
Inspiring Excellence

CSE422 - Artificial Intelligence

**Assignment**

# Table Of Contents

# 1. Introduction

This project focuses on developing a machine learning model to predict the likelihood of lung cancer based on various health-related and lifestyle features. The dataset used includes detailed information about 15 features: Gender, Age, Smoking, Yellow Fingers, Anxiety, Peer Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness of Breath, Swallowing Difficulty, Chest Pain, and Lung Cancer (target variable). The primary goal

of this project is to provide an accurate and efficient predictive tool to assist in the early detection of lung cancer.

The motivation behind this project stems from the growing global health concern regarding lung cancer, which remains one of the leading causes of cancer-related deaths. Early detection is crucial for improving survival rates and optimizing treatment plans. By leveraging machine learning techniques, we aim to create a scalable and reliable model that can analyze complex patterns in health data, thereby facilitating early diagnosis and contributing to better health outcomes.

---

## 2. Dataset Description

**Source**

- **Link**: https://www.kaggle.com/datasets/nancyalaswad90/lung-cancer/data
    - **Reference**: Kaggle

# Dataset Overview

- **Features:** The dataset contains 2 categorical features such as 'GENDER', 'LUNG_CANCER' and 14 Numerical Variables such as 'AGE', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY', 'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING', 'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH', 'SWALLOWING DIFFICULTY', 'CHEST PAIN'
    - **Type of Problem:** This is a classification problem where the target is to predict Lung cancer whether a is **YES** or **NO**.

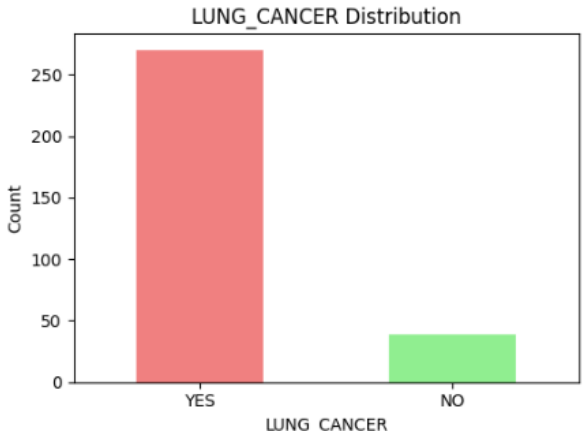
- **Rows and columns:** There are 309 rows and instances and 16 columns in the dataset.
- **Feature Types**: 2 features are categorical & 14 features are Numerical . There are no quantitative features.
- **Datapoint:** There are 4944 data points in this dataset

## Balanced/Imbalanced Dataset Analysis:

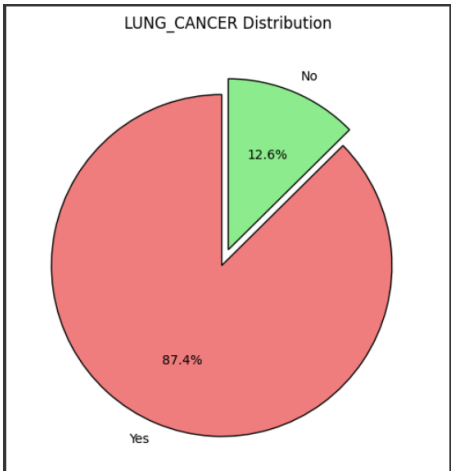
The target feature (Lung_Cancer) has two types of instances:

- **Yes**: 270 instances
- **No**: 39 instances

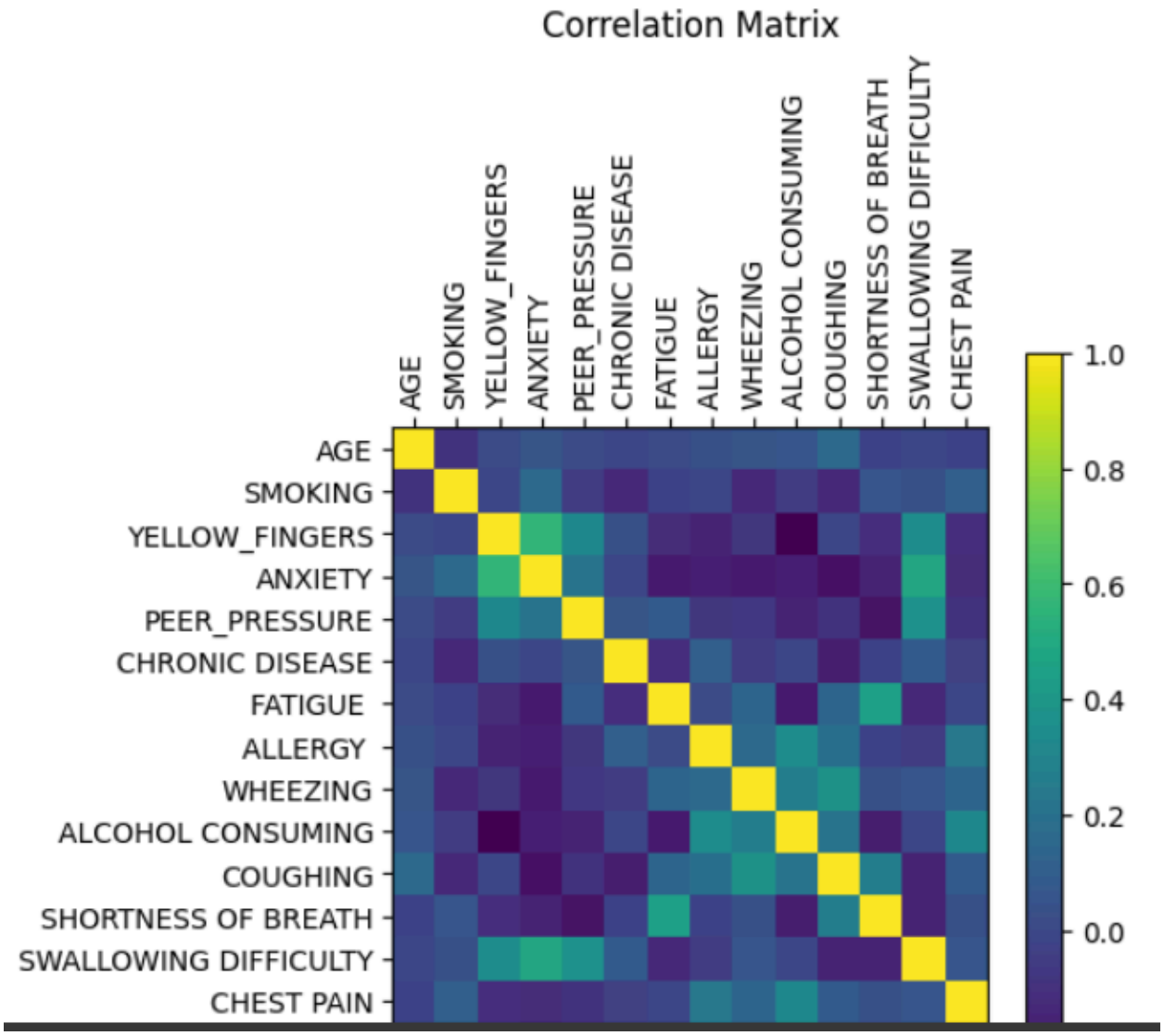
The dataset is imbalanced, as confirmed by a bar chart & pie chart representing the distribution of classes. The target feature, Lung_Cancer, has a significant disparity in instances: **Yes** (270 instances, ~87.4%) and **No** (39 instances, ~12.6%).




**tures :**




To understand the relationships among features and their relevance to the target variable (LUNG_CANCER), a correlation heatmap was generated using the Matplotlib library. This visualization highlighted key features, such as YELLOW_FINGERS, which strongly correlated with ANXIETY (0.5658), and FATIGUE, which showed a strong correlation with SHORTNESS OF BREATH (0.4417). Conversely, CHRONIC DISEASE displayed minimal correlation with most features and was considered for removal to optimize model performance.




## 3. Dataset Pre-processing

**Adding Null Values**: Null values were introduced in the "AGE" column by setting every third value to NaN. This was done to simulate real-world missing data scenarios and test the model's ability to handle them.

**Removing Duplicates**: Duplicate rows were identified and removed from the dataset to prevent any bias during model training and ensure that each record is unique.

**Handling Missing Values**: The missing values in the "AGE" column were filled with the mean of the available values. This method is commonly used to impute missing data in numerical columns and preserve the overall distribution of the data.

**Encoding Categorical Variables**: The "GENDER" column, which was categorical, was encoded using one-hot encoding to convert the non-numeric values into a suitable format for machine learning models. This approach creates a binary column for each category, helping the model interpret the data effectively.

**Label Encoding for Target Variable**: The "LUNG_CANCER" column, which contains binary categorical values ('YES' and 'NO'), was converted into numerical values (1 and 0) for easier processing by machine learning algorithms.

After these steps, the dataset was ready for further analysis and model training. The shape of the dataset is now:

# 4. Dataset Splitting

The dataset was divided into training and testing subsets:

- **Training Set**: 70% of the data (196 instances) was used for model training.
- **Test Set**: 30% of the data (85 instances) was reserved for model evaluation.

A **stratified** split was implemented to maintain the class distribution across both subsets, ensuring that the proportions of edible and poisonous mushrooms remained consistent. It removes any kind of bias that may influence the model.

---

# 5. Model Training & Testing

The following models were implemented to predict lung cancer:

1. **Decision Tree**:
    - This model captures complex non-linear relationships in the data and is well-suited for categorical features.
    - Hyperparameters such as max depth and criterion were optimized to improve accuracy and reduce overfitting.
2. **Random Forest**:
    - An ensemble method that builds multiple decision trees and combines their outputs for robust predictions.
    - Hyperparameters like the number of estimators were tuned for optimal performance.
3. **K-Nearest Neighbors (KNN)**:
    - A distance-based algorithm that classifies a data point based on the majority class among its nearest neighbors.
    - The value of (number of neighbors) affects accuracy.

4. **Support Vector Machine (SVM)**:
    - A powerful classifier that finds the optimal hyperplane separating classes in high-dimensional space..

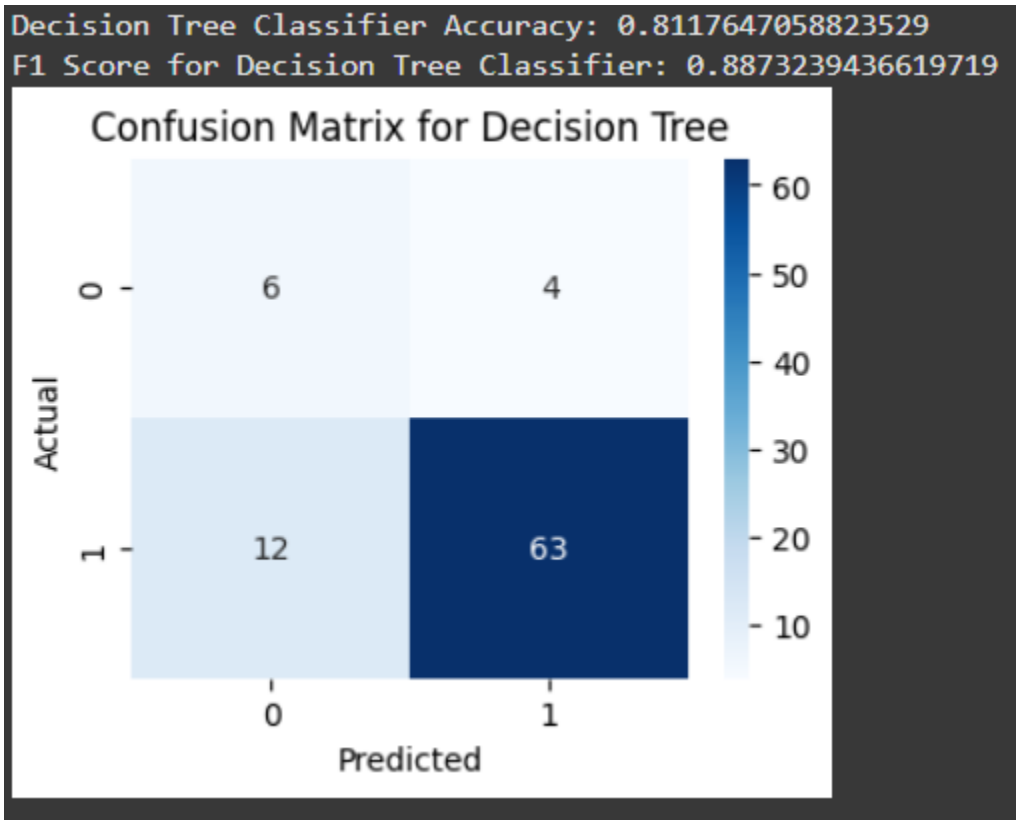
Each model was trained on the training set (70%) and evaluated on the test set(30%) using various performance metrics such as accuracy, precision, recall, f1 score. Scaling worked significantly for the SVM model.

---

# 6. Model Selection/Comparison Analysis

**Decision Tree Classifier**

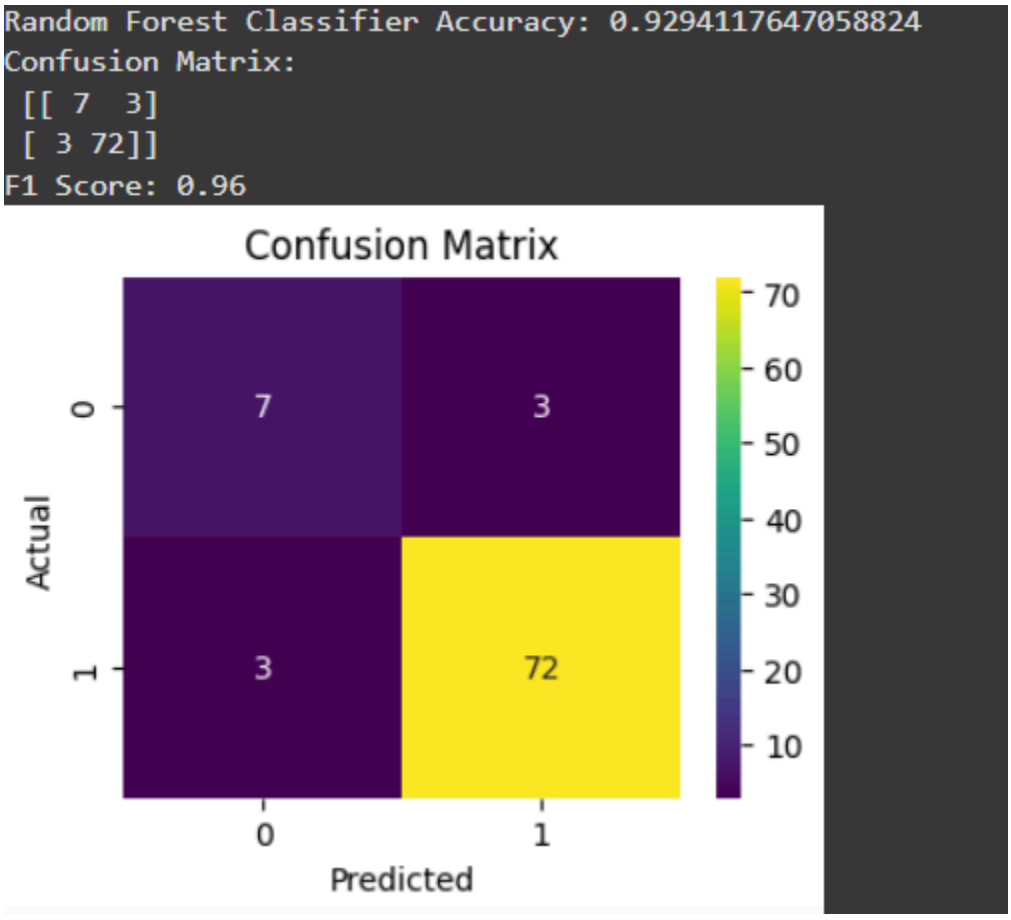
The Decision Tree model demonstrated great performance in lung cancer prediction classification, an accuracy of 81.18%. The F1-scores of 99.73% for both classes indicate a good model performance. The confusion matrix showing a few misclassification. While the model shows good predictive capability, its performance suggests some limitations in capturing all the complex patterns in the data. The Decision Tree's performance, while impressive, falls slightly short of the more sophisticated ensemble methods, possibly due to its tendency to overfit on specific patterns in the training data.




**Random Forest Classifier**

The Random Forest model exhibited exceptional performance, achieving an accuracy of 92.94%., the model achieved very high precision and recall scores of 1. The F1-scores were equally impressive at 0.96, demonstrating the model's ability to maintain balance between

precision and recall. The confusion matrix showed a few misclassifications. This superior performance can be attributed to the ensemble nature of Random Forest, which combines multiple decision trees to reduce overfitting and capture complex patterns in the data. The model's ability to handle the categorical features and maintain high accuracy suggests it effectively learned the underlying patterns while avoiding overfitting, making it one of the top performers among all tested models.



**K-Nearest Neighbors Model**

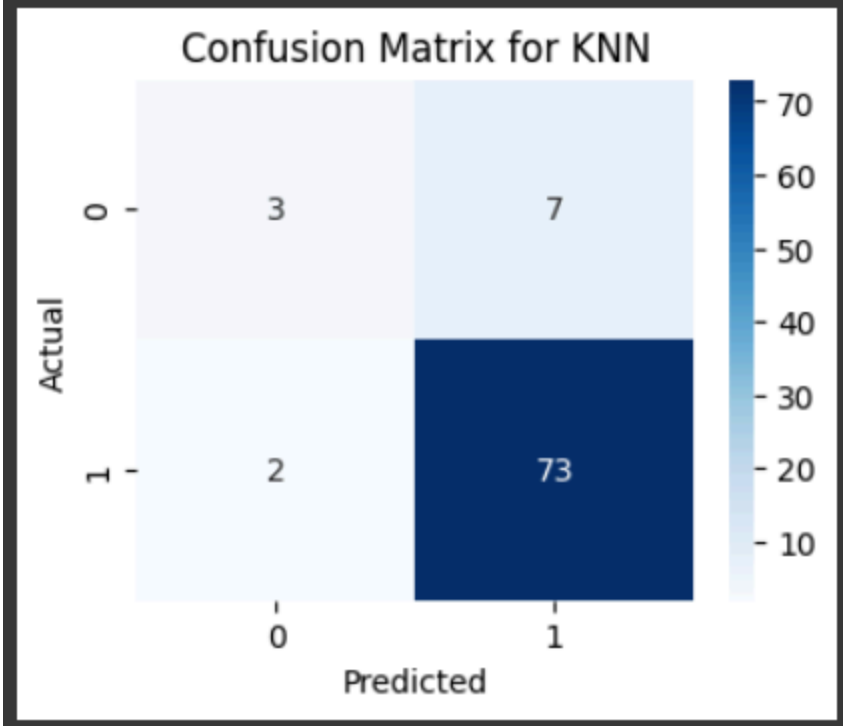
The K-Nearest Neighbors (KNN) model performed very well for the most cases with 91.23% correct prediction. Furthermore, the model's outstanding precision-recall balance is highlighted by the F1-score of 0.942. The confusion matrix provides additional evidence thatmost cases were correctly categorized. This great performance implies that the dataset is quite separable, enabling KNN to detect patterns even in the absence of feature scaling. But as the model might not translate well to new data, these findings also raise questions regarding possible overfitting. Although KNN is straightforward, easy to understand, and very successful in this situation, its dependence on distance computation restricts scalability for larger datasets, and scaling may be necessary for real-world applications to achieve the best results.

```
KNN Classifier Accuracy: 0.8941176470588236
Classification Report for KNN:
              precision    recall  f1-score   support

           0       0.60      0.30      0.40        10
           1       0.91      0.97      0.94        75

    accuracy                           0.89        85
   macro avg       0.76      0.64      0.67        85
weighted avg       0.88      0.89      0.88        85

F1 Score: 0.9419354838709677
```




**Support Vector Machine Model**

An outstanding 96.49% accuracy by the Support Vector Machine (SVM) model. The' F1-scores of 0.98 demonstrated a good balance between recall and precision. On the other hand, 3

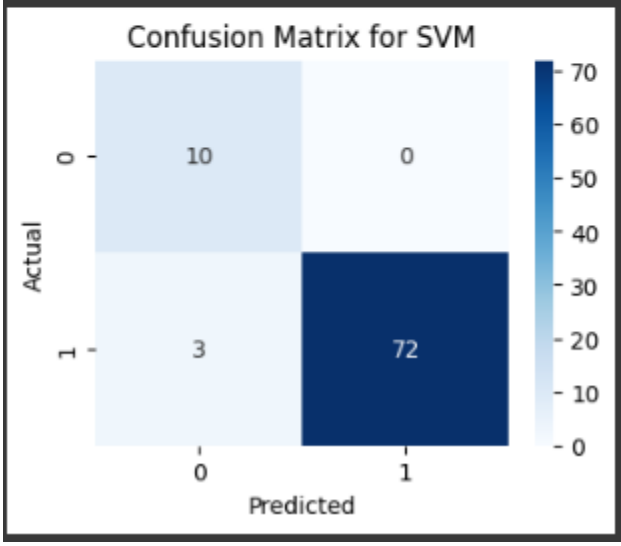
misclassifications were found in the confusion matrix (3 false negatives for and 0 false positive ). Precision, recall, and F1-scores for both classes approached near-perfect values.

```
SVM Accuracy: 0.9647058823529412
Confusion Matrix for SVM:
 [[10  0]
 [ 3 72]]
Classification Report for SVM:
              precision    recall  f1-score   support

           0       0.77      1.00      0.87        10
           1       1.00      0.96      0.98        75

    accuracy                           0.96        85
   macro avg       0.88      0.98      0.92        85
weighted avg       0.97      0.96      0.97        85

F1 Score: 0.9795918367346939
```




**Comparison with other models (Decision Tree)**

Decision trees provide an obvious solution to classification difficulties since they are very interpretable and simple to use. They tend to overfit, especially when dealing with complicated or noisy datasets, but they work well on smaller datasets with distinct patterns. In contrast to models like SVM or Random Forest, this may result in a reduction in the generalizability of the models, rendering them less resilient. Techniques like pruning and ensemble approaches can significantly increase the performance of the decision tree algorithm. In certain use scenarios, especially when model transparency is essential, Decision Trees offer a decent balance between accuracy and interpretability despite their propensity to overfit.

**Comparison with other models (Random Forest)**

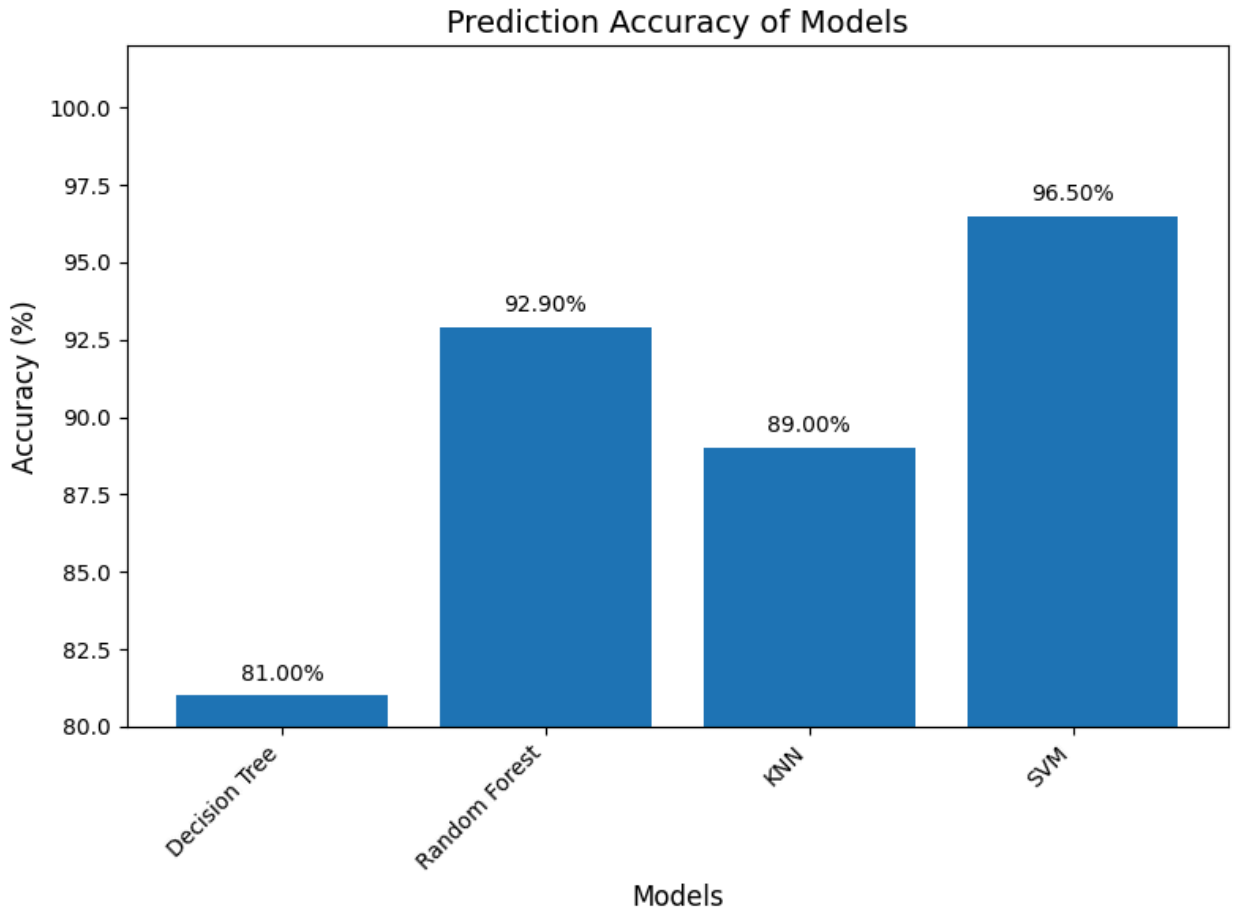
Multiple decision trees are combined in the Random Forest ensemble approach to increase resilience against noise, decrease overfitting, and improve accuracy. Large, high-dimensional datasets and challenging classification issues are two areas in which it excels. On well-structured datasets, Random Forest may still fall short of models like SVM in terms of precision and recall, even if it typically performs better than individual decision trees and Naive Bayes. The strengths of Random Forest are its capacity to manage feature interactions, prevent overfitting, and generalize effectively. Nevertheless, it requires more computing power than more straightforward models like Decision Trees or Naive Bayes. Compared to models like SVM, Random Forest may need more resources and training times, but it can attain outstanding accuracy in large-scale datasets.

**Comparison with other models (SVM)**

Because Support Vector Machines (SVM) can handle high-dimensional data and efficiently establish decision boundaries, they offer a reliable answer to classification challenges. Compared to more straightforward models like Decision Trees, they may have trouble with interpretability, although they typically perform well on complicated datasets with obvious separability. SVMs use kernel functions to capture complicated patterns, making them very useful for large and complex datasets. Less transparency and more processing demands, however, may result from this. SVMs frequently perform better in terms of accuracy and generalizability than more straightforward models like Decision Trees, despite these trade-offs. They are a popular option in situations where accuracy and durability are crucial due to their capacity to identify the ideal hyperplane.

**Comparison with other models (KNN)**

In this instance, the KNN model outperforms models like Decision tree in terms of outcomes because of the dataset's structure. However, some models may be more robust to noisy or real-world data and less sensitive to hyperparameters like the choice of k. Because of its remarkable performance in this scenario, KNN is a desirable choice for classification problems of a similar kind. Larger datasets and more cross-validation testing would be prudent to ensure dependability for more popular applications.


Prediction Accuracy of Models
Accuracy (%)
100.0
97.5
95.0
92.5
90.0
87.5
85.0
82.5
80.0
81.00%
92.90%
89.00%
96.50%
Decision Tree
Random Forest
KNN
SVM
Models

## 8. Conclusion

In this project, we successfully developed and evaluated machine learning models to predict lung cancer based on given features features. Leveraging a comprehensive dataset, we systematically preprocessed the data, including feature selection, encoding, to optimize model performance. 4 models—Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were implemented and compared using various metrics.

The Random Forest and SVM models emerged as top performers, achieving almost flawless accuracy due to their ability to handle complex patterns. KNN also demonstrated great accuracy. Decision Tree, though accurate, was prone to overfitting and was surpassed by its ensemble counterpart, Random Forest.

The results highlight the potential of machine learning in predicting lung cancer. Each model brought unique strengths, underscoring the importance of aligning model selection with dataset characteristics and application needs. Future improvements could involve expanding the dataset to include more entries and exploring advanced ensemble techniques for enhanced generalizability.

This project exemplifies the transformative power of AI in solving real-world problems, paving the way for safer foraging and medical sector. By automating the prediction process, we have taken a step toward mitigating health risks associated with preventing this deadly disease.