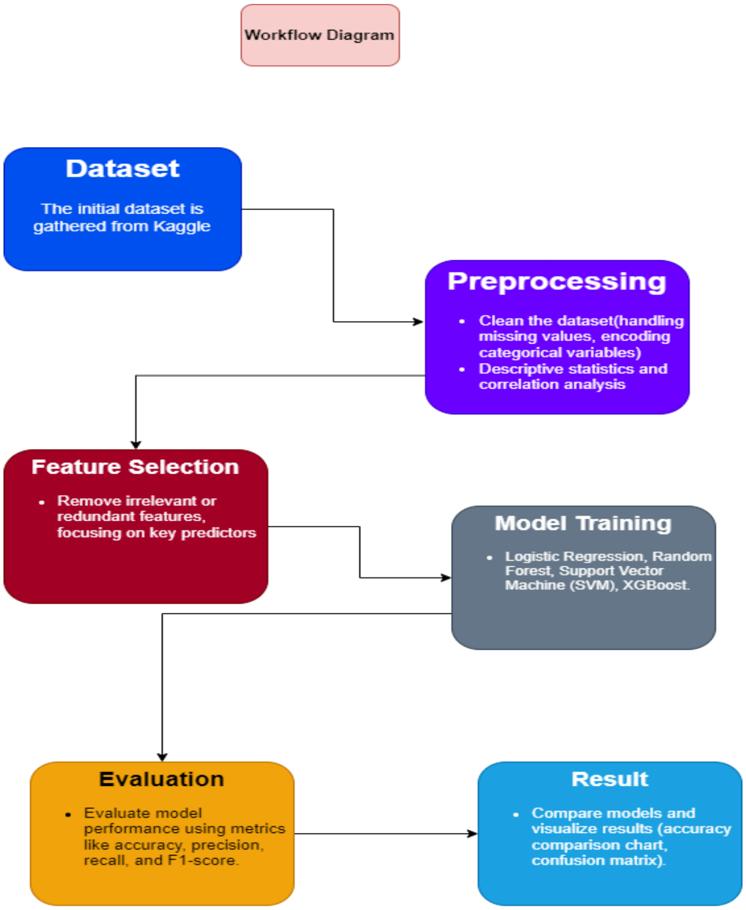
```
# correlation_matrix = df.corr()
correlation_matrix = numeric_columns.corr()

print(correlation_matrix)
```

```
                         GENDER       AGE   SMOKING  YELLOW_FINGERS   ANXIETY  \
GENDER                 1.000000  0.021306  0.036277       -0.212959 -0.152127   
AGE                    0.021306  1.000000 -0.084475        0.005205  0.053170   
SMOKING                0.036277 -0.084475  1.000000       -0.014585  0.160267   
YELLOW_FINGERS        -0.212959  0.005205 -0.014585        1.000000  0.565829   
ANXIETY               -0.152127  0.053170  0.160267        0.565829  1.000000   
PEER_PRESSURE         -0.275564  0.018685 -0.042822        0.323083  0.216841   
CHRONIC DISEASE       -0.204606 -0.012642 -0.141522        0.041122 -0.009678   
FATIGUE               -0.083560  0.012614 -0.029575       -0.118058 -0.188538   
ALLERGY                0.154251  0.027990  0.001913       -0.144300 -0.165750   
WHEEZING               0.141207  0.055011 -0.129426       -0.078515 -0.191807   
ALCOHOL CONSUMING      0.454268  0.058985 -0.050623       -0.289025 -0.165750   
COUGHING               0.133303  0.169950 -0.129471       -0.012640 -0.225644   
SHORTNESS OF BREATH   -0.064911 -0.017513  0.061264       -0.105944 -0.144077   
SWALLOWING DIFFICULTY -0.078161 -0.001270  0.030718        0.345904  0.489403   
CHEST PAIN             0.362958 -0.018104  0.120117       -0.104829 -0.113634   
LUNG_CANCER            0.067254  0.089465  0.058179        0.181339  0.144947   

                       PEER_PRESSURE  CHRONIC DISEASE   FATIGUE   ALLERGY  \
GENDER                     -0.275564        -0.204606 -0.083560  0.154251   
AGE                         0.018685        -0.012642  0.012614  0.027990   
```

1.5 Workflow-Diagram


Workflow Diagram
Dataset
The initial dataset is gathered from Kaggle
Preprocessing
Clean the dataset(handling missing values, encoding categorical variables)
Descriptive statistics and correlation analysis
Feature Selection
Remove irrelevant or redundant features, focusing on key predictors
Model Training
Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost.
Evaluation
Evaluate model performance using metrics like accuracy, precision, recall, and F1-score.
Result
Compare models and visualize results (accuracy comparison chart, confusion matrix).

1.6 Methods (Algorithm used)

- **Logistic Regression**:

  A statistical method for binary classification, used to predict the probability of a binary outcome. It models the relationship between a dependent binary variable and one or more independent variables using a logistic function.

- **Random Forest**:

  An ensemble learning technique that constructs multiple decision trees during training and outputs the mode of the classes. It helps improve predictive accuracy by reducing overfitting and increasing generalization.

- **Support Vector Machine (SVM)**:

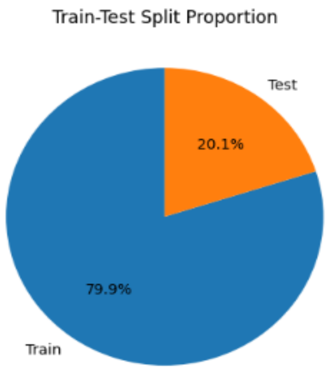
  A supervised learning algorithm that finds the optimal hyperplane that separates data into different classes. It maximizes the margin between the support vectors, which are the closest data points to the hyperplane.

- **XGBoost**:

  A powerful machine learning algorithm based on gradient boosting, known for its efficiency and accuracy, particularly with tabular data. It optimizes decision trees sequentially, with each new tree focusing on the errors made by previous ones.

1.7 Results

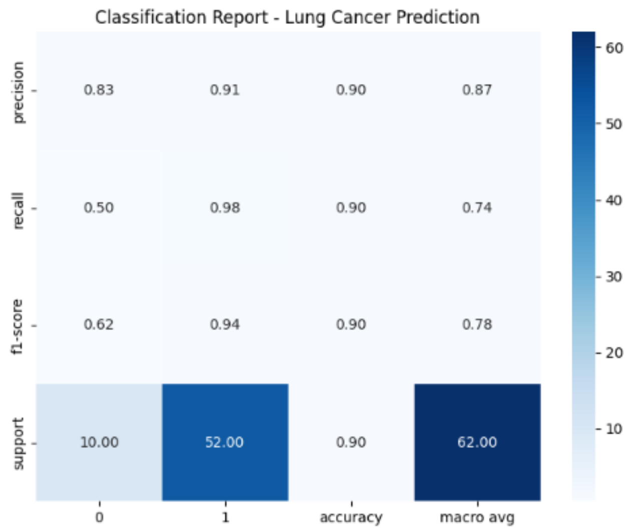
1.7.1 Train-Test Split Pie Chart**:**



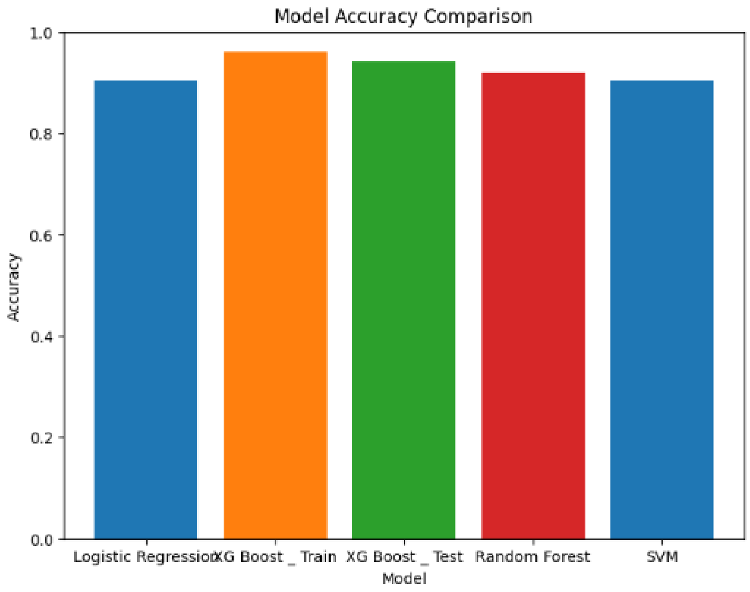
This chart visually represents the proportion of data allocated for training and testing. It shows the split between training data and testing data, typically 80% for training and 20% for testing.

1.7.2 Logistic Regression’s Heatmap for Classification Report**:**

The heatmap visualizes the performance metrics (precision, recall, F1-score) of the Logistic Regression model, providing insights into classification accuracy.



1.7.3 Bar Chart Comparing Model Accuracies:



A bar chart displaying the accuracy comparison across different models: Logistic Regression, Random Forest, SVM, and XGBoost.

1.8 Conclusion

This project focused on predicting lung cancer likelihood using various machine learning models. The models tested, including Logistic Regression, Random Forest, SVM, and XGBoost, demonstrated varying degrees of performance. The results showed that XGBoost achieved the highest accuracy, followed by Random Forest and SVM. This study underscores the potential of machine learning in early detection of lung cancer, which is crucial for improving patient outcomes. Future improvements could involve using larger datasets or more advanced algorithms to further enhance prediction accuracy and robustness.

1.9 Reference

[1]. https://academic.oup.com/gpb/article/20/5/850/7230459

[2] https://www.mdpi.com/2076-3417/12/13/6517

[3] https://ieeexplore.ieee.org/document/8869001