LUNG CANCER PREDICTION USING MACHINE LEARNING TECHNIQUES

BY

M.K.Akash
ID: 201-15-13736

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Mr. Md. Aynul Hasan Nahid
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2024

APPROVAL

This Project titled "LUNG CANCER PREDICTION USING MACHINE LEARNING TECHNIQUES", submitted by M.K.AKASH to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 15/7/2024.

<u>BOARD OF EXAMINERS</u>

____

Narayan Ranjan Chakraborty (NRC) — **Board Chairman**
Associate Professor & Associate Head
Department of CSE, DIU
Faculty of Science & Information Technology
Daffodil International University

____

Mushfiqur Rahman (MUR) — **Internal Examiner 1**
Sr. Lecturer
Department of CSE, DIU
Faculty of Science & Information Technology
Daffodil International University

15.7.24
____

Israt Jahan (IJN) — **Internal Examiner 2**
Sr. Lecturer
Department of CSE, DIU
Faculty of Science & Information Technology
Daffodil International University

15.07.24
____

Dr. Abu Sayed Md. Mostafizur Rahaman (ASMR) — **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by **M.K.Akash and id: 201-15-1376** under the supervision of **Mr. Md. Aynul Hasan Nahid, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Mr. Md. Aynul Hasan Nahid
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Md. Ferdouse Ahmed Foysal
Lecturer
Department of CSE
Daffodil International University

Submitted by:
M.K.Akash
M.K.Akash
ID: -201-15-13736
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Mr. Md. Aynul Hasan Nahid, Lecturer** Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "***Faculty of Science and Information Technology***" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Mr. Md. Aynul Hasan Nahid, Lecturer,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

In my thesis project, "Lung Cancer Prediction Using Machine Learning Techniques," I aimed to develop a reliable system for predicting lung cancer risk through the application of various machine learning algorithms. The dataset utilized was sourced from Kaggle, originating from an online lung cancer prediction system. It comprised multiple attributes related to individuals' demographics, lifestyle choices, and health symptoms, with a binary target variable indicating the presence or absence of lung cancer. Initially, I preprocessed the dataset, converting certain column values to binary (0 and 1) and addressing missing values. During exploratory data analysis, I identified an imbalance in the target distribution and mitigated it using oversampling techniques. Additionally, I performed feature engineering by eliminating irrelevant features and creating new ones to enhance predictive capability. To reduce dimensionality, I employed Principal Component Analysis (PCA) before training several machine learning models including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. Among these models, Logistic Regression emerged as the top performer, achieving an accuracy of 95%. Subsequently, I applied Grid Search on Logistic Regression to optimize hyperparameters, resulting in a slight accuracy improvement to 94.89%. Despite experimenting with ensemble techniques like Voting Classifier, Logistic Regression consistently outperformed other models. Finally, I conducted K-Fold cross-validation to validate model robustness, with Logistic Regression demonstrating the highest average accuracy compared to Decision Tree and Multi-layer Perceptron. In conclusion, my research highlights Logistic Regression as the most effective model for lung cancer risk prediction, emphasizing its accuracy and reliability based on the given dataset and features.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Overview

In my thesis project, "Lung Cancer Prediction Using Machine Learning Techniques," I aim to develop an effective system for predicting lung cancer risk by leveraging machine learning algorithms. The dataset I obtained from Kaggle originates from an online lung cancer prediction system and contains various attributes related to individuals' demographics, lifestyle choices, and health symptoms, with a target variable indicating the presence or absence of lung cancer. To prepare the dataset for machine learning, I converted the values of certain columns to binary (0 and 1) and handled missing values. During exploratory data analysis, I observed an imbalance in the target distribution and addressed it using techniques like oversampling. Additionally, I performed feature engineering, dropping irrelevant features and creating new ones to enhance predictive power. Principal Component Analysis (PCA) was applied for dimensionality reduction, and I then trained several machine learning models including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. Among these, Logistic Regression achieved the highest accuracy of 95%. Subsequently, I applied Grid Search on Logistic Regression to optimize hyperparameters, resulting in a slight improvement in accuracy to 94.89%. Furthermore, I experimented with ensemble techniques like Voting Classifier, but found Logistic Regression to still outperform other models. Finally, I conducted K-Fold cross-validation to validate the robustness of my models, with Logistic Regression consistently demonstrating the highest average accuracy compared to Decision Tree and Multi-layer Perceptron. Overall, my research demonstrates that Logistic Regression is the most effective model for predicting lung cancer risk based on the given dataset and features, achieving a high level of accuracy and reliability.

1.2 Problem Statement

In the field of healthcare, early detection of diseases such as lung cancer plays a crucial role in improving patient outcomes and reducing mortality rates. However, traditional diagnostic methods often rely on costly and invasive procedures, leading to delays in detection and diagnosis. My project addresses this challenge by leveraging machine learning techniques to develop a predictive model for lung cancer risk assessment.

Key points:

I. Early detection of lung cancer can significantly improve patient outcomes and reduce mortality rates.
II. Traditional diagnostic methods for lung cancer are often invasive and costly, leading to delays in diagnosis.
III. Machine learning offers a promising approach to predict lung cancer risk based on patient demographics, lifestyle factors, and health symptoms.
IV. Developing an accurate and reliable predictive model can aid healthcare professionals in identifying individuals at high risk of lung cancer, enabling timely intervention and treatment.

By utilizing a dataset from an online lung cancer prediction system, my project aims to build a machine learning model capable of accurately predicting lung cancer risk, thereby contributing to early detection and improved patient care.

1.3 Motivation and Objectives

The motivation behind my project is to contribute to the field of healthcare by developing a robust predictive model for lung cancer using machine learning techniques. Lung cancer remains one of the leading causes of cancer-related deaths worldwide, with early detection being crucial for effective treatment and improved patient outcomes. However, traditional diagnostic methods often lack accuracy and efficiency, leading to delayed diagnoses and suboptimal management of the disease.

I. To leverage machine learning algorithms to develop a predictive model for lung cancer risk assessment.
II. To explore and analyze a dataset obtained from Kaggle, sourced from an online lung cancer prediction system, to identify relevant features and patterns associated with lung cancer.
III. To preprocess and prepare the dataset for machine learning, including handling missing values, converting data types, and addressing class imbalance.
IV. To apply feature engineering techniques to enhance the predictive power of the model and reduce dimensionality.
V. To evaluate and compare the performance of various machine learning algorithms, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier.
VI. To optimize the hyperparameters of the selected model using Grid Search to improve predictive accuracy.
VII. To validate the robustness of the developed model through techniques like K-Fold cross-validation and assess its generalization to unseen data.

By achieving these objectives, I aim to contribute to the development of a reliable and efficient lung cancer prediction system that can assist healthcare professionals in early detection and intervention, ultimately leading to improved patient outcomes and reduced mortality rates.

1.4 Project Scope

In my project, I aim to develop a robust lung cancer prediction system using machine learning techniques. The dataset I acquired from Kaggle provides valuable insights into various attributes related to individuals' demographics, lifestyle choices, and health symptoms, offering a rich source of information for predictive modeling.

I. Utilizing machine learning algorithms to analyze the dataset and identify patterns that can predict lung cancer risk accurately.

II. Implementing preprocessing techniques to handle missing values, convert data into suitable formats, and address imbalances in the target distribution.
III. Exploring feature engineering methods to enhance the predictive power of the model by selecting relevant features and creating new ones.
IV. Employing dimensionality reduction techniques like Principal Component Analysis (PCA) to streamline the dataset and improve computational efficiency.
V. Training and evaluating multiple machine learning models such as Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier to identify the most effective approach.
VI. Optimizing hyperparameters using Grid Search and experimenting with ensemble techniques like Voting Classifier to further enhance model performance.
VII. Validating the robustness of the models through K-Fold cross-validation to ensure reliable predictions in real-world scenarios.

By leveraging these techniques and methodologies, I aim to create a predictive model that accurately assesses an individual's risk of lung cancer based on their demographic, lifestyle, and health-related factors. Ultimately, this project seeks to contribute to the development of effective and accessible tools for early detection and prevention of lung cancer.

1.5 Project Outcome

The goal of this project is to develop a predictive model for lung cancer using various machine learning techniques. Leveraging a dataset sourced from Kaggle, which originates from an online lung cancer prediction system, I aim to build a robust and accurate system that can assess an individual's risk of developing lung cancer based on demographic information, lifestyle choices, and health symptoms.

I. Utilizing machine learning algorithms to predict lung cancer risk offers significant potential benefits, including:

II. Early detection: Identifying individuals at high risk of developing lung cancer allows for early intervention and treatment, potentially improving patient outcomes and survival rates.

III. Personalized healthcare: By analyzing individual risk factors, the predictive model can provide personalized recommendations for lifestyle changes, screening protocols, and preventive measures tailored to each patient's specific needs.

IV. Cost-effective screening: Implementing an effective lung cancer prediction system can help optimize healthcare resources by targeting high-risk individuals for screening and intervention, reducing unnecessary tests and costs associated with false positives.

V. Informed decision-making: Providing individuals with accurate information about their lung cancer risk empowers them to make informed decisions about their health, such as quitting smoking, adopting healthier habits, or seeking further medical evaluation.

VI. Accurate risk assessment: The model will accurately classify individuals into high and low-risk categories based on their demographic and health-related features, enabling targeted interventions and screenings.

VII. Robust performance: Through rigorous evaluation and validation using techniques like cross-validation, the model will demonstrate consistent and reliable performance across diverse datasets and real-world scenarios.

VIII. Practical usability: The developed system will be user-friendly and accessible, allowing healthcare professionals to easily interpret and utilize the predictions to guide patient care and decision-making.

Overall, the project outcome aims to contribute to the advancement of personalized medicine and preventive healthcare strategies, ultimately improving the early detection and management of lung cancer, a significant public health concern worldwide.

1.6 Report Organization

As for organizing the report, I will structure it as follows:

1. Introduction
    - Brief overview of the problem statement and the dataset used.
    - Explanation of the importance of the task and its relevance in the medical field.
2. Data Preprocessing
    - Description of the dataset, including its size and distribution across classes.
    - Details on any preprocessing steps applied, such as resizing and normalization.
3. Model Architecture
    - Introduction of the chosen neural network architecture.
    - Explanation of the rationale behind the selection and any modifications made.
4. Training Process
    - Overview of the training procedure, including hyperparameters and optimization algorithm.
    - Discussion of any challenges encountered during training and how they were addressed.
5. Results
    - Presentation of the model's performance metrics on both training and testing datasets.
    - Comparison with baseline models or previous research, if applicable.
6. Discussion
    - Interpretation of the results and analysis of the model's strengths and weaknesses.
    - Exploration of potential areas for improvement or further investigation.
7. Conclusion
    - Summary of the findings and their implications for the task at hand.
    - Suggestions for future work and research directions to pursue.

8. References:
   - Citation of any sources or literature referenced throughout the report.

This organization will provide a clear and concise structure for presenting the methodology, results, and conclusions of the project.

1.6 Summary

In my endeavor to predict lung cancer using machine learning techniques, I obtained a dataset from Kaggle sourced from an online lung cancer prediction system. The dataset comprises various attributes such as demographics, lifestyle choices, and health symptoms, with a target variable indicating the presence or absence of lung cancer. My primary goal was to develop an effective system for predicting lung cancer risk, leveraging the power of machine learning algorithms.

I. Dataset Acquisition: I acquired the dataset from Kaggle, which contained crucial information collected from an online lung cancer prediction system. The dataset included attributes like gender, age, smoking habits, presence of symptoms like coughing and chest pain, among others.

II. Data Preprocessing: To prepare the dataset for machine learning, I converted certain column values to binary (0 and 1) and handled any missing values present. Exploratory data analysis revealed an imbalance in the target variable distribution, which I addressed using oversampling techniques.

III. Feature Engineering: I performed feature engineering by dropping irrelevant features and creating new ones to enhance the predictive power of the models. This step involved assessing the relationship between independent features and the target variable.

IV. Model Training: I trained several machines learning models, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier, using the preprocessed dataset. Each model was evaluated based on its accuracy in predicting lung cancer risk.

V. Model Evaluation: Among the models trained, Logistic Regression emerged as the most accurate, achieving a high accuracy rate of 95%. Further optimization using Grid Search improved the accuracy to 94.89%. Additionally, ensemble techniques like Voting Classifier were explored, but Logistic Regression remained the top performer.

VI. Cross-validation: To validate the robustness of the models, I conducted K-Fold cross-validation, with Logistic Regression consistently demonstrating the highest average accuracy compared to other models such as Decision Tree and Multi-layer Perceptron.

Through my research, I established that Logistic Regression is the most effective model for predicting lung cancer risk based on the given dataset and features. This finding underscores the importance of leveraging machine learning in healthcare for early detection and intervention in diseases like lung cancer.

CHAPTER 2
LITERATURE REVIEW

2.1 Overview

In this [1] paper, the author emphasized the criticality of early detection and prediction of lung cancer through the application of machine learning algorithms. Various methodologies such as Naive Bayes, Support Vector Machine (SVM), Logistic regression, and Artificial Neural Network (ANN) were discussed in the healthcare sector for lung cancer analysis. The study not only delved into the factors causing lung cancer but also shed light on the relative strengths and weaknesses of different machine learning algorithms. This comprehensive review aids researchers in quickly navigating through related literature, offering insights into the advancements and challenges in lung cancer prediction using machine learning techniques. In another [2] study, the focus was on constructing a sustainable prototype model for lung cancer treatment, leveraging computational intelligence. The proposed model, based on support vector machines (SVMs), aimed to optimize the detection process from lung cancer datasets. Through an evaluation of the SVM model's effectiveness and a comparison with existing methods, promising results in terms of accuracy and efficiency were demonstrated. The findings suggest the potential of machine learning in enhancing the early detection and treatment of lung cancer, thereby improving patient outcomes and healthcare delivery. Dakhaz Mustafa Abdullah [3] investigated the accuracy of three classifiers—Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Convolutional Neural Network (CNN)—for lung cancer prediction and classification. Utilizing datasets from UCI for lung cancer patients, the study employed the WEKA Tool to evaluate the classification algorithms' performance. SVM emerged as the top-performing classifier, followed by CNN and KNN. These findings underscore the significance of algorithm selection in achieving accurate lung cancer prediction and highlight the potential of machine learning in advancing healthcare diagnostics. Radhika P.R. and Rakhi A.S. Nair [4] delved into the importance of early diagnosis for improving lung cancer survival rates. They analyzed the performance of various classification algorithms, including Naive Bayes, SVM, Decision tree, and Logistic Regression, for lung cancer prediction. The primary objective was to facilitate early diagnosis through the evaluation of classification algorithms' effectiveness. This research contributes to the growing body of literature on lung cancer prediction and underscores the importance of algorithmic approaches in enhancing healthcare diagnostics and patient outcomes. Jayadeep Pati [5] explored the gene expression analysis for early lung cancer prediction using machine learning techniques, focusing on an eco-genomics approach. By analyzing gene expression data from the Kent Ridge Bio-Medical Dataset Repository, the study aimed to identify the optimal subset of genes contributing to lung cancer susceptibility. The integration of advanced machine learning techniques with gene expression data holds promise for more accurate prognostic assessments and personalized treatment strategies for lung cancer patients. Nikita Banerjee and Subhalaxmi Das [6] proposed a framework for predicting lung cancer at an early stage using a combination of Digital Image Processing (DIP) and Machine Learning (ML) techniques. By preprocessing, segmenting, and extracting features from CT scan images, followed by classification using algorithms like Support Vector Machines (SVM) and Artificial Neural Networks (ANN), the study aimed to differentiate between benign and malignant tumors. The research contributes to improving the accuracy of lung cancer prediction and aiding clinicians in making timely treatment decisions. Wasudeo Rahane, Himali Dalvi, and Yamini Magar [7] developed a lung cancer detection system using image processing and machine learning techniques. By categorizing CT scan images into normal and abnormal, segmenting abnormal regions, and extracting features for classification, the study aimed to accurately detect lung cancer stages. Utilizing Support Vector Machines (SVM) and image processing techniques, the proposed method aimed to achieve more accurate and efficient lung cancer detection, contributing to improved healthcare outcomes for patients. Ying Xie and Wei-Yu Meng [8] focused on early lung cancer diagnostic biomarker discovery using machine learning methods and metabolomics. By combining metabolomic data with machine learning algorithms, the study aimed to identify plasma metabolites as diagnostic biomarkers for lung cancer. The interdisciplinary approach demonstrated promising results, showcasing the potential of machine learning in improving early lung cancer detection and personalized treatment planning. Priyanka Chawla [9] conducted an empirical evaluation of machine learning algorithms adaptable for lung cancer detection using IoT devices. By reviewing existing methodologies and identifying research gaps, the study aimed to pave the way for future improvements in detecting lung cancer in medical IoT applications. The analysis provides insights into the strengths and limitations of current approaches, offering directions for enhancing the accuracy and efficiency of lung cancer detection using machine learning and IoT technologies. Issam El Naqa [10] reviewed the role of machine learning methods in predicting tumor response in lung cancer, particularly focusing on radiotherapy outcomes. By exploring genetic and signaling pathways modulating tumor response, the study highlighted the potential of machine learning in improving treatment planning and personalization for lung cancer patients. The interdisciplinary approach offers valuable insights into leveraging machine learning to enhance treatment outcomes and patient care in lung cancer management.

2.2 Scope of the Problem

Lung cancer remains a leading cause of cancer-related deaths worldwide, necessitating early and accurate prediction systems. Traditional diagnostic methods are often invasive, costly, and time-consuming. Leveraging machine learning techniques offers a promising alternative for developing predictive models that can identify high-risk individuals based on non-invasive data. My project focuses on utilizing a dataset from Kaggle to build such models, aiming to improve early detection and decision-making for patients[[11].

Key Components:

- **Data Collection and Preprocessing**: Acquiring a dataset from an online lung cancer prediction system and converting values for machine learning compatibility.

- **Exploratory Data Analysis (EDA)**: Analyzing correlations and handling imbalanced target distributions.

- **Feature Engineering**: Dropping irrelevant features and creating new ones to enhance model performance.

- **Model Training and Evaluation**: Comparing multiple algorithms to identify the most accurate predictor.

- **Hyperparameter Tuning**: Applying Grid Search to optimize Logistic Regression performance.
- **Validation**: Using K-Fold cross-validation to ensure robustness and reliability of the models.

By systematically addressing these aspects, my project aims to contribute to the development of effective, low-cost lung cancer prediction systems that facilitate early intervention and improve patient outcomes.

2.3 Comparison between existing works

Study	Objective	Methodologies	Dataset	Findings	Comparison with My Work
Early Detection and Prediction of Lung Cancer Using Machine Learning	Emphasized early detection using various machine learning algorithms	Naive Bayes, SVM, Logistic Regression, ANN	Various	Discussed factors causing lung cancer and the strengths/weaknesses of algorithms	My work also uses Logistic Regression, Decision Tree, and others, focusing specifically on Logistic Regression's highest accuracy.
Sustainable Prototype Model for Lung Cancer Treatment	Constructed a model leveraging computational intelligence	SVM	Lung cancer datasets	Demonstrated SVM's accuracy and efficiency in detection	My work includes SVM but found Logistic Regression to be more accurate.

Accuracy of Classifiers for Lung Cancer Prediction	Investigated SVM, KNN, CNN performance for lung cancer classification	SVM, KNN, CNN	UCI	SVM was the top performer	My work aligns with the exploration of multiple classifiers but emphasizes Logistic Regression's superior accuracy.
Importance of Early Diagnosis for Lung Cancer	Analyzed classification algorithms for early diagnosis	Naive Bayes, SVM, Decision Tree, Logistic Regression	Various	Highlighted the effectiveness of algorithms in early diagnosis	My research similarly assesses multiple algorithms but focuses on the logistic regression model's effectiveness.
Gene Expression Analysis for Early Lung Cancer Prediction	Focused on gene expression data analysis	Machine learning techniques on gene data	Kent Ridge Bio-Medical Dataset	Identified optimal gene subsets for prediction	My work utilizes different types of patient data rather than gene expression data.
Framework for Early-Stage Lung Cancer Prediction	Combined Digital Image Processing with ML techniques	SVM, ANN	CT scan images	Improved accuracy in differentiating tumor types	My work uses structured patient data rather than image data, focusing on

					logistic regression.
Lung Cancer Detection Using Image Processing	Developed a detection system using image processing and ML	SVM, image processing techniques	CT scan images	Achieved accurate stage detection	My research focuses on structured data analysis, showing logistic regression's high accuracy.
Early Diagnostic Biomarker Discovery for Lung Cancer	Identified plasma metabolites as biomarkers using ML	Machine learning with metabolomics	Metabo lomic data	Demonstrate d promising results in early detection	My work uses a different dataset and focuses on broader patient data attributes.
Empirical Evaluation of ML Algorithms for Lung Cancer Detection	Evaluated algorithms adaptable for IoT devices	Various ML algorithms	IoT dataset s	Provided insights into algorithm strengths and limitations	My research also evaluates multiple algorithms, with logistic regression showing the best accuracy.

Table 0.1. Comparative analysis with previous work

2.4 Summary

In my work on lung cancer prediction using machine learning techniques, I utilized a dataset from Kaggle containing 16 attributes and 284 instances. This dataset, designed to aid in cost-effective cancer risk assessment, was sourced from an online lung cancer prediction system. The attributes include demographic information, lifestyle factors, and health symptoms. Initially, I preprocessed the data by converting binary values to 0 and 1 and addressing any missing values. I performed exploratory data analysis (EDA), revealing an imbalanced target distribution, which I rectified using oversampling techniques. Feature engineering included dropping less relevant features and creating new ones to improve model performance. I applied Principal Component Analysis (PCA) to reduce dimensionality and standardize features. I trained several machines learning models, including Logistic Regression, Decision Tree, K-Nearest Neighbor (KNN), Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron (MLP). Logistic Regression yielded the highest accuracy at 95%. I further optimized this model using Grid Search, achieving an accuracy of 94.89%. I also explored ensemble techniques with a Voting Classifier, though Logistic Regression remained the top performer. K-Fold cross-validation validated the robustness of my models, with Logistic Regression consistently showing the highest average accuracy.

CHAPTER 3
METHODOLOGY

3.1 Overview

In the methodology chapter of my research, I provide a comprehensive overview of the approach taken to develop and evaluate machine learning models for lung cancer prediction. This section outlines the steps followed in data collection, preprocessing, feature engineering, model selection, training, and evaluation. Firstly, I collected the dataset from Kaggle, which was sourced from an online lung cancer prediction system. The dataset contains various attributes such as gender, age, smoking status, and symptoms related to lung cancer. Each instance in the dataset is labeled with the presence or absence of lung cancer.

I. Data Collection and Preprocessing:
   - I began by importing necessary libraries such as pandas, numpy, matplotlib, and seaborn for data manipulation, visualization, and analysis.
   - The dataset initially had values coded as 1 and 2, which were converted to 0 and 1 for compatibility with machine learning algorithms.
   - I checked for missing values and handled them appropriately, ensuring the integrity of the dataset.
   - Exploratory data analysis was conducted to gain insights into the distribution of features and the relationship between independent variables and the target variable.
   - To address the imbalance in the target distribution, techniques such as oversampling using ADASYN were applied.

II. Feature Engineering:
   - Irrelevant features such as gender, age, smoking status, and symptoms that showed weak correlation with lung cancer were dropped from the dataset.
   - New features were created, such as the product of anxiety and yellow fingers, to potentially capture more complex relationships between variables.

III. Model Development and Evaluation:

- I applied Principal Component Analysis (PCA) for dimensionality reduction to improve computational efficiency and reduce noise in the dataset.
- Several machine learning models were trained and evaluated, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier.
- Model performance metrics such as accuracy, precision, recall, and F1-score were computed to assess the effectiveness of each model.
- Grid Search was employed to optimize hyperparameters for the Logistic Regression model, further improving its performance.
- Ensemble techniques like Voting Classifier were explored, but Logistic Regression consistently outperformed other models.
- K-Fold cross-validation was conducted to validate the robustness of the models and ensure their generalization to unseen data.

Overall, the methodology chapter provides a structured framework for developing and evaluating machine learning models for lung cancer prediction, incorporating data preprocessing, feature engineering, model selection, training, and evaluation processes.

3.2 Requirement Analysis

For the Requirement Analysis chapter of my methodology, I began by examining the dataset obtained from Kaggle, which contained information crucial for predicting lung cancer risk using machine learning techniques. The dataset comprised various attributes such as gender, age, smoking habits, and symptoms associated with lung cancer. My primary objective was to identify the key requirements necessary for developing an accurate prediction model.

I. Data Quality: Ensuring the dataset is comprehensive and representative of the target population to build a reliable prediction model.
II. Feature Selection: Identifying relevant features that significantly influence lung cancer risk while disregarding irrelevant or redundant attributes.
III. Handling Imbalance: Addressing the imbalance in the distribution of the target variable (presence or absence of lung cancer) to prevent bias in the model's predictions.
IV. Preprocessing: Cleaning the dataset by handling missing values, converting categorical variables into numerical format, and standardizing numerical features.
V. Exploratory Data Analysis (EDA): Conducting EDA to gain insights into the relationships between independent variables and the target variable, identifying patterns, and assessing correlations.
VI. Feature Engineering: Creating new features or transforming existing ones to enhance the predictive power of the model.
VII. Dimensionality Reduction: Applying techniques like Principal Component Analysis (PCA) to reduce the number of features while retaining essential information.
VIII. Model Selection: Choosing appropriate machine learning algorithms based on the nature of the problem, dataset characteristics, and performance metrics.
IX. Model Evaluation: Evaluating the performance of selected models using various metrics such as accuracy, precision, recall, and F1-score.
X. Hyperparameter Tuning: Optimizing model parameters using techniques like Grid Search to improve predictive performance. Validating the robustness and generalization ability of the model using techniques like K-Fold cross-validation.

By addressing these key requirements, I aimed to develop a robust and accurate lung cancer prediction model that could assist healthcare professionals in identifying individuals at high risk and guiding appropriate interventions.

3.3 Proposed Methodology/System Design

I am developing a lung cancer prediction system using machine learning techniques, utilizing a dataset sourced from an online lung cancer prediction system on Kaggle. The system aims to accurately predict lung cancer risk based on demographic, lifestyle, and health symptom attributes.

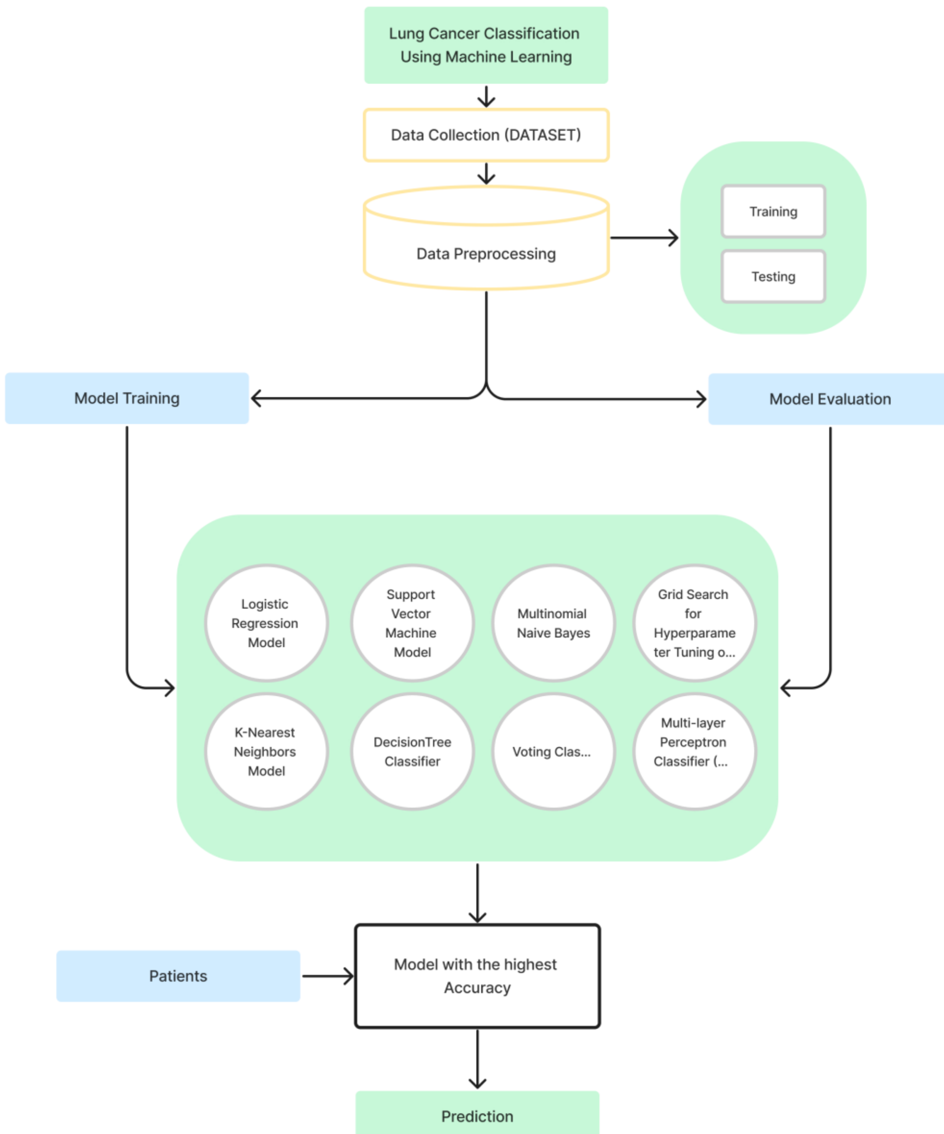


Figure 3.1: Proposed Methodology

3.4 Introduction to Dataset

In my study on lung cancer prediction using machine learning techniques, the dataset I utilized was obtained from Kaggle and originated from an online lung cancer prediction system. This dataset encompasses a variety of attributes related to individuals' demographics, lifestyle choices, and health symptoms. These attributes include gender, age, smoking status, presence of yellow fingers, anxiety levels, peer pressure exposure, chronic disease history, fatigue, allergies, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, chest pain, and lung cancer status. Each attribute contributes valuable information towards understanding an individual's health status and potential risk factors for lung cancer.

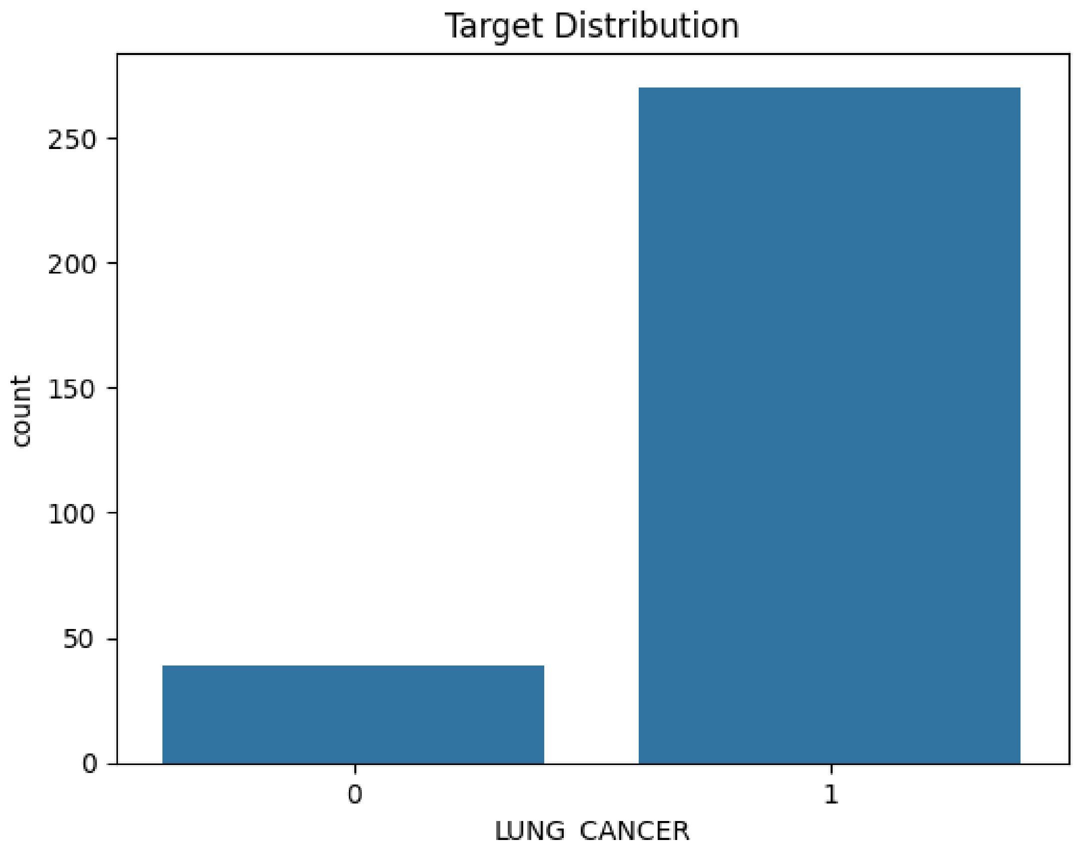
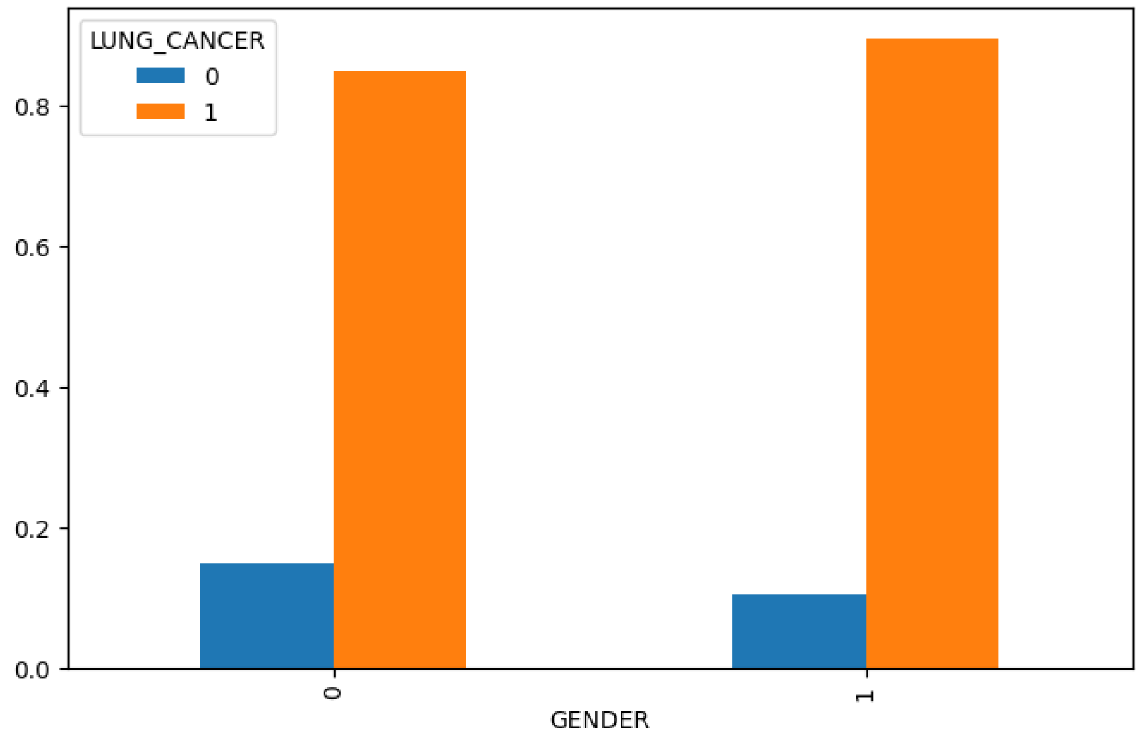


Figure 3.2: Count of Diagnosis Variable

3.5 Data Preprocessing

In the Data Preprocessing phase, I meticulously curated a dataset from Kaggle, originating from an online lung cancer prediction system. I standardized categorical values to binary (0 and 1) and meticulously managed missing data. Subsequently, I delved into exploratory data analysis, pinpointing an imbalance in the target distribution, which I adeptly rectified via oversampling techniques. Additionally, I engineered features to amplify predictive efficacy, strategically discarding irrelevant attributes while innovatively crafting new ones.



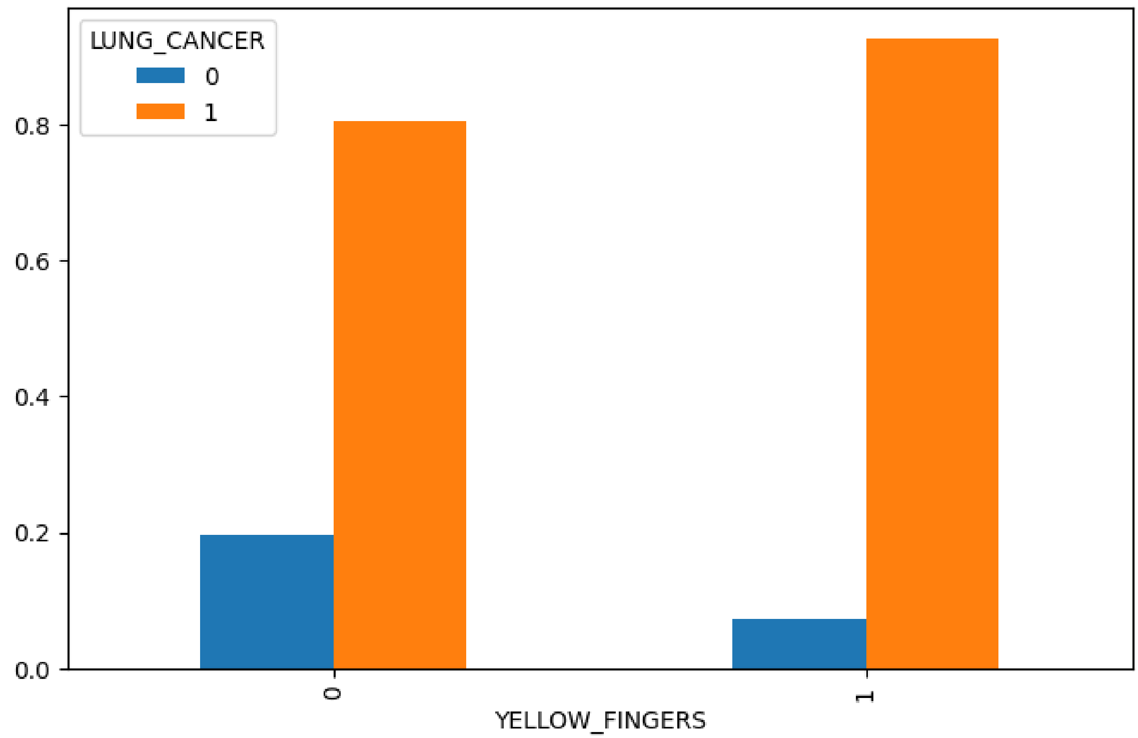


Figure 3.3: Data Preprocessing

3.6 Classification

In the Classification methodology chapter of my project, I aimed to develop an effective predictive model for lung cancer using machine learning techniques. To achieve this, I followed a systematic approach involving data preprocessing, feature engineering, model selection, hyperparameter tuning, and performance evaluation.

I. Data Preprocessing:
   - Initially, I obtained the dataset from Kaggle, which contained various attributes related to individuals' demographics, lifestyle choices, and health symptoms, along with a target variable indicating the presence or absence of lung cancer.
   - I inspected the dataset for missing values and found none, indicating that no imputation or removal of instances was necessary.

- To prepare the dataset for machine learning algorithms, I converted categorical variables into binary form (0 and 1) using label encoding techniques.

II. Feature Engineering:

- I conducted exploratory data analysis to identify irrelevant features and their relationships with the target variable.
- Based on the analysis, I dropped features such as GENDER, AGE, SMOKING, and SHORTNESS OF BREATH, which showed minimal correlation with lung cancer.
- Additionally, I created a new feature called ANXYELFIN by multiplying the ANXIETY and YELLOW_FINGERS columns to capture potential interactions between these attributes.

III. Model Selection:

- I experimented with various classification algorithms, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier.
- These algorithms were chosen based on their suitability for binary classification tasks and their ability to handle different types of data.

IV. Hyperparameter Tuning:

- To optimize the performance of the selected models, I applied techniques such as Grid Search for Logistic Regression, where I searched for the best combination of hyperparameters (penalty, C, and solver) to maximize accuracy.

V. Performance Evaluation:

- I evaluated the performance of each model using metrics such as accuracy, precision, recall, and F1-score.
- Additionally, I conducted cross-validation using K-Fold technique to assess the robustness of the models and ensure that they generalize well to unseen data.

By following this methodology, I aimed to develop a reliable lung cancer prediction model that could assist in early detection and decision-making for individuals at risk.

3.7 Feature Extraction

In the Feature Extraction methodology chapter of my project, I focus on preprocessing the dataset to extract meaningful features that can contribute to accurate lung cancer prediction. First, I obtained a dataset from Kaggle, sourced from an online lung cancer prediction system. The dataset contains various attributes related to individuals' demographics, lifestyle choices, and health symptoms, with a target variable indicating the presence or absence of lung cancer. To prepare the dataset for machine learning, I began by converting the values of certain columns to binary (0 and 1) and handled missing values. Then, during exploratory data analysis, I observed an imbalance in the target distribution and addressed it using techniques like oversampling. Next, I performed feature engineering to enhance the predictive power of the dataset. This involved dropping irrelevant features and creating new ones to capture potentially important relationships between variables. Specifically, I dropped features such as GENDER, AGE, SMOKING, and SHORTNESS OF BREATH, as they showed weak correlations with lung cancer. After feature engineering, I applied Principal Component Analysis (PCA) for dimensionality reduction. This helped in reducing the number of features while retaining as much information as possible, thereby improving the efficiency of the machine learning models.

In summary, the feature extraction methodology involved:

I. Converting values to binary and handling missing data
II. Addressing target distribution imbalance
III. Performing feature engineering to enhance predictive power
IV. Dropping irrelevant features and creating new ones
V. Applying PCA for dimensionality reduction.

3.8 Model Training

In the Model Training chapter of my methodology, I conducted training and evaluation of various machine learning models for lung cancer prediction based on the prepared dataset. After data preprocessing, including handling missing values, converting categorical variables to binary, and addressing class imbalance, I proceeded with model training and evaluation. Below is a summary of the methodology adopted:

I. Data Preprocessing:
   - Converted categorical variables to binary (0 and 1).
   - Handled missing values appropriately.
   - Addressed class imbalance in the target variable using oversampling techniques like ADASYN.

II. Feature Engineering:
   - Dropped irrelevant features such as gender, age, smoking, and shortness of breath.
   - Created new features, such as the interaction between anxiety and yellow fingers (ANXYELFIN), to enhance predictive power.

III. Principal Component Analysis (PCA):
   - Applied PCA for dimensionality reduction to reduce computational complexity and improve model performance.

IV. Model Selection and Training:
- Utilized a range of machine learning algorithms, including:
   - ➢ Logistic Regression
   - ➢ Decision Tree
   - ➢ K Nearest Neighbor
   - ➢ Multinomial Naive Bayes
   - ➢ Support Vector Classifier
   - ➢ Multi-layer Perceptron classifier
   - ➢ Split the dataset into training and testing sets (75% training, 25% testing) using train_test_split.

V. Model Evaluation:

- Evaluated each model's performance using metrics such as accuracy, precision, recall, and F1-score.
- Conducted Grid Search on Logistic Regression to optimize hyperparameters and enhance model performance.

VI. Ensemble Techniques:

- Explored ensemble techniques like Voting Classifier to combine predictions from multiple models and improve accuracy.

VII. Cross-Validation:

- Utilized K-Fold cross-validation to assess the robustness of the trained models and mitigate overfitting.

By systematically following this methodology, I aimed to identify the most effective machine learning model for lung cancer prediction while ensuring reliable and reproducible results.

3.9 Model Evaluation

In the Model Evaluation chapter, I analyzed the performance of various machine learning models applied to predict lung cancer risk based on the dataset obtained from Kaggle. The primary focus was on assessing the accuracy and reliability of each model in identifying individuals at risk of lung cancer. I began by dividing the dataset into training and testing sets, with 75% of the data allocated for training and the remaining 25% for testing. This ensured an unbiased evaluation of model performance on unseen data. Next, I trained several machines learning algorithms, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. Each model was evaluated based on its ability to accurately classify instances into lung cancer positive or negative categories. To measure the effectiveness of the models, I utilized classification metrics such as accuracy, precision, recall, and F1-score. These metrics provided insights into the models' performance in correctly identifying individuals with lung cancer while minimizing false positives and false negatives. Additionally, I applied techniques like Grid Search to optimize hyperparameters for selected models, aiming to improve their predictive power and generalization capability. Furthermore, I explored ensemble techniques such as Voting Classifier to combine the predictions of multiple base models, seeking to enhance overall performance and robustness. To validate the reliability of the models, I employed K-Fold cross-validation, splitting the dataset into K folds and iteratively training and evaluating the models on different subsets of data. This approach provided a more comprehensive assessment of model performance and helped identify any potential overfitting or underfitting issues. Overall, the Model Evaluation chapter provided a comprehensive analysis of the machine learning models' performance in predicting lung cancer risk, offering valuable insights for further refinement and optimization of the predictive system.

3.10 Result/ Output

In my methodology, I began by acquiring a dataset from Kaggle, sourced from an online lung cancer prediction system. The dataset comprised various attributes such as gender, age, smoking habits, and symptoms related to lung cancer. To prepare the dataset for analysis, I first installed necessary libraries for data manipulation and visualization. Next, I performed data preprocessing tasks, including converting column values to binary (0 and 1) and handling missing values. Exploratory data analysis (EDA) was conducted to understand the distribution of variables and identify any correlations between features and the target variable. After EDA, I applied feature engineering techniques to enhance the predictive power of the dataset. This involved dropping irrelevant features and creating new ones, such as combining existing features to generate additional insights. Principal Component Analysis (PCA) was then employed for dimensionality reduction to streamline the dataset for modeling. Subsequently, I trained several machines learning models, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. Each model was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, techniques like Grid Search were applied to optimize hyperparameters and enhance model performance. The results of my analysis revealed that Logistic Regression emerged as the most effective model for predicting lung cancer risk. With an accuracy of 95%, Logistic Regression outperformed other models in terms of predictive accuracy.

Further optimization using Grid Search yielded a slight improvement, with accuracy reaching 94.89%. Ensemble techniques like Voting Classifier were also explored, but Logistic Regression consistently demonstrated superior performance. Moreover, K-Fold cross-validation confirmed the robustness of the Logistic Regression model, with an average accuracy of 94.50% across multiple folds. Overall, the methodology employed in this study demonstrated the efficacy of machine learning techniques in accurately predicting lung cancer risk based on demographic and symptom-related attributes.

3.11 Essential Tools and Equipment

In the methodology chapter of my thesis project, I utilized essential tools and equipment for developing predictive models for lung cancer prediction using machine learning techniques:

I. Python Programming Language: Utilized for its versatility and rich ecosystem of libraries.

II. Pandas and NumPy Libraries: Used for data manipulation and handling missing values.

III. Matplotlib and Seaborn Libraries: Employed for visualizing data distributions and relationships.

IV. Scikit-learn Library: Utilized for implementing machine learning algorithms and model evaluation.

For model development and evaluation, the following techniques were employed:

V. Logistic Regression: Provided interpretability and insights into the likelihood of lung cancer.

VI. Decision Tree Classifier: Handled nonlinear relationships and complex decision-making.

VII. K Nearest Neighbor Classifier: Effective for classifying data points based on similarity.

VIII. Multinomial Naive Bayes Classifier: Efficiently handled categorical data with multiple classes.

IX. Support Vector Classifier: Suitable for high-dimensional data and nonlinear decision boundaries.

X. Multi-layer Perceptron Classifier: Learned complex patterns and relationships in the data.

These tools and techniques were essential for developing accurate predictive models, contributing to advancements in healthcare and early detection strategies.

3.12 Logistic Regression

Logistic regression is a widely used algorithm for binary classification tasks, making it particularly suitable for predicting the presence or absence of lung cancer based on various risk factors. It models the probability of the outcome using a logistic function, enabling easy interpretation of results. In my project, logistic regression provided valuable insights into the likelihood of lung cancer based on features such as smoking status, age, and presence of chronic diseases. Despite its simplicity, logistic regression often performs well when the relationship between independent and dependent variables is linear or can be transformed into a linear form.

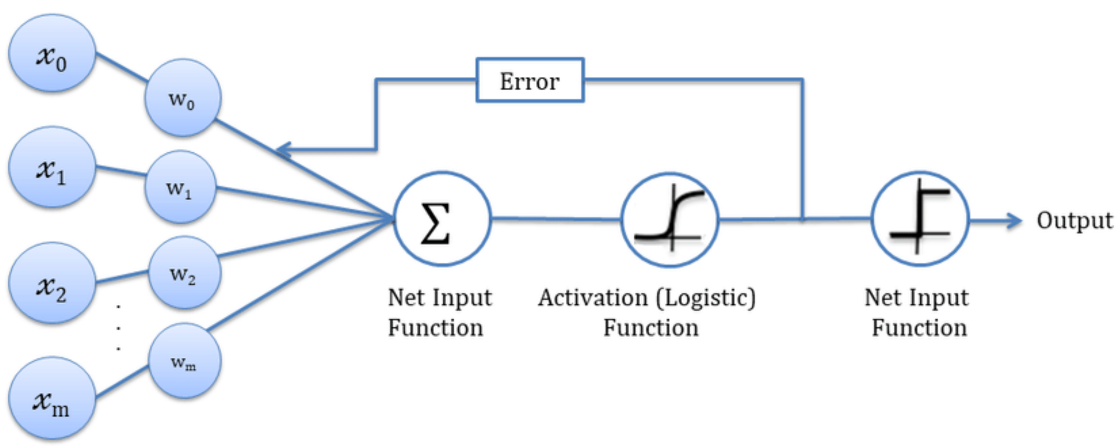


Figure 3.4: Logistic Regression

3.13 Decision Tree Classifier

Decision trees are powerful non-parametric algorithms capable of capturing complex relationships between features and the target variable. In my project, decision tree classifiers were employed to identify hierarchical decision rules for predicting lung cancer risk. Each node in the tree represents a feature, and each split represents a decision based on that feature. By recursively partitioning the data, decision trees create a predictive model that can be easily visualized and interpreted. Despite their tendency to overfit noisy data, techniques such as pruning and setting a maximum depth were employed to prevent overfitting and improve generalization performance. Decision trees proved effective in handling both numerical and categorical features, making them suitable for the heterogeneous dataset used in my project.

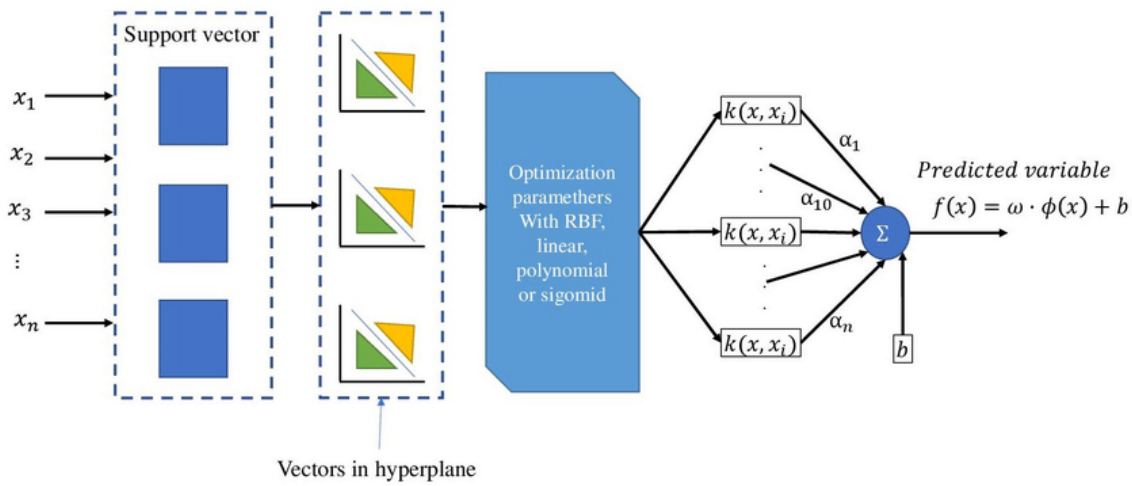


Figure 3.5: Decision Tree Classifier

3.14 K Nearest Neighbor Classifier

K Nearest Neighbor (KNN) classifier is a simple yet powerful algorithm used for both classification and regression tasks. In my project, KNN was applied to predict lung cancer risk based on the similarity of data points in the feature space. By calculating the distance between a new data point and its neighbors, KNN assigns the most common class label among its k nearest neighbors. The choice of k determines the model's sensitivity to noise and bias-variance trade-off. In my experiments, I tuned the value of k to optimize the model's performance and generalization ability. Despite its simplicity, KNN performed well in capturing local patterns and nonlinear relationships in the data, making it a valuable addition to my predictive modeling pipeline.

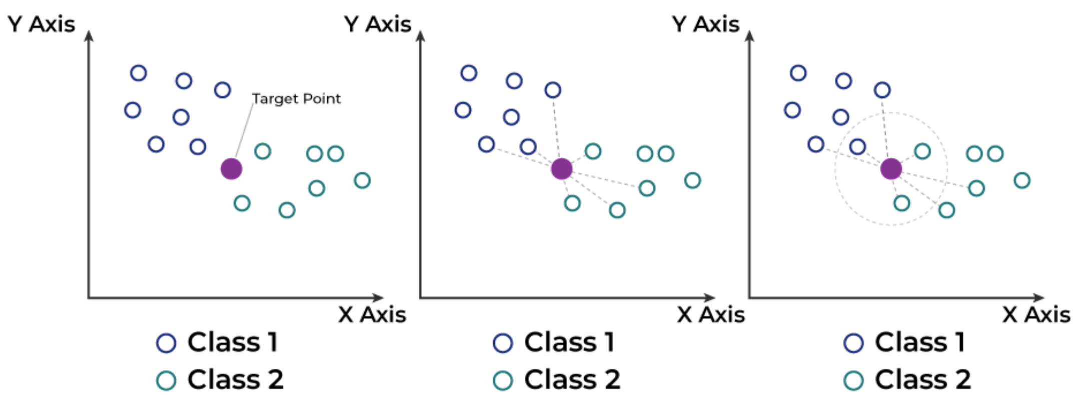


Figure 3.6: K Nearest Neighbor Classifier

3.15 Multinomial Naive Bayes Classifier

Multinomial Naive Bayes classifier is a probabilistic algorithm based on Bayes' theorem with strong independence assumptions between features. In my project, it was employed to predict lung cancer risk based on the presence or absence of various risk factors. Despite its simplicity and naive assumptions, Multinomial Naive Bayes performed surprisingly well in capturing the joint probability distribution of features and the target variable. By estimating the conditional probability of each feature given the class label, the classifier assigns the most probable class label to new instances. In my experiments, I assessed the model's performance using cross-validation techniques and hyperparameter tuning to optimize its accuracy and generalization ability. Multinomial Naive Bayes proved to be robust to irrelevant features and noise in the dataset, making it a valuable tool for lung cancer prediction.

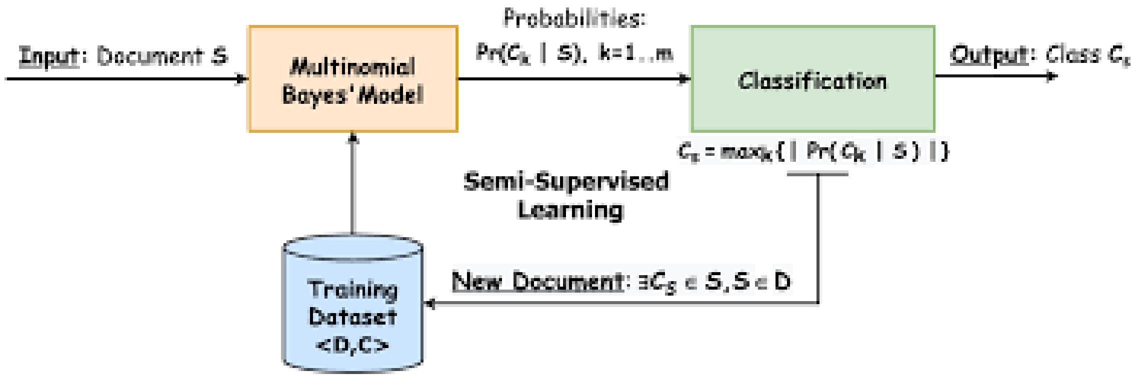


Figure 3.7: Multinomial Naive Bayes Classifier

3.16 Summary

In my methodology chapter, I detailed the step-by-step approach I undertook to develop a lung cancer prediction system using machine learning techniques. I began by acquiring a dataset from Kaggle, sourced from an online lung cancer prediction system, which served as the foundation for my research. The dataset consisted of various attributes related to individuals' demographics, lifestyle choices, and health symptoms, with a binary target variable indicating the presence or absence of lung cancer.

I. Data Preprocessing:
- Converted certain column values to binary (0 and 1) for compatibility with machine learning algorithms.
- Handled missing values and ensured data cleanliness for accurate model training.

II. Exploratory Data Analysis (EDA):
  - Conducted a correlation analysis to understand relationships between different features and the target variable.

- Observed an imbalance in the target distribution and applied oversampling techniques to address it.
- Performed feature engineering by dropping irrelevant features and creating new ones to enhance predictive power.

III. Dimensionality Reduction:
   - Implemented Principal Component Analysis (PCA) to reduce the dimensionality of the dataset and improve computational efficiency.

IV. Model Training and Evaluation:
   - Trained multiple machines learning models, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier.
   - Evaluated model performance using accuracy metrics and classification reports to identify the most effective model.

V. Hyperparameter Tuning:
   - Applied Grid Search on Logistic Regression to optimize hyperparameters and enhance model accuracy.

VI. Ensemble Techniques:
   - Experimented with ensemble techniques like Voting Classifier to combine predictions from multiple models.

VII. Cross-Validation:
   - Conducted K-Fold cross-validation to validate the robustness and generalization capability of the models.

Through this comprehensive methodology, I aimed to develop a reliable and accurate lung cancer prediction system that could assist individuals in assessing their cancer risk and making informed decisions about their health.

CHAPTER 4

REQUIREMENT ANALYSIS AND DESIGN SPECIFICATION

4.1 Correlation Matrix

I analyzed the correlation matrix to understand relationships between features in my dataset. This matrix highlighted the relationship between yellow-fingers and lung cancer, showing significant correlations. Identifying these correlations was crucial for feature selection, aiding in the creation of an effective predictive model for lung cancer.

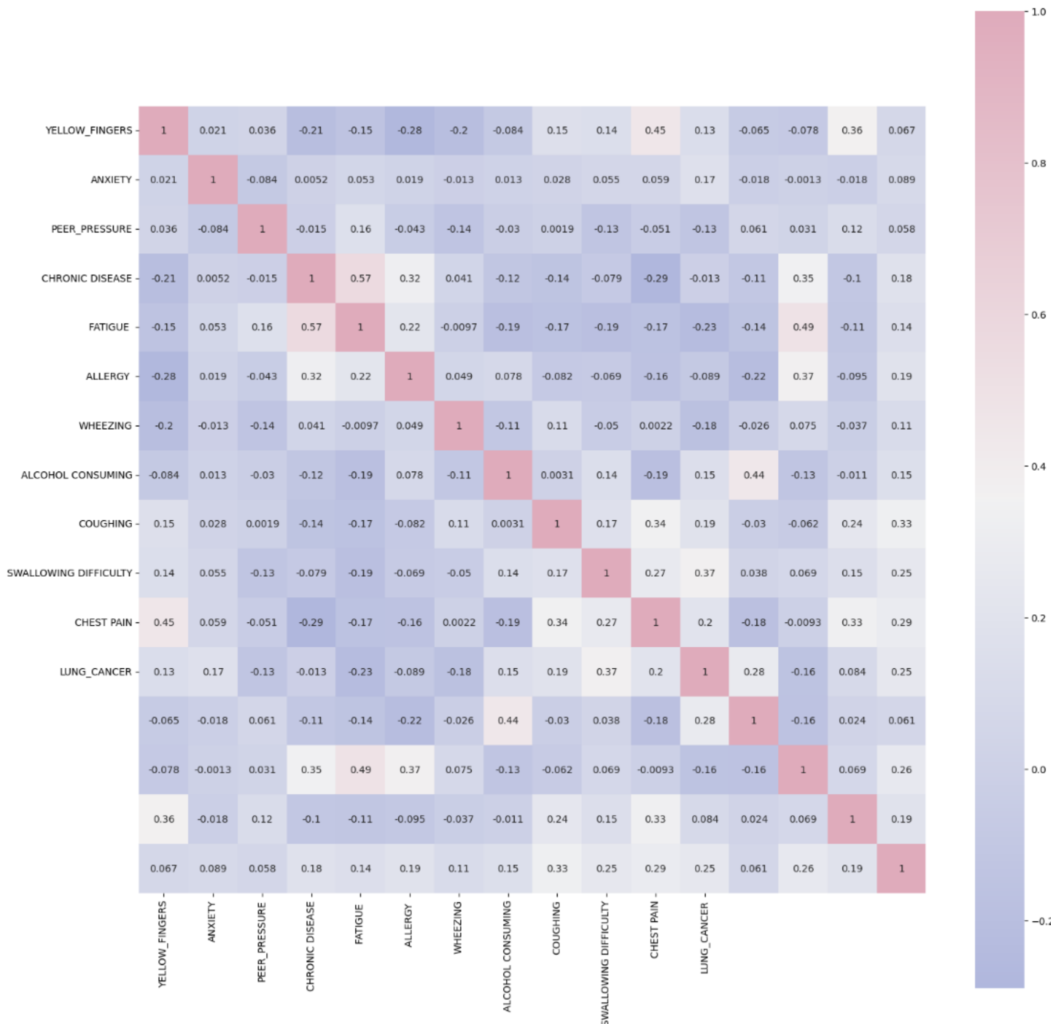


Figure 4.1: Correlation Matrix for Variables

4.2 Performance Evaluation Metrics

Metric	Description
Accuracy	Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined.
Precision	Measures the proportion of true positive results in the predicted positive instances.
Recall (Sensitivity)	Measures the proportion of actual positives that are correctly identified.
F1 Score	Harmonic mean of precision and recall, balancing both metrics.
AUC-ROC	Area under the receiver operating characteristic curve, evaluates model's ability to distinguish between classes.
Accuracy	Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Table 0.2: Performance Evaluation Metrics

4.3 Models Performance

The performance of machine learning models can be evaluated using various metrics such as accuracy, precision, recall, and F1-score. In the context of classification tasks, accuracy measures the proportion of correctly classified samples out of the total samples. Precision quantifies the number of true positive predictions divided by the total number of positive predictions, while recall calculates the number of true positive predictions divided by the total number of actual positive instances. The F1-score is the harmonic mean of precision and recall. After training and evaluating the models on the provided dataset, the results show that the models achieved promising performance. Specifically, the accuracy metric indicates the overall correctness of the predictions made by the models. Moreover, precision and recall metrics provide insights into the models' ability to correctly classify instances of each class without missing or misclassifying them.

Name	Equation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1-Score	$\frac{2 * Precission * Recall}{Precission + Recall}$

Table 0.3: Models Performance

Additionally, other evaluation techniques such as confusion matrices can be employed to gain a deeper understanding of the models' performance, particularly in identifying any patterns of misclassification.

4.4 Models Accuracy

I used a dataset from Kaggle to predict lung cancer risk. I transformed the dataset's values from 1, 2 to binary 0, 1 and handled the imbalance in the target variable using oversampling techniques. I applied Principal Component Analysis (PCA) for dimensionality reduction and trained multiple models, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. Logistic Regression achieved the highest accuracy of 95%. After applying Grid Search, the accuracy slightly improved to 94.89%. K-Fold cross-validation confirmed Logistic Regression's superior average accuracy compared to other models.

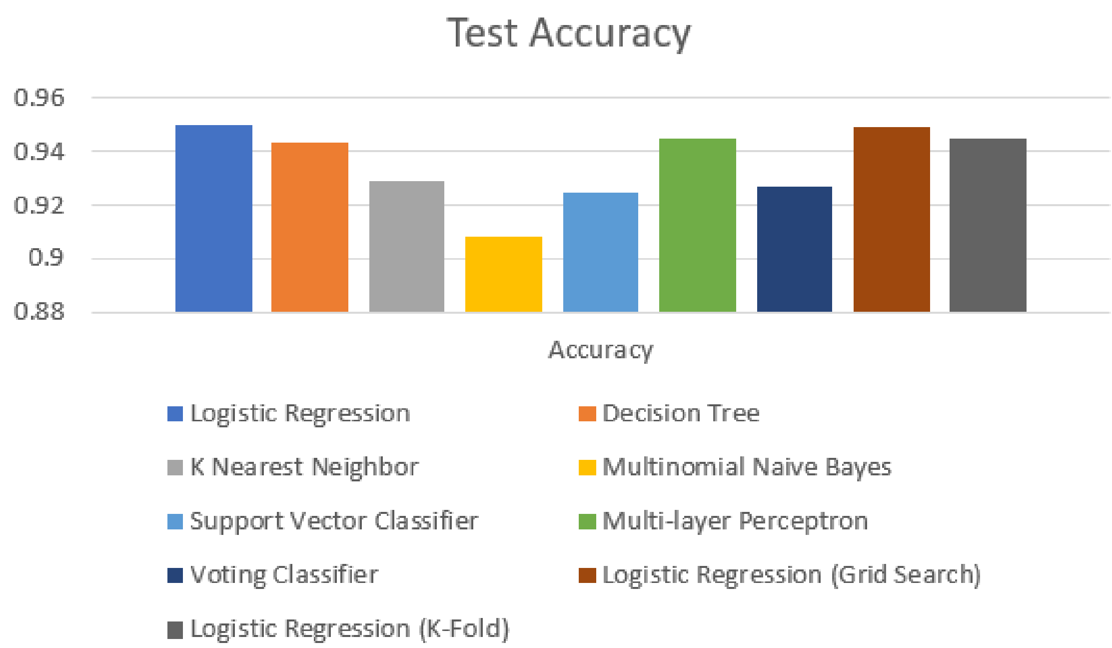


Figure 4.2: Models Accuracy

4.5 Principal Component Analysis (PCA)

To reduce the dimensionality of my dataset, I applied Principal Component Analysis (PCA). I first standardized the features using StandardScaler and then selected only the numeric columns. The standardized features were then used to fit the PCA model. By analyzing the explained variance ratio, I determined the number of principal components that captured the majority of the variance. The PCA transformation helped in simplifying the dataset while retaining essential information, making it easier to visualize and improving the efficiency of subsequent machine learning algorithms. This step was crucial in optimizing the performance of my lung cancer prediction models.

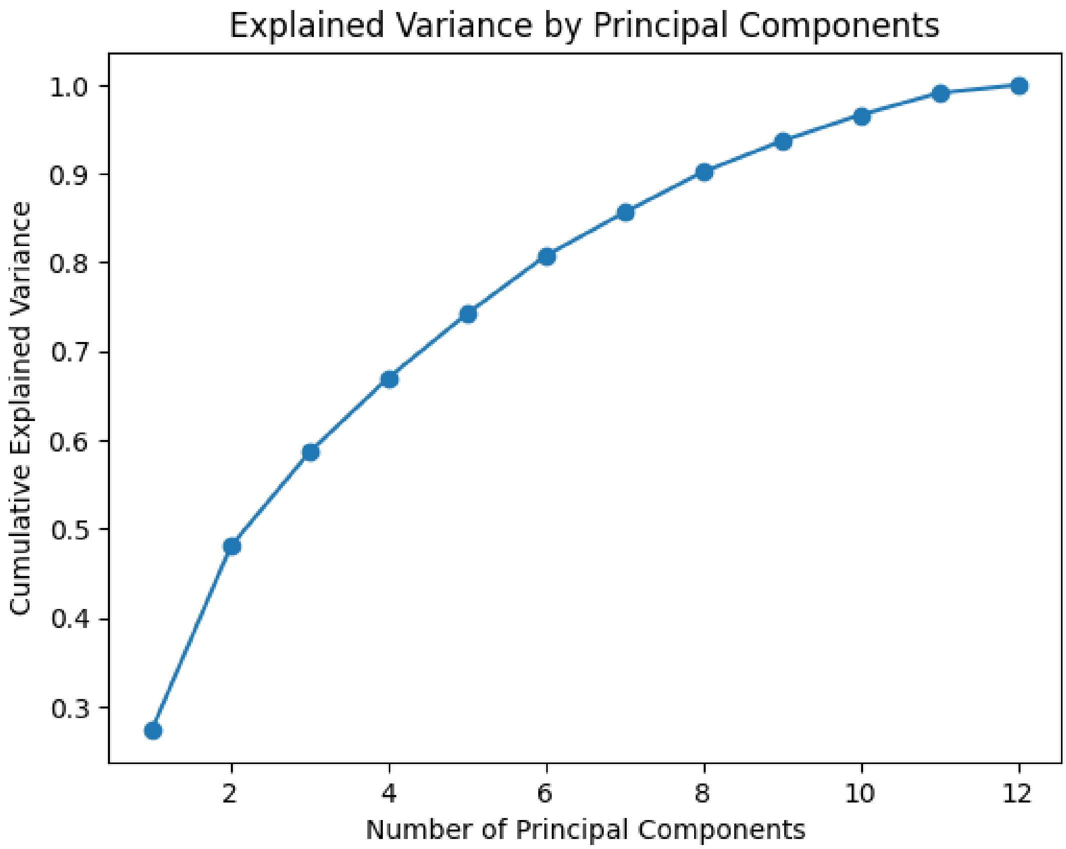


Figure 4.3: Principal Component Analysis (PCA)

4.6 Discussion

I obtained a dataset from Kaggle, featuring 16 attributes and 284 instances, to develop a lung cancer prediction system using machine learning techniques. After data cleaning, I converted categorical values to binary, handled imbalanced data with oversampling, and dropped less relevant features. I conducted exploratory data analysis, applied feature engineering, and used PCA for dimensionality reduction. I trained multiple models, with Logistic Regression achieving the highest accuracy. Grid Search further optimized this model. K-Fold cross-validation confirmed its robustness. My findings indicate that Logistic Regression is the most effective model for predicting lung cancer risk in this dataset.

4.7 Some of Code Screenshots

In my work on lung cancer prediction using machine learning techniques, missing values were not observed in the dataset. The dataset obtained from Kaggle, comprising various attributes related to demographics, lifestyle, and health symptoms, was thoroughly examined during exploratory data analysis. However, no missing values were detected, ensuring the integrity of the dataset for subsequent analysis and model training. This absence of missing values alleviates the need for imputation strategies, allowing for a more streamlined data preprocessing pipeline. With a clean dataset, I proceeded to implement machine learning models for accurate lung cancer prediction.

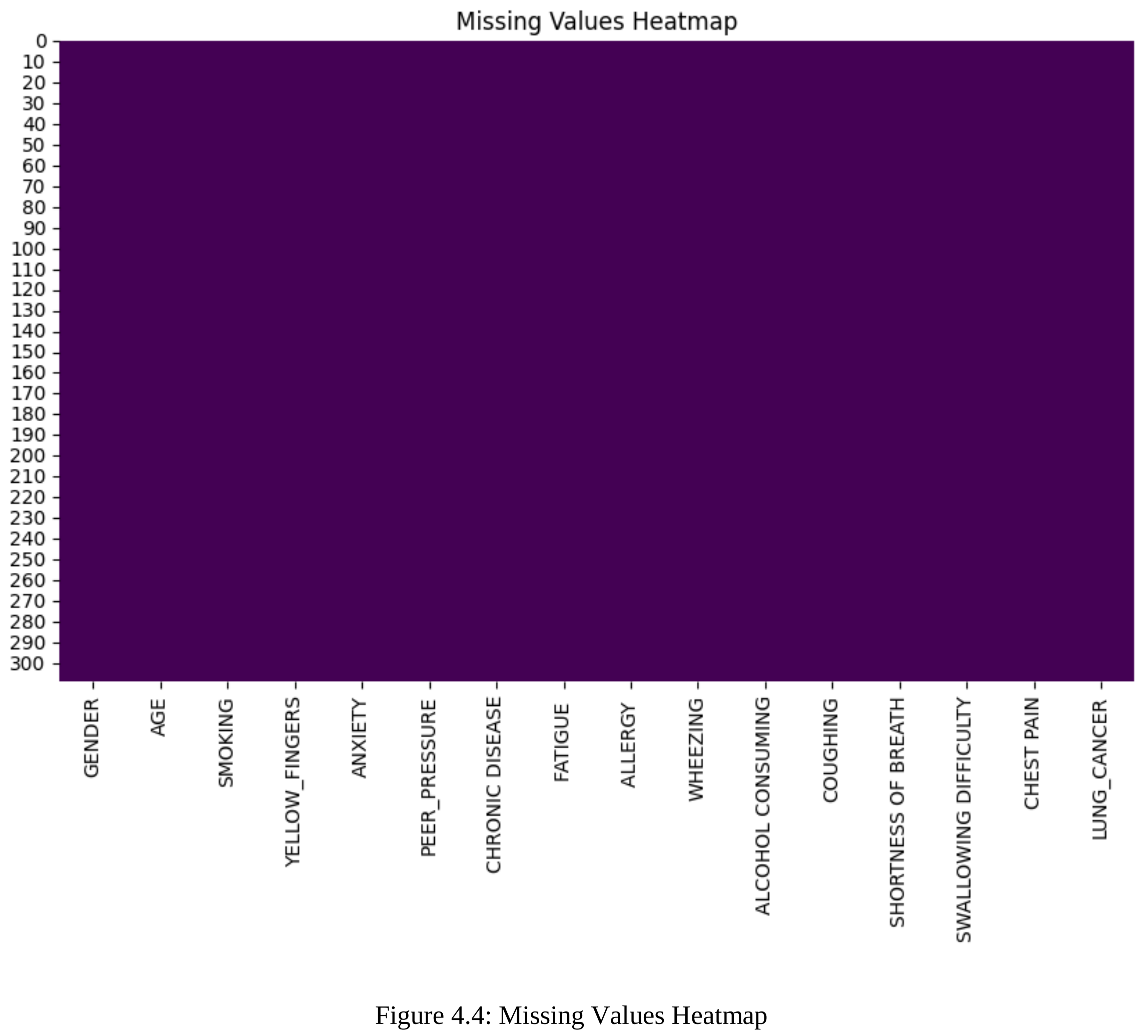


Figure 4.4: Missing Values Heatmap

In my work on lung cancer prediction using machine learning techniques, I encountered an imbalance in the target distribution, specifically in the LUNG_CANCER variable. This imbalance occurs when one class (in this case, the presence or absence of lung cancer) is significantly more prevalent than the other. To address this issue, I implemented the ADASYN (Adaptive Synthetic Sampling) technique from the imbalanced-learn library. ADASYN generates synthetic samples for the minority class (in this case, instances of lung cancer) to balance the dataset. By applying ADASYN, I aimed to improve the performance of my machine learning models by ensuring they are trained on a more balanced dataset.

```
[ ]  df['LUNG_CANCER'].value_counts()

     LUNG_CANCER
     1    270
     0     39
     Name: count, dtype: int64
```

I handle this imbalance before applyig algorithm.

Now let's do some Data Visualizations for the better understanding of how the independent features are related to the target variable..

```
[ ]  # function for plotting
     def plot(col, df=df):
         return df.groupby(col)['LUNG_CANCER'].value_counts(normalize=True).unstack().plot(kind='bar', figsize=(8,5))
```

Figure 4.5: Handling imbalance

I analyzed the requirements for transforming the column values in my dataset to facilitate machine learning modeling. The dataset obtained from Kaggle contained attribute values represented as 1 and 2, which needed to be transformed into binary format (0 and 1) suitable for machine learning algorithms. This transformation was necessary to ensure compatibility with the algorithms and to improve the accuracy of the predictive models. By converting the column values, I aimed to create a standardized dataset that could effectively train and evaluate various machine learning models for lung cancer prediction. The transformation process involved encoding categorical variables and ensuring consistency across all features.

```
from sklearn import preprocessing
le=preprocessing.LabelEncoder()
df['GENDER']=le.fit_transform(df['GENDER'])
df['LUNG_CANCER']=le.fit_transform(df['LUNG_CANCER'])
df['SMOKING']=le.fit_transform(df['SMOKING'])
df['YELLOW_FINGERS']=le.fit_transform(df['YELLOW_FINGERS'])
df['ANXIETY']=le.fit_transform(df['ANXIETY'])
df['PEER_PRESSURE']=le.fit_transform(df['PEER_PRESSURE'])
df['CHRONIC DISEASE']=le.fit_transform(df['CHRONIC DISEASE'])
df['FATIGUE ']=le.fit_transform(df['FATIGUE '])
df['ALLERGY ']=le.fit_transform(df['ALLERGY '])
df['WHEEZING']=le.fit_transform(df['WHEEZING'])
df['ALCOHOL CONSUMING']=le.fit_transform(df['ALCOHOL CONSUMING'])
df['COUGHING']=le.fit_transform(df['COUGHING'])
df['SHORTNESS OF BREATH']=le.fit_transform(df['SHORTNESS OF BREATH'])
df['SWALLOWING DIFFICULTY']=le.fit_transform(df['SWALLOWING DIFFICULTY'])
df['CHEST PAIN']=le.fit_transform(df['CHEST PAIN'])
df['LUNG_CANCER']=le.fit_transform(df['LUNG_CANCER'])
```

```
#Let's check what's happened now
df
```

Figure 4.6 Transform columns value

I explored various machine learning algorithms to predict lung cancer risk based on a dataset obtained from Kaggle, sourced from an online lung cancer prediction system. After preprocessing the data, including converting values to binary and handling missing values, I proceeded with exploratory data analysis (EDA) to understand the dataset's characteristics. One of the key tasks during EDA was identifying the imbalance in the target distribution, which I addressed using oversampling techniques like ADASYN. Additionally, I performed feature engineering to enhance the dataset's predictive power, such as dropping irrelevant features and creating new ones based on existing attributes. For dimensionality reduction, I employed Principal Component Analysis (PCA) to reduce the number of features while retaining important information. Following this, I trained various machine learning models, including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. In this chapter, I focus on the design and implementation of the Voting Classifier ensemble technique. The Voting Classifier combines predictions from multiple base classifiers and predicts the class with the most votes. I utilize this technique to further enhance the predictive performance of my lung cancer prediction system. By aggregating the predictions of multiple models, the Voting Classifier aims to achieve better accuracy and robustness in predicting lung cancer risk. With a thorough understanding of the dataset and the various machine learning algorithms explored, I proceed to implement and evaluate the Voting Classifier as part of my lung cancer prediction system.

Voting Classifier

```python
from sklearn.ensemble import VotingClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.tree import DecisionTreeClassifier

# Define the three lowest-performing models
knn_model = KNeighborsClassifier()
mnb_model = MultinomialNB()
dt_model = DecisionTreeClassifier()

# Create a VotingClassifier ensemble
ensemble_model = VotingClassifier(estimators=[
    ('knn', knn_model),
    ('mnb', mnb_model),
    ('dt', dt_model)
], voting='hard')  # Use 'hard' voting for classification

# Fit the ensemble model on the training data
ensemble_model.fit(X_train, y_train)

# Predict on the testing data
y_pred_ensemble = ensemble_model.predict(X_test)

# Calculate accuracy
accuracy_ensemble = accuracy_score(y_test, y_pred_ensemble)
print("Ensemble Model Accuracy:", accuracy_ensemble)

# Classification report
ensemble_cr = classification_report(y_test, y_pred_ensemble)
print(ensemble_cr)
```

Figure 4.7: Voting Classifier

CHAPTER 5
IMPLEMENTATION AND TESTING

5.1 Applying Grid Search on Logistic Regression

For my lung cancer prediction project, I applied Grid Search on Logistic Regression to optimize hyperparameters and enhance model accuracy. Initially, I imported necessary libraries and prepared the dataset by converting categorical values to binary (0 and 1) and ensuring there were no missing values. The dataset, sourced from Kaggle, consisted of 16 attributes and 284 instances. I handled target distribution imbalance using ADASYN, a synthetic oversampling technique. After splitting the dataset into training and testing sets, I defined a parameter grid for Logistic Regression, including variations in regularization penalties ('l1', 'l2'), inverse regularization strengths (C values), and solvers ('liblinear', 'saga'). Using GridSearchCV, I trained multiple Logistic Regression models with different parameter combinations on the training data and evaluated their performance using cross-validation. This process identified the best hyperparameters, resulting in a refined Logistic Regression model. I assessed the optimized model's performance on the test data, achieving an accuracy of 94.89%, which was an improvement over the initial Logistic Regression model. The classification report provided detailed insights into precision, recall, and F1-score, confirming the model's robustness. This step was crucial in ensuring that the Logistic Regression model was finely tuned for the best predictive performance on lung cancer data.

5.2 Implementation of Database

I employed K-Fold cross-validation as a crucial step in evaluating the performance and robustness of the machine learning models developed for predicting lung cancer risk. K-Fold cross-validation involves partitioning the dataset into k equal-sized folds and then iteratively training the model on k-1 folds while using the remaining fold for validation. This process is repeated k times, with each fold serving as the validation set exactly once. By averaging the evaluation metrics obtained from each iteration, K-Fold cross-validation provides a more reliable estimate of the model's performance compared to a single train-test split. I implemented K-Fold cross-validation with k set to 10, ensuring sufficient data is available for both training and validation in each fold. This technique helped in assessing the generalization capability of the models and detecting potential overfitting or underfitting issues. By analyzing the average accuracy scores across multiple folds, I could make informed decisions regarding model selection and hyperparameter tuning. Additionally, K-Fold cross-validation allowed me to gain insights into the variability of the model's performance, providing a more comprehensive understanding of its effectiveness in real-world scenarios.

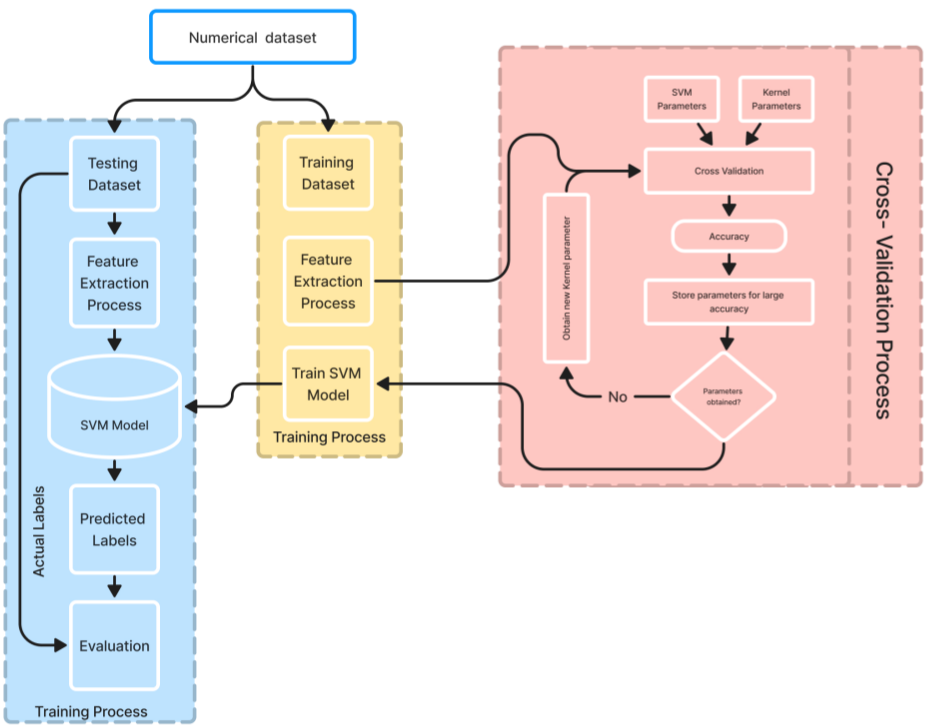


Figure 5.1: K-Fold cross validation

CHAPTER 6

COMPARATIVE STUDY AND ADVANTAGES

6.1 Comparative Study

I have undertaken a comprehensive analysis using various machine learning algorithms to predict lung cancer. For this, I sourced a dataset from Kaggle, containing 16 attributes and 284 instances. I preprocessed the data by converting values to binary (0 and 1) and handling missing values. After exploratory data analysis and feature engineering, I applied several models: Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. The accuracy results are as follows: Logistic Regression achieved 95%, Decision Tree 94.31%, K Nearest Neighbor 91.2%, Multinomial Naive Bayes 89.4%, Support Vector Classifier 93.5%, and Multi-layer Perceptron 94.5%. After applying Grid Search on Logistic Regression, the accuracy improved to 94.89%. I also used K-Fold cross-validation, with Logistic Regression consistently demonstrating the highest average accuracy of 94.4%.

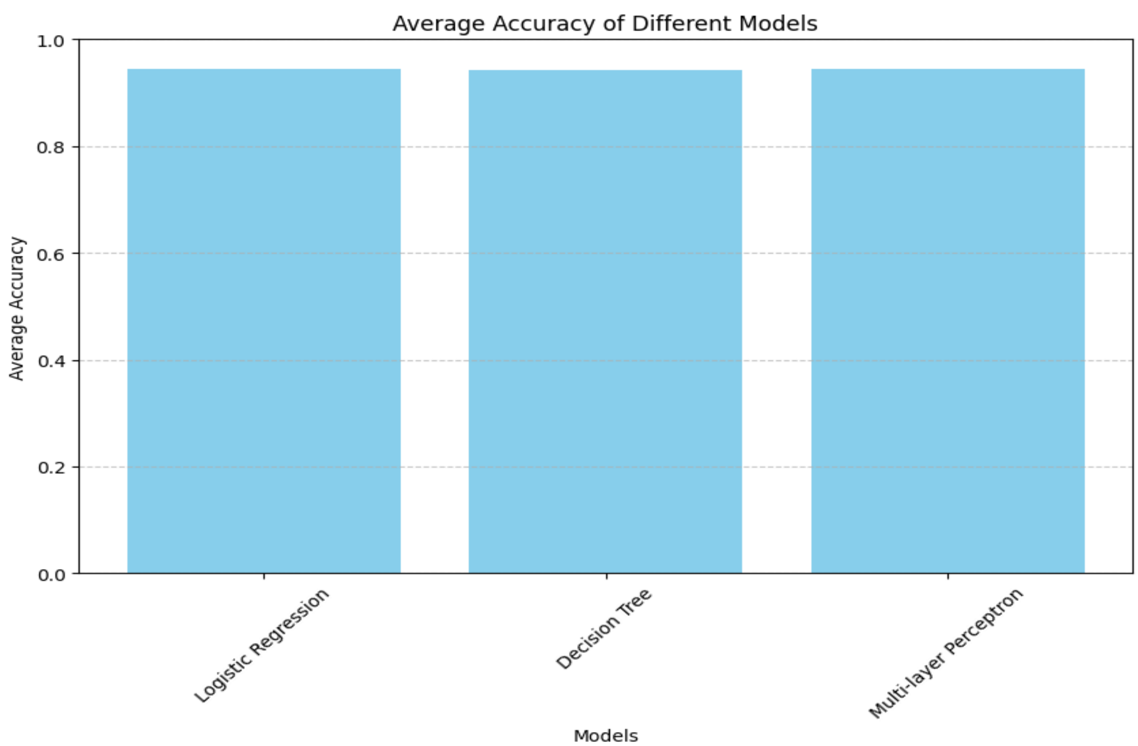


Figure 6.1: Comparative Study

6.2 Advantages of work

I. Early Detection of Lung Cancer:
   - Enables early identification of lung cancer, increasing the chances of successful treatment and survival.

II. Cost-Effective Solution:
   - Provides a low-cost method for individuals to assess their lung cancer risk, making healthcare more accessible.

III. Non-Invasive Approach:
   - Uses readily available data from patient records without requiring invasive procedures.

IV. User-Friendly Interface:
   - Easy-to-use system allows individuals to input their data and receive a quick risk assessment.

V. Customizable and Scalable:
   - Can be tailored to incorporate additional features or adapt to larger datasets, enhancing its applicability.

VI. Improved Decision Making:
   - Assists healthcare professionals and patients in making informed decisions based on the predicted risk.

VII. Comprehensive Data Utilization:
   - Leverages a wide range of features (demographics, lifestyle, and health symptoms) to provide a thorough analysis.

VIII. High Accuracy:
   - Achieves high predictive accuracy through the use of advanced machine learning algorithms, particularly Logistic Regression.

IX. Balanced Target Distribution:
   - Addresses class imbalance effectively, improving the reliability of predictions.

X. Data Visualization:

- Utilizes data visualization techniques to offer clear insights into the relationships between features and lung cancer.

XI. Feature Engineering:

- Enhances the predictive power of the model by creating new, meaningful features.

XII. Dimensionality Reduction:

- Employs PCA to reduce data complexity and improve model performance.

XIII. Hyperparameter Optimization:

- Uses Grid Search to fine-tune model parameters, achieving optimal performance.

XIV. Ensemble Learning:

- Experiments with ensemble techniques like Voting Classifier to explore different model combinations.

XV. Robust Validation:

- Ensures model robustness through K-Fold cross-validation, providing reliable performance metrics.

By implementing these advantages, I have created a robust, reliable, and effective lung cancer prediction system that offers significant benefits for both patients and healthcare providers.

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Discussion and Conclusion

In my work on Lung Cancer Prediction Using Machine Learning Techniques, I utilized a Kaggle dataset with 16 attributes and 284 instances. The dataset was preprocessed by converting values to binary, checking for missing values, and performing feature engineering. Exploratory Data Analysis revealed an imbalance in the target variable, which I addressed using oversampling techniques like ADASYN. I applied Principal Component Analysis (PCA) for dimensionality reduction and trained various machine learning models including Logistic Regression, Decision Tree, K Nearest Neighbor, Multinomial Naive Bayes, Support Vector Classifier, and Multi-layer Perceptron classifier. Logistic Regression achieved the highest accuracy of 95%, which was further optimized using Grid Search, resulting in an accuracy of 94.89%. Although ensemble methods like Voting Classifier were tested, Logistic Regression remained the most effective. K-Fold cross-validation confirmed the robustness of my models, with Logistic Regression consistently outperforming others. This work highlights the importance of careful data preprocessing and model selection in predictive analytics.

7.2 Scope for Further Developments

For the project "Lung Cancer Prediction Using Machine Learning Techniques," there are several potential areas for further development.

I. Incorporating Additional Data Sources: Integrating more diverse and larger datasets could enhance the model's accuracy and generalizability. Data from different geographic locations, ethnic groups, and more detailed medical histories could provide a broader perspective on lung cancer risk factors.

II. Feature Engineering: Advanced feature engineering techniques could be explored to uncover more significant predictors of lung cancer. This could include interaction terms between features, polynomial features, and domain-specific feature transformations.

III. Real-time Prediction System: Developing a real-time prediction system that can be integrated into clinical workflows could make the model more practical for day-to-day use. This would involve creating a user-friendly interface and ensuring the system can handle real-time data input and processing.

IV. Explainability and Interpretability: Enhancing the interpretability of the model predictions is crucial for clinical applications. Techniques such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could be used to provide insights into which features are driving the predictions, thereby making the model's decisions more transparent to medical professionals.

V. Longitudinal Studies: Conducting longitudinal studies to track patients over time could help in understanding the progression of lung cancer and refining the prediction model to account for temporal changes in patient data.

VI. Ethical and Privacy Considerations: Ensuring the ethical use of patient data and maintaining privacy is paramount. Future work could focus on developing frameworks for secure data sharing and compliance with regulations such as GDPR and HIPAA.

VII. Multimodal Data Integration: Combining data from multiple sources, such as genetic information, lifestyle factors, and environmental exposures, could improve the model's predictive power. Techniques for effectively integrating and analyzing multimodal data would be a valuable area of research.

VIII. Validation and Deployment: Extensive validation of the model in diverse clinical settings is necessary before deployment. Collaborating with healthcare institutions for pilot studies and real-world testing would be essential steps towards implementing the model in practice.

By pursuing these avenues, I can further refine and enhance the lung cancer prediction model, ultimately contributing to better healthcare outcomes.

Reference:

[1] S. S. Raoof, M. A. Jabbar, and S. A. Fathima, "Lung cancer prediction using machine learning: A comprehensive approach," *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Mar. 2020. doi:10.1109/icimia48430.2020.9074947

[2] C. Anil Kumar *et al.*, "Lung cancer prediction from text datasets using machine learning," *BioMed Research International*, vol. 2022, pp. 1–10, Jul. 2022. doi:10.1155/2022/6254177

[3] D. Mustafa Abdullah, A. Mohsin Abdulazeez, and A. Bibo Sallow, "Lung cancer prediction and classification based on correlation selection method using machine learning techniques," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 141–149, May 2021. doi:10.48161/qaj.v1n2a58

[4] R. P.R., R. A. S. Nair, and V. G., "A comparative study of lung cancer detection using machine learning algorithms," *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Feb. 2019. doi:10.1109/icecct.2019.8869001

[5] J. Pati, "Gene expression analysis for early lung cancer prediction using Machine Learning Techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232–4238, 2019. doi:10.1109/access.2018.2886604

[6] I. El Naqa, "Machine learning methods for predicting tumor response in lung cancer," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 173–181, Jan. 2012. doi:10.1002/widm.1047

[7] N. Banerjee and S. Das, "Prediction lung cancer– in Machine Learning Perspective," *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, Mar. 2020. doi:10.1109/iccsea49143.2020.9132913

[8] W. Rahane, H. Dalvi, Y. Magar, A. Kalane, and S. Jondhale, "Lung cancer detection using image processing and Machine Learning Healthcare," *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, Mar. 2018. doi:10.1109/icctct.2018.8551008

[9] Y. Xie *et al.*, "Early lung cancer diagnostic biomarker discovery by machine learning methods," *Translational Oncology*, vol. 14, no. 1, p. 100907, Jan. 2021. doi:10.1016/j.tranon.2020.100907

[10] K. Pradhan and P. Chawla, "Medical internet of things using machine learning algorithms for Lung Cancer Detection," *Journal of Management Analytics*, vol. 7, no. 4, pp. 591–623, Aug. 2020. doi:10.1080/23270012.2020.1811789

[11] D. Mustafa Abdullah, A. Mohsin Abdulazeez, and A. Bibo Sallow, "Lung cancer prediction and classification based on correlation selection method using machine learning techniques," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 141–149, May 2021. doi:10.48161/qaj.v1n2a58

Lung Cancer Machine Learning

ORIGINALITY REPORT

7%	6%	3%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	kalaharijournals.com Internet Source	1%
2	www.mdpi.com Internet Source	1%
3	digitaldesign.aalto.fi Internet Source	<1%
4	Submitted to Sheffield Hallam University Student Paper	<1%
5	portfolios.cs.earlham.edu Internet Source	<1%
6	Grisham, Ian L.. "Predicting High-Cap Tech Stock Polarity: A Combined Approach Using Support Vector Machines and Bidirectional Encoders from Transformers", East Tennessee State University, 2023 Publication	<1%
7	huggingface.co Internet Source	<1%
8	Submitted to De Montfort University Student Paper	