Lung Cancer Prediction

GROUP 05

Kavindi Chamathka - s16367

Nethmi Sansala - s16252

Kavindu Weerasekara - s16076

ABSTRACT

Lung cancer is one of the leading causes of cancer-related deaths worldwide, highlighting the urgent need for early detection and effective prediction methods. Machine learning-based techniques are utilized in this report to develop a predictive framework for lung cancer diagnosis. This report also aims to present the findings of the exploratory data analysis and advanced analysis conducted on '**Lung Cancer Prediction Dataset**' obtained from Kaggle. Our investigation focuses on building an efficient prediction model to help medical practitioners make well-informed decisions about diagnosis and treatment through rigorous data preparation, model selection, and performance evaluation.

CONTENTS

Table of Contents

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Lung cancer is one of the most common cancers globally and a leading cause of mortality in the world. In most cases, lung cancer is detected with the least symptoms at its later stage. Hence, diagnosing this at the correct stage for suitable medication is important. Thus, beforehand detection, prediction and diagnosis of lung cancer has become essential as it improves survival rates and treatment effectiveness. In that case, machine learning techniques opens an opportunity for an effortless process. In addressing these challenges, the project seeks to develop a predictive model for risk assessment by identifying the key drivers to be included in this model and identifying the relationships and patterns between them. Also, this will be beneficial for individuals at risk, healthcare professionals and policy makers, allowing them to make important decisions.

"Cancer is a word, not a sentence."— John Diamond

DESCRIPTION OF THE QUESTION

The primary objective of this study is to develop a reliable and interpretable classification model that can predict the likelihood of lung cancer in patients based on various clinical and behavioral factors. By accurately distinguishing between high-risk and low-risk individuals, the model aims to support early diagnosis and enable timely intervention. This study also aims to achieve the following key goals:

- Identify the most influential predictors that contribute to lung cancer risk.
- Evaluate and compare multiple classification models based on their predictive performance.
- Provide a user-friendly and explainable tool to assist healthcare professionals in identifying potential lung cancer cases and making informed clinical decisions.

DESCRIPTION OF THE DATASET

The "Lung Cancer Prediction" dataset is taken from the Kaggle website and contains 5000 records with 18 variables (3 numerical and 15 categorical), out of which the response "PULMONARY_DISEASE" is a categorical variable with two levels.

Variable Name	Description
AGE	Age of the individual (in years).
GENDER	Gender of the individual (male, female).
SMOKING	Whether the individual has smoking habits (Yes/No).
FINGER_DISCOLORATION	Change in the color of fingers (Yes/No).
MENTAL_STRESS	Whether the individual is suffering from mental stress (Yes/No).
EXPOSURE_TO_POLLUTION	Whether the individual contacts with environmental pollutants (Yes/No).
LONG_TERM_ILLNESS	Whether the individual suffers from long term illness (Yes/No).
ENERGY_LEVEL	Overall physical and mental vitality of the individual.
IMMUNE_WEAKNESS	Whether the individual is having a weakness in the immunity (Yes/No).
BREATHING_ISSUE	Whether the individual is suffering from breathing issues (Yes/No).
ALCOHOL_CONSUMPTION	Whether the individual consumes alcohol (Yes/No).
THROAT_DISCOMFORT	Whether the individual suffers from a discomfort in the throat (Yes/No).
OXYGEN_SATURATION	Percentage of oxygen in the blood.
CHEST_TIGHTNESS	Whether the individual feels a tightness in chest area (Yes/No).
FAMILY_HISTORY	Whether the individual has a family history of lung cancer (Yes/No).
SMOKING_FAMILY_HISTORY	Whether a family member has a history of smoking (Yes/No).
STRESS_IMMUNE	Whether the stress impacts on individual's immunity (Yes/No).
PULMONARY_DISEASE	Whether the individual has pulmonary disease (Yes/No).

Table 1: Dataset Description

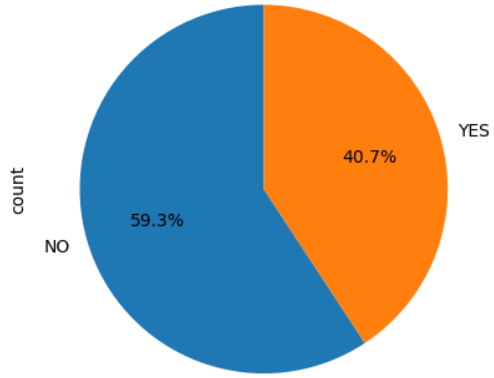
DATA PRE-PROCESSING

- No missing values were found in the dataset.

- No duplicate records were found in the dataset.
- Checked for outliers. A very small number of outliers were identified only in energy level (31) and oxygen saturation (30), indicating that majority of the data points fall within the expected range. Therefore, these outliers will not be removed.
- Then the dataset was splitted into training (80%) and test (20%) sets such that the training set consisted of 4000 records of patients.
- Afterward, the descriptive analysis and advanced analysis were conducted using the training set.

IMPORTANT RESULTS OF THE DESCRIPTIVE ANALYSIS

Response Variable - PULMONARY_DISEASE



The response variable is a binary categorical variable containing two categories, whether the individual has pulmonary disease or not. This suggests that pulmonary disease is relatively common in the sample, affecting about 4 out of every 10 individuals. The dataset is not heavily imbalanced, and it didn't lead to a biased predictive model reducing accuracy. Therefore, SMOTE was not applied.

Figure 1: Distribution of response variable

Distribution of numerical variables

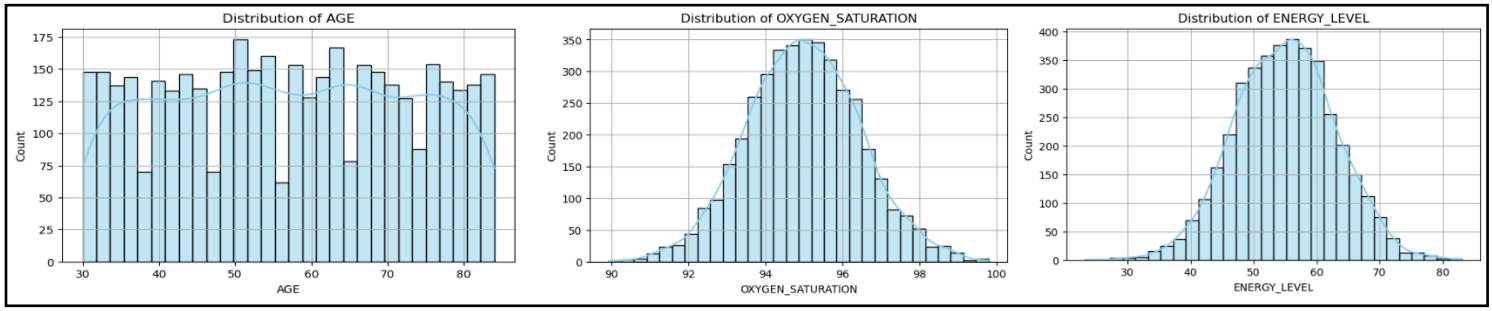


Figure 2: Distribution of numerical variables

The age distribution shows a uniform spread between approximately 30 and 85 years. Similar frequencies in most age intervals indicate a balanced representation across different adult age

groups. Oxygen saturation and energy level follow normal distributions. Distribution of oxygen saturation reflects a healthy population trend, as normal oxygen saturation levels are typically between 92% and 98%. Most individuals have an energy level within the range of 40 to 70, showing that extremely low or high energy levels are rare. The shapes of the features suggest a minimal need for transformation.

Distribution of each numerical variable by response variable - pulmonary disease

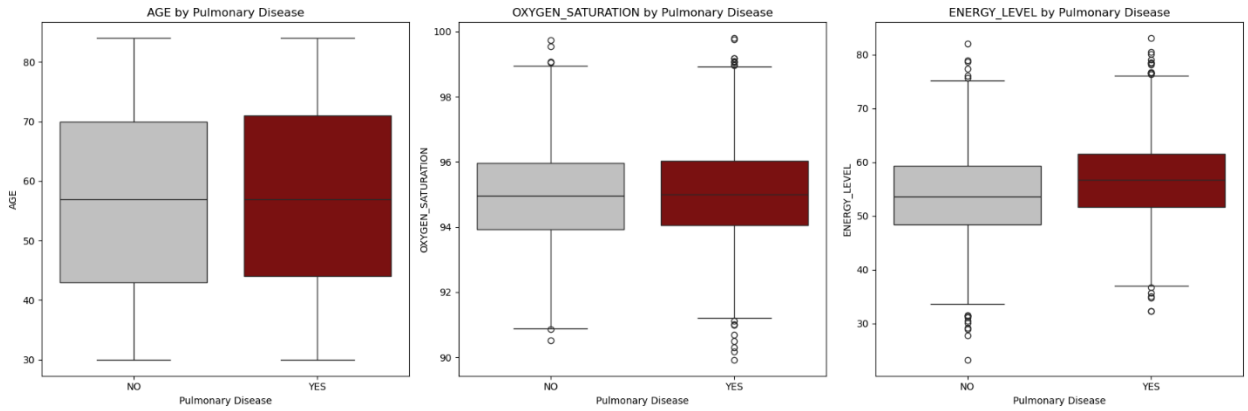


Figure 3: Box plots of numerical variables by Pulmonary Disease

The age distributions are very similar for both groups (Yes and No). Median age is nearly identical, and the interquartile ranges are largely overlapping. Both groups have very similar median oxygen saturation and the IQRs are also overlapping. Slightly more outliers below 92% are seen in the affected group, which is expected clinically. Median energy level is slightly higher in the affected group. Both distributions have a similar range, but the affected group appears to have a higher upper whisker, indicating more people with high energy levels.

Statistical Hypothesis Testing (Mann–Whitney U Test at 5% Significance Level)

To verify whether the above observed differences in distributions between individuals with and without pulmonary disease are statistically significant, Mann–Whitney U tests were conducted for the three numerical variables.

Variable	P Value	Conclusion
AGE	0.4934	Not Significant
OXYGEN_SATURATION	0.0874	Not Significant
ENERGY_LEVEL	0.0000	Significant

Table 2: Hypothesis testing results for numerical variables

The results confirm the visual interpretations obtained through the boxplots. Age and Oxygen Saturation do not show significant group differences, even though minor visual variations were present. This highlights that Energy Level has a significant impact in identifying pulmonary disease in this dataset.

Correlation among numerical variables

The correlation heatmap obtained here highlights that there is no significant linear correlation between any of the pairs of numerical variables in this dataset as all cross-variable correlations are close to zero. Therefore, it indicates that this dataset is lack of multicollinearity.

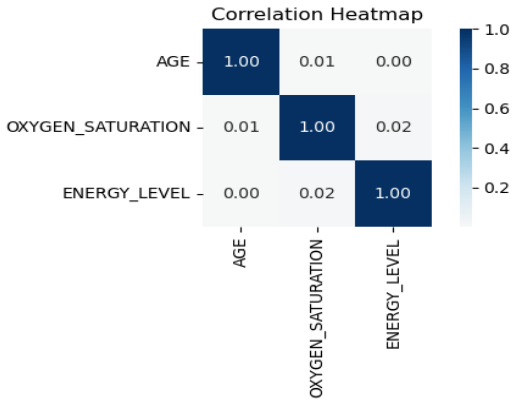
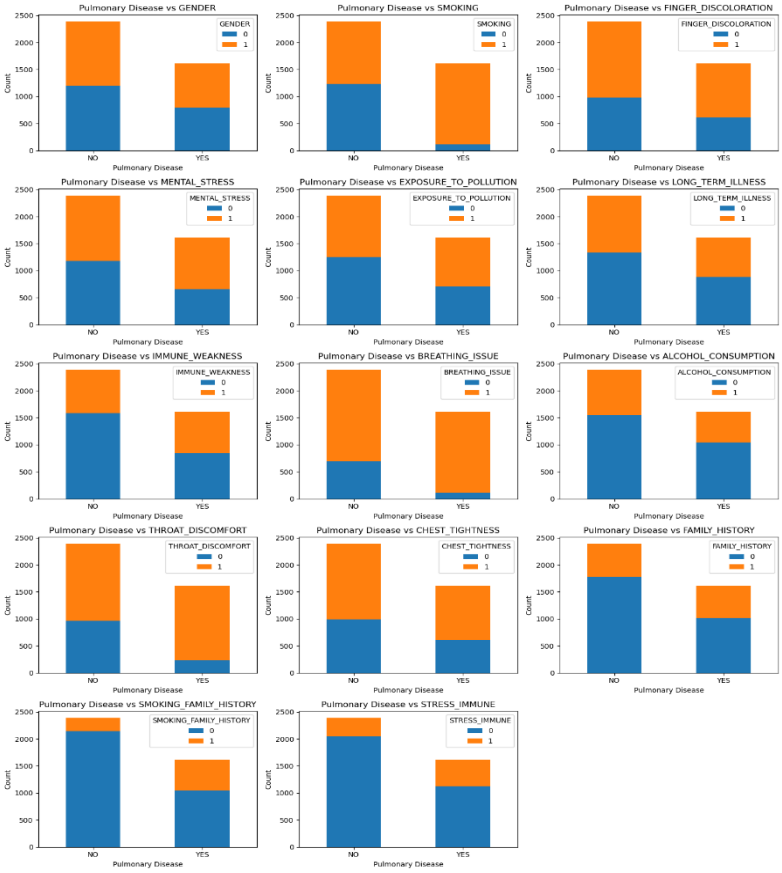


Figure 4: Correlation Heatmap

Distribution of each categorical variable by response variable - pulmonary disease



These stacked bar plots provide initial insights into potential risk factors and associations with pulmonary disease within this dataset. Features like smoking, finger discoloration, exposure to pollution, throat discomfort, immune weakness, breathing issues and smoking family history appear to be associated with a higher likelihood of having pulmonary disease.

Figure 5: Distribution of categorical variables

Chi-Square Test of Independence

To check the association between various features and the presence of pulmonary disease significantly, a series of Chi-Square tests of independence were conducted under 0.05 significance level. The results are summarized below.

Feature	Chi-Square	P-Value	Significant
SMOKING	838.223	0.0000	Yes
MENTAL_STRESS	27.600	0.0000	Yes
EXPOSURE_TO_POLLUTION	23.928	0.0000	Yes
IMMUNE_WEAKNESS	74.306	0.0000	Yes
BREATHING_ISSUE	288.917	0.0000	Yes
THROAT_DISCOMFORT	301.954	0.0000	Yes
FAMILY_HISTORY	55.452	0.0000	Yes
SMOKING_FAMILY_HISTORY	370.627	0.0000	Yes
STRESS_IMMUNE	139.340	0.0000	Yes
CHEST_TIGHTNESS	3.719	0.0538	No
FINGER_DISCOLORATION	3.138	0.0765	No
LONG_TERM_ILLNESS	0.265	0.6064	No
GENDER	0.209	0.6476	No
ALCOHOL_CONSUMPTION	0.002	0.9634	No

Table 3: Hypothesis Testing results for categorical variables

FAMD and K-Means Clustering

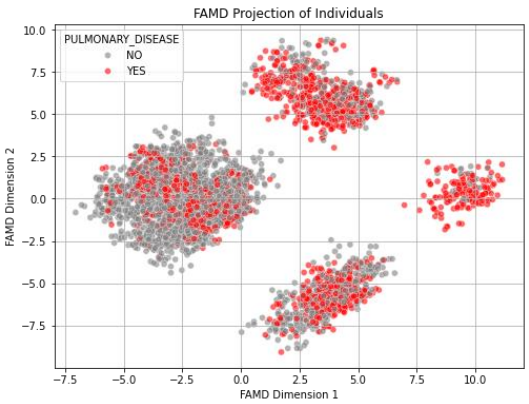


Figure 6: FAMD projection of individuals

To explore whether any clusters in the dataset containing both quantitative and qualitative variables, Factor Analysis of Mixed Data (FAMD) was applied.

The plot here reveals distinct groupings, with some clear separation between individuals with and without pulmonary disease, indicating that the combination of variables used in FAMD is capturing important underlying differences related to the condition.

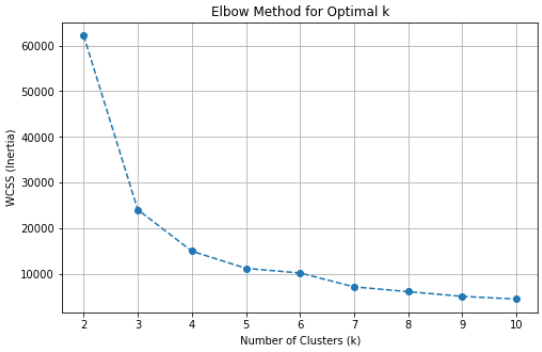


Figure 7: Elbow method for optimal k

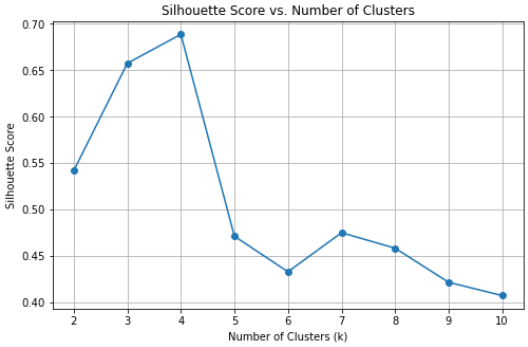
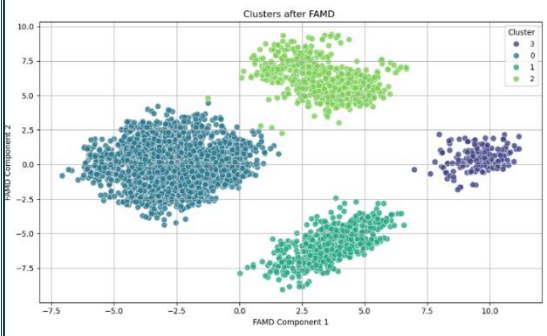


Figure 8: Silhouette Plot

To determine the optimal number of clusters the Elbow method and K-means clustering were used.



Based on the Elbow Method applied to this data, k=4 is the point where the improvement in WCSS starts to level off, suggesting that using more than 4 clusters might lead to overfitting or unnecessary complexity. Also, based on Silhouette Score plot maximum silhouette score is given by k=4 suggesting optimal number of clusters is 4. Figure 9 clearly depicts the four clusters that were identified using the K-Means clustering approach.

Figure 9: Clusters after FAMD

Final Findings of the Descriptive Analysis

- Among the three numerical variables, energy level was only significant.
- Smoking, mental stress, exposure to pollution, immune weakness, breathing issues, throat discomfort, family history, smoking family history and immune stress are significantly associated with pulmonary disease.
- The individuals perform four distinct clusters.

IMPORTANT RESULTS OF THE ADVANCED ANALYSIS

Cluster-wise Model Fitting

Separate models were trained on each of the four clusters to improve prediction accuracy and adapt the model to the unique characteristics of each subgroup. Before applying the models that are considered below, the variables were transformed. To make things easier a pipeline was created for the predictor variables of mixed data types, where the categorical variables were encoded using the one hot encoder and the numerical variables were transformed using the standard scaler. For each cluster, multiple classification models were evaluated, including random forests, logistic regression, support vector machines, decision trees and XGboost. The best-performing model for each cluster was selected based on accuracy on the test set as follows.

Cluster 0					
	Random Forest	Logistic	SVC	Decision Tree	XGboost
Training	1.0000	0.9048	0.9159	1.0000	1.0000
Testing	0.9214	0.9149	0.9198	0.8412	0.9200
Cluster 1					
	Random Forest	**Logistic**	SVC	Decision Tree	XGboost
Training	1.0000	0.8922	0.9192	1.0000	1.0000
Testing	0.8830	0.9006	0.8596	0.7251	0.8300
Cluster 2					
	Random Forest	Logistic	SVC	Decision Tree	XGboost
Training	1.0000	0.9134	0.9289	1.0000	1.0000
Testing	0.7719	0.7310	0.6608	0.7661	0.6500
Cluster 3					

	Random Forest	Logistic	SVC	Decision Tree	XGboost
Training	1.0000	0.8667	0.8606	1.0000	1.0000
Testing	0.9362	0.9149	0.9149	0.7660	0.8900

Table 4: R Squared for the fitted models of the four clusters

To further optimize each cluster-specific model, Recursive Feature Elimination with Cross Validation (RFECV) was applied for feature selection. However, a slight decline in model performance was observed following automatic feature elimination. To address this, the importance score plots were analyzed and based on these insights, key features were selected for each cluster and refitted the models accordingly using a new pipeline. Also, the parameters were fine-tuned using GridSearchCV and given below are the results of the refitted models after feature elimination.

Cluster	Before tunning	After tunning
Cluster 0 – Random Forest		
Training	1.0000	0.9357
Testing	0.9231	0.9247
Cluster 1 - Logistic		
Training	0.8922	0.8922
Testing	0.8947	0.9006
Cluster 2 – Random Forest		
Training	1.0000	0.9243
Testing	0.8012	0.7953
Cluster 3 – Random Forest		
Training	1.0000	1.0000
Testing	0.9362	0.9362

Table 5: Accuracy of models before and after hyperparameter tuning

The obtained importance score plots for each cluster are shown below.

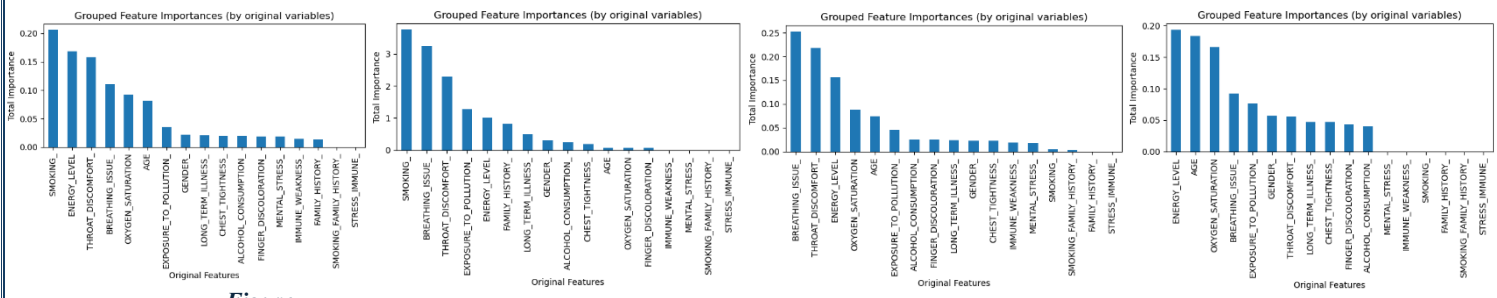


Figure 10:Importance score plots for the four clusters

ISSUES ENCOUNTED AND PROPOSED SOLUTIONS

1. Random Forest and XGBoost models showed perfect accuracy on training data (1.0000), suggesting overfitting.

Solution: Hyperparameters were fine-tuned for the fitted models based on selected features using GridSearchCV.

2. RFECV resulted in a drop in accuracy of the fitted models, maybe due to the removal of weak but jointly informative variables.
   Solution: Manual feature re-selection was done based on importance score plots to gain essential predictors.

DISCUSSION AND CONCLUSION

This study focused on predicting pulmonary disease using clinical and lifestyle data with a range of machine learning methods. Descriptive analysis revealed that energy level and the factors smoking, mental stress, exposure to pollution, immune weakness, breathing issues, throat discomfort, family history, smoking family history and immune stress were significantly linked to disease risk. Clustering helped to identify four subgroups, improving model performance when classifiers were applied separately. Random Forest performed best for three clusters and Logistic Regression performed best for the remaining cluster. The overall framework illustrates a good predictive ability and practical value. Combining cluster-wise modeling with machine learning can effectively predict lung cancer risk. Future work can focus on adding more medical data, improving interpretability, and exploring deeper models for improving accuracy.

REFERENCES

1. https://pythonfordatascienceorg.wordpress.com/chi-square-python/
2. https://www.nature.com/articles/s41598-024-58345-8
3. https://www.datacamp.com/tutorial/random-forests-classifier-python
4. https://maxhalford.github.io/prince/
5. https://www.w3schools.com/python/python_ml_k-means.asp
6. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html
7. https://scikitlearn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html
8. https://machinelearningmastery.com/calculate-feature-importance-with-python/
9. ST3011 Lecture notes
10. https://iarjset.com/wp-content/uploads/2024/06/IARJSET.2024.115112.pdf

APPENDIX

1. Link for the dataset: Lung Cancer Prediction Dataset
2. Click here for Python codes