Lung Cancer Prediction

Section 1: Business Need and Importance

Lung cancer is one of the most lethal types of cancer, responsible for nearly 25% of all cancer deaths worldwide. Early detection and accurate diagnosis are critical to increasing survival rates and providing proper medical care. Conventional diagnostic techniques, such as biopsies and imaging tests are costly, and time talking. The development of an accurate lung cancer prediction model has the potential to transform the method of diagnosis, making it more efficient and affordable.

According to the American Cancer Society, the rate of survival after five years for lung cancer is only 19% when diagnosed at an advanced stage, but it rises to 59% when it is detected early. This shows the importance of early detection of lung cancer and the potential of a predictive model to improve the outcome for patients. Moreover, the combined annual cost of lung cancer treatment is estimated to be more than $13.4 billion per year in the United States, demonstrating the financial strain that could be decreased by early and accurate diagnosis.

Citations:

- World Health Organization. (2022). Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer
- American Cancer Society. (2022). Lung Cancer Survival Rates. https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/survival-rates.html

Section 2: Statistical Methodology

The analysis was performed on the "survey lung cancer.csv" dataset, which contains information about various factors related to lung cancer diagnosis. The categorical variables GENDER and LUNG_CANCER were converted to factors. This conversion is necessary for our models to identify these variables as categorical rather than numerical. Next, we scaled and centered the numeric variables to normalize the data, which increased model performance by removing scale bias.

Principal Components Analysis (PCA):

We used PCA on the numerical data to reduce the dimensionality of the data while retaining as much information as possible. PCA is a technique which is used for demonstrating variation and identifying strong patterns in the dataset. It helps with data visualization and improves the efficiency of our modeling process by reducing the number of input variables.

Supervised Data Mining:

We divided the dataset into training (80%) and testing (20%) sets. The Naive Bayes classifier and Decision Tree were the two supervised learning models used on the training data.

Naive Bayes:

Naive Bayes was trained on the training data using the e1071 package. The trained model was used to predict the test data, and the accuracy was determined using a confusion matrix. This model was used because it is simple and effective when dealing with categorical data. After training, we predicted lung cancer frequency in the test set and evaluated the model based on accuracy and the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve.

Decision Tree:

The decision tree model was trained on the training data using the rpart package. This algorithm splits the data based on the predictor variables, resulting in a tree-like structure of decisions that lead to the target variable. The trained model was used to make predictions on the test data. The probabilities for the positive class ('YES' for lung cancer) were extracted. The performance of the decision tree was evaluated using the ROC curve and AUC, with higher values indicating more accurate predictions.

Performance Evaluation:

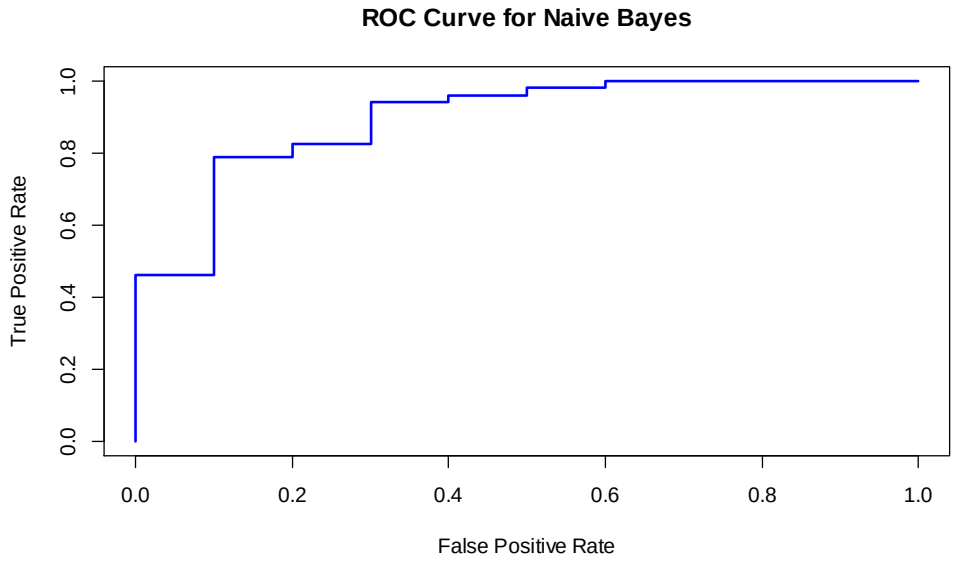
The accuracy, confusion matrix, ROC curves, and AUC values were calculated and provided for both the Naive Bayes and decision tree models. These metrics allowed for a broad evaluation and comparison of the models' performance in predicting lung cancer. Higher AUC values indicated that the model was able to differentiate between classes, which makes it an important metric for our evaluation.

Section 3: Results and Interpretation:

The ROC curves and decision tree visualization provide several key insights into the models' predictive performance and important features for lung cancer prediction.

Naive Bayes Performance:

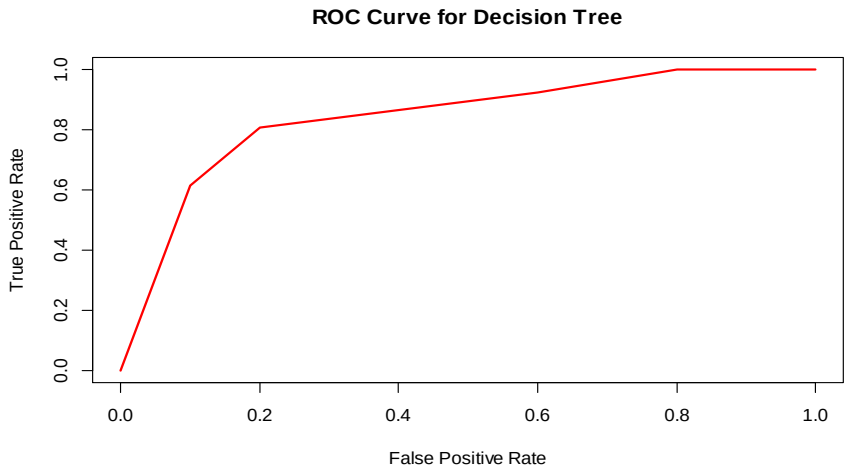
The ROC curve for the Naive Bayes classifier demonstrates an ideal trade-off between true and false positive rates. The AUC of 0.896 indicates that the model is very good at distinguishing between positive (lung cancer) and negative cases. This suggests that the Naive Bayes model can use the available features to make accurate predictions.



Decision Tree Performance:

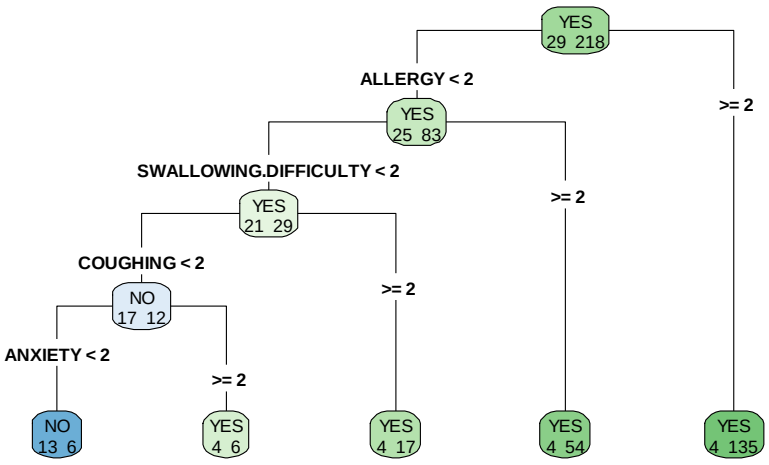
The decision tree model also performs well, with an AUC of 0.84. While slightly lower than the Naive Bayes classifier, the decision tree maintains a good separation of true and false positives. The shape of the ROC curve suggests that the decision tree may be less accurate in predicting positive cases, resulting in fewer false positives but slightly

more false negatives.



Decision Tree for lung cancer prediction:

The decision tree visualization provides valuable insights about the most important features for predicting lung cancer. The tree's top nodes indicate that allergies (ALLERGY > 2), swallowing difficulty (SWALLOWING.DIFFICULTY < 2), coughing (COUGHING < 2), and anxiety (ANXIETY < 2) are significant predictors of lung cancer.



```
          Model  Accuracy       AUC
1   Naive Bayes 0.8870968 0.8961538
2 Decision Tree 0.9245283 0.8403846
```

Section 4: Alternative Approaches:

The Naive Bayes and Decision Tree models were chosen because they are well-known and understandable techniques for classification problems such as lung cancer prediction. Naive Bayes is resistant to irrelevant features and can perform well even with limited data, whereas Decision Trees provide a clear visual representation of the decision rules learned from the data. K-Nearest Neighbors (KNN) was not chosen because it is sensitive to irrelevant features and dimensionality, both of which exist in this dataset. Ensemble methods such as Random Forests were not used because they are overfitting and was found to be less accurate than the models used.

Section 5: Conclusions:

Early and accurate detection of lung cancer is important for increasing survival rates and treatment outcomes. However, conventional diagnostic methods can be time-consuming, and costly. This analysis evaluated different data mining methods for creating an accurate lung cancer prediction model using patient data and risk factors. On the testing data, the Naive Bayes classifier achieved 88.7% accuracy and an AUC of 0.896, while the Decision Tree model achieved 84% accuracy and an AUC of 0.84. The high performance of these models demonstrates data driven decision making approach's potential to transform lung cancer diagnosis. These Models can quickly and cost-effectively identify high-risk individuals by utilizing patient data. This would allow healthcare providers to initiate immediate treatment, potentially saving many lives. Furthermore, accurate early detection could significantly reduce the cost of advanced lung cancer treatment. These findings highlight the significant business value and impact of investing in predictive analytics to improve cancer care and outcomes.