Lung Cancer Prediction

Anirudh Tiku (20224698) - 19at77@queensu.ca

Anneth Sivakumar (20320973) - 21as221@queensu.ca

Avi Mohan (20175700) - 19am3@queensu.ca

Braedon Van Wiechen - 21bjvw@queensu.ca

Zain Zaidi (20257425) - zain.zaidi@queensu.ca

April 2024

1.0 INTRODUCTION

Lung cancer remains a significant global health concern, and its prevalence and consequences continue to pose considerable concerns. Lung cancer is the most diagnosed cancer in Canada, accounting for most cancer-related deaths among both men and women [1]. According to the International Agency for Research on Cancer (IARC), lung cancer is the leading cause of cancer-related deaths globally, resulting in over 1.8 million fatalities (18%) in 2020 [2]. The disease is frequently detected at an advanced stage, resulting in restricted treatment options and an undesirable prognosis [2].

Given the severity of lung cancer and its impact on individuals and the healthcare system, there is an urgent need to better understand its risk factors and early indicators. This motivates our study, which aims to analyze the risk factors associated with lung cancer using a dataset containing information on patients diagnosed with the disease. Understanding the risk factors for lung cancer is important for a variety of reasons. For starters, it can assist in identifying high-risk individuals who could benefit from targeted screening programs, resulting in earlier detection and better outcomes. Second, it can help to shape public health policies targeted at reducing the incidence of lung cancer by targeting modifiable risk factors such as smoking and air pollution exposure. Finally, it can contribute to the development of individualized treatments for people who are at risk of developing the disease.

In this report, we chose a dataset called Lung Cancer Prediction from Kaggle [3]. While previous studies have focused on computer-assisted systems, including advancements in deep learning techniques for interpreting computed tomography (CT) imaging, additional research is required to explore numerous factors and their interactions that may contribute to the development of lung cancer [4]. Thus, by leveraging statistical tools, our approach aims to give a thorough assessment of many factors, including lifestyle, environmental, and genetic predisposition, shedding insight into putative mechanisms underlying lung cancer development.

2.0 DATASET DESCRIPTION AND VISUALIZATION

While treating lung cancer presents substantial challenges, there are various strategies that can help mitigate its impact on patients. Adopting a healthy lifestyle, limiting carcinogen exposure, and scheduling regular medical screenings, can play a crucial role. Early diagnosis is pivotal in facilitating lifestyle changes and enabling more effective treatments.

We describe the dataset in ***Table 1*** to reveal general information about each group that generates 24 variables with a 1000-sample size. The dataset initially contained 1000 records with 26

attributes. However, on further inspection, we realized that the first and second attributes i.e. 'Index' and 'Patient.Id' were redundant for our analysis, so we dropped them from our dataset.

Table 1. Dataset Summary

Variable Name	Mean	Maximum	Minimum	Variable Type
Age	37.170	73	14	numerical
Gender	1.402	2	1	categorical
Air Pollution	3.840	8	1	categorical
Alcohol Use	4.563	8	1	categorical
Dust Allergy	5.165	8	1	categorical
Occupational Hazards	4.840	8	1	categorical
Genetic Risk	4.580	7	1	categorical
Chronic Lung Disease	4.380	7	1	categorical
Balanced Diet	4.491	7	1	categorical
Obesity	4.465	7	1	categorical
Smoking	3.948	8	1	categorical
Passive Smoker	4.195	8	1	categorical
Chest Pain	4.438	9	1	categorical
Coughing of Blood	4.859	9	1	categorical
Fatigue	3.856	9	1	categorical
Weight Loss	3.855	8	1	categorical
Shortness of Breath	4.240	9	1	categorical
Wheezing	3.777	8	1	categorical
Swallowing Difficulty	3.746	8	1	categorical
Clubbing of fingernails	3.923	9	1	categorical
Frequent Cold	3.536	7	1	categorical
Dry Cough	3.853	7	1	categorical
Snoring	2.926	7	1	categorical
Level	"Medium"	"High"	"Low"	categorical

2.1 Demographic Distribution of Lung Cancer

In ***Figure 1*** and ***Figure 2*** reveals the distribution of Lung Cancer levels by gender and age. Specifically, in ***Figure 1***, it can be observed that the levels of Lung Cancer across males and females (1 and 2 respectively) is relatively the same. However, there is a slightly higher chance of males developing higher-stage cancer whereas females have a slightly higher chance of developing lower-stage cancer.

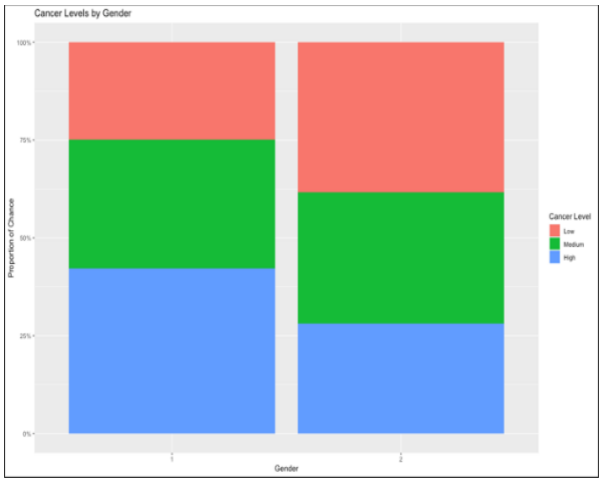


Figure 1. Stacked Bar Chart of Lung Cancer Levels by Gender

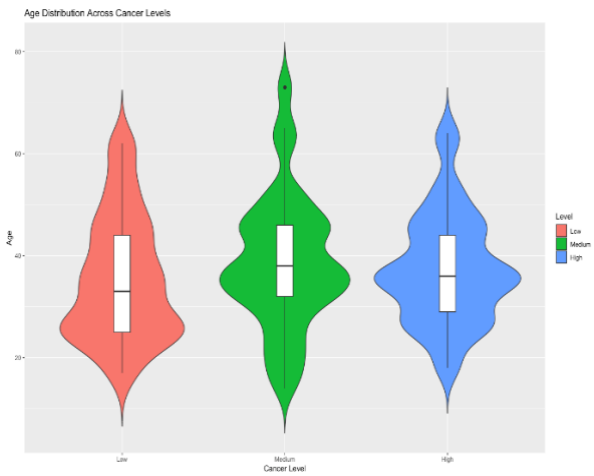
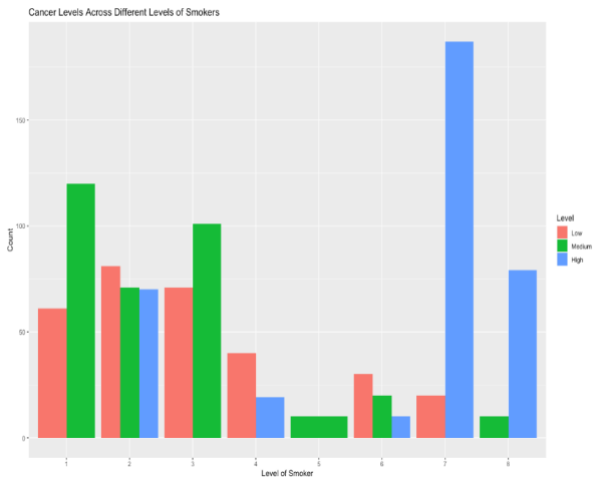


Figure 2. Violin Plot of Age Distribution Across Lung Cancer Levels

2.2 Smoking and Lung Cancer Statistics

Figure 3 and ***Figure 4*** display the impact of smoking on Lung Cancer. It is very well known that smoking is carcinogenic, and we can easily infer that from the histogram in ***Figure 3***. The count of lower-stage lung cancer among smokers is high, and for heavy smokers, there is an obviously increased count of end-stage lung cancer. This is also apparent with patients who have exposure to passive smoking, as can be observed in ***Figure 4***.



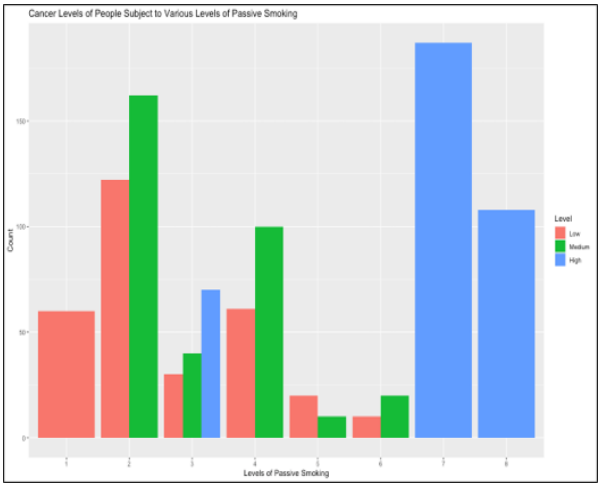


Figure 3. Histogram of Lung Cancer Levels Across Different Levels of Smokers

Figure 4. Histogram of Lung Cancer Levels Based on Exposure to Different Levels of Passive Smoking

2.3 Impact of Eating Habits on Lung Cancer

Obesity and lack of a healthy and balanced diet have also been correlated with the severity of lung cancer, as seen in **Figure 5** and **Figure 6**. Counts of patients with higher-stage lung cancer are higher with worse levels of obesity. This suggests that obesity is associated with the severity of lung cancer – either as a precursor or a side effect and that having good health and a more balanced diet could reduce the chances of getting lung cancer.

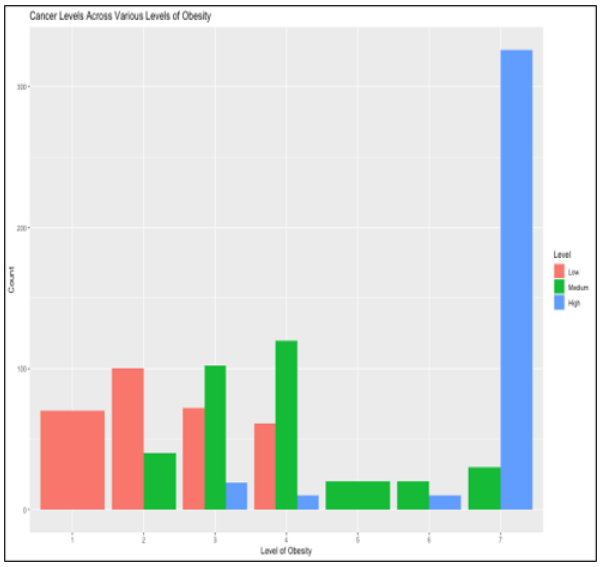


Figure 5. Histogram of Lung Cancer Levels Based on Different Levels of Obesity

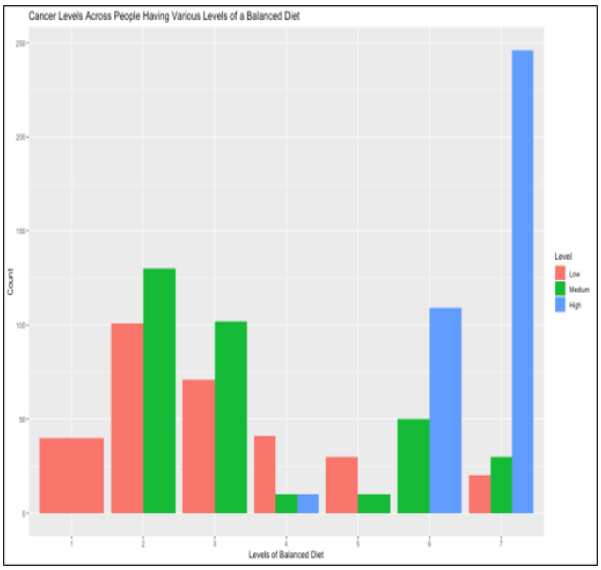


Figure 6. Histogram of Lung Cancer Levels Based on Levels of Balanced Diets

2.4 Respiratory Responses to Lung Cancer

In this dataset, we explored the association of common respiratory illnesses and their association with lung cancer. First, we explored the association with lung cancer and chronic lung disease. In **Figure 7**, we found that the severity of lung cancer is worse as the level of pre-existing chronic illness is worse. On further investigation, this trend makes sense as chronic lung illness weakens the area near the respiratory tract and increases the inflammatory response of the body. This makes conditions more favorable for the tumor to expand and grow, thus increasing the severity of lung cancer. This trend is also seen in the other respiratory system in **Figure 8** and **Figure 9**,

with blood coughs and wheezing. Because the area near the respiratory tract is weakened, the severity of blood coughs and wheezing is greater with the bigger tumor in higher stage lung cancers.

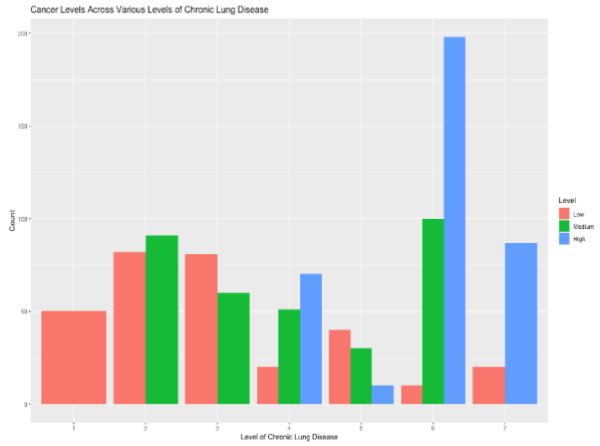


Figure 7. Histogram of Lung Cancer Levels Across Different Levels of Chronic Lung Disease

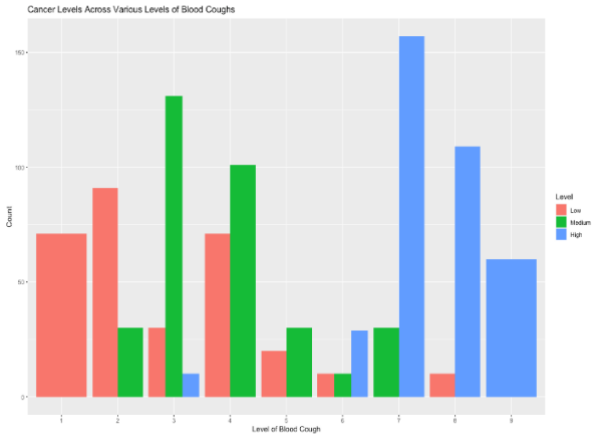


Figure 8. Histogram of Lung Cancer Levels Across Different Levels of Blood Coughs

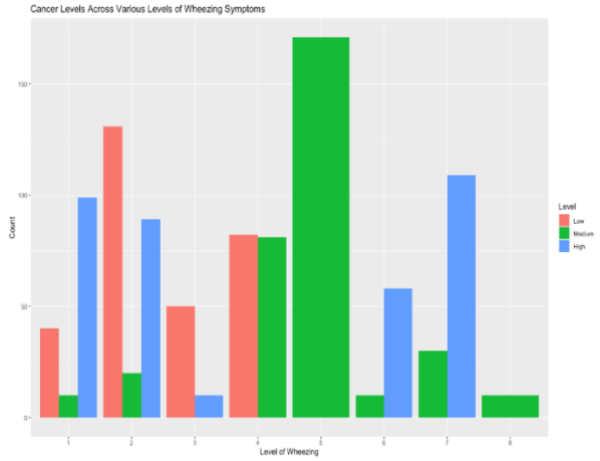


Figure 9. Histogram of Lung Cancer Levels Across Different Levels of Wheezing

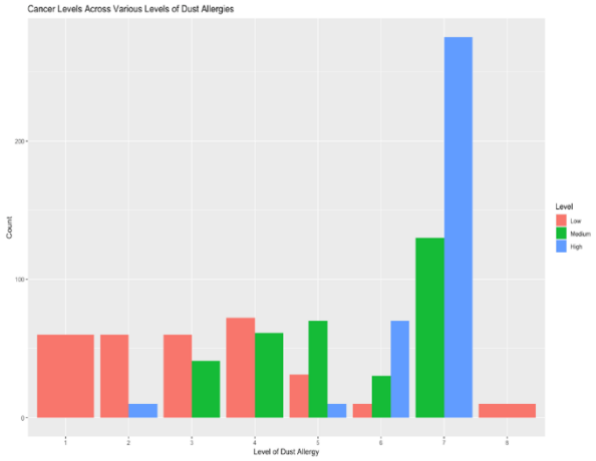


Figure 10. Histogram of Lung Cancer Levels for Different Levels of Dust Allergies

3.0 METHODOLOGY:

3.1 K-Nearest Neighbors:

Hypothesis:

Our hypothesis is that there is a significant relationship between the patient's attributes and the likelihood of having lung cancer.

Model Explanation:

The KNN model is a machine learning algorithm used to perform classification and regression tasks [5]. For each data point, we attempt to determine a value, k, that represents its nearest neighbors using Euclidean distance. After identifying the k nearest neighbors, we correctly assign responses to the testing data we have. We next compare the predicted data to the actual testing data, assessing accuracy by counting the number of correct predictions. In the context of our lung cancer dataset, KNN is an appropriate method for this hypothesis because it can capture complicated correlations between patient features and the probability of developing lung cancer without assuming a certain underlying data distribution [5]. This analysis may provide useful insights for future research, or interventions focused at preventing or treating lung cancer.

Preprocessing:

During the preprocessing step of the analysis, the dataset is standardized using min-max normalization, and is then divided into training and testing sets to assess the model's performance. This division is performed at a 70-30 split, where 70% of the data is used for training and 30% for testing. The data is then filtered to remove instances where the ‘Level’ variable is ‘Medium’, as logistic regression can only be performed with two categorical variables. The ‘Level’ variable is finally converted to a binary outcome, with ‘High’ represented as 1 and ‘Low’ as 0, and this new binary variable is added to the training set as ‘Level binary’.

Result:

The KNN classification model yielded highly accurate results for predicting the severity of lung cancer in patients. The confusion matrix in ***Figure 10*** indicates that the model accurately divided all events into three severity levels (Medium, High, and Low). Moreover, the model's overall statistics show excellent performance, with an accuracy of 1.0, suggesting that it accurately predicted all cases in the test set. Furthermore, the model's sensitivity, specificity, positive predictive value, and negative predictive value were all perfect (1.0) for each severity level as seen in ***Table 2***, demonstrating that the model accurately identified instances of each class while avoiding misclassification.

Table 2. Sensitivity and Specificity of KNN Model

	Medium	High	Low
Sensitivity	1.0	1.0	1.0
Specificity	1.0	1.0	1.0

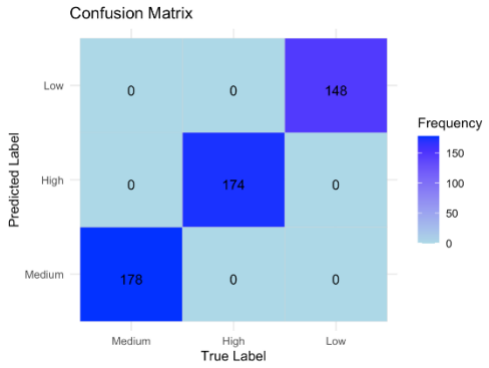


Figure 11. Confusion Matrix of KNN Model

3.2 Logistic Regression:

Hypothesis:

Our hypothesis is that there exists a statistically significant association between the patient's attributes and the probability of having lung cancer, as quantified by the odds ratios of the logistic regression model.

Model Explanation:

Logistic regression is a statistical model used to analyze the relationship between a categorical dependent variable and one or more independent variables [6]. It estimates the probability of the dependent variable occurring based on the independent variables [6]. For our dataset, logistic regression is suitable for this hypothesis because it can model the relationship between patient attributes and the probability of having lung cancer [6]. This allows for quantifying the strength and direction of the association between each attribute and the likelihood of lung cancer, which can help identify potential risk factors.

Preprocessing:

During the preprocessing step of the analysis, the dataset is divided into training and testing sets to assess the model's performance. Specifically, the dataset is then randomly split into training and test sets using a 70-30 split with min-max normalization. The data is then filtered to remove instances where the 'Level' variable is 'Medium', as logistic regression can only be performed with two categorical variables. The 'Level' variable is finally converted to a binary outcome, with 'High' represented as 1 and 'Low' as 0, and this new binary variable is added to the training set as 'Level binary'.

Result:

The performance of the logistic regression model in identifying individuals with diabetes is characterized by a confusion matrix and ROC curve. The confusion matrix in ***Figure 12*** reveals that logistic regression can perfectly classify patients who have "High" or "Low" levels of lung cancer. The accuracy of the prediction yields 100%, revealing how extremely efficient logistic

regression is at classifying patients. Furthermore, the ROC curve in ***Figure 11*** reveals that the logistic regression is accurate at classifying patients as the curve is at the top left corner of the graph, forming a perfect right angle. Additionally, the ROC curve produces an AUC value of 1.0. These results further demonstrate the effectiveness of logistic regression in classifying patients with varying levels of severity.

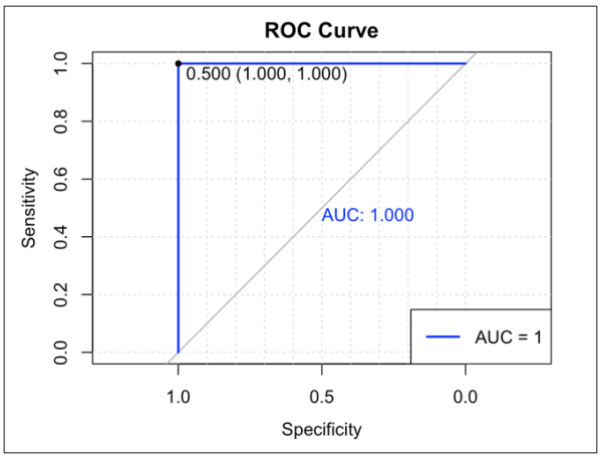


Figure 12. ROC Curve and AUC of Logistic Regression Model

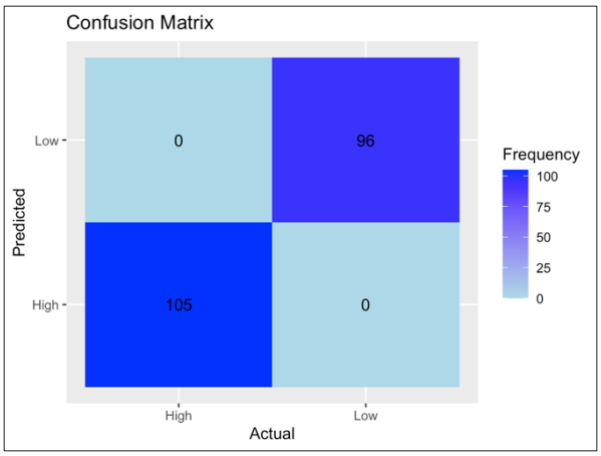


Figure 13. Confusion Matrix of Logistic Regression Model

3.3 Decision Tree

Hypothesis:

Our hypothesis is that the decision tree model can accurately classify the three levels of lung cancer by learning the nonlinear relationships and interactions among various risk factors and symptoms.

Model Explanation:

Decision trees are a powerful machine learning method that excels in modeling nonlinear relationships without assuming any specific distribution of variables [7]. This flexibility allows them to effectively model nonlinear relationships, making them well-suited for datasets where the relationships between features and the response cannot easily captured by linear models [7]. In the context of our lung cancer dataset, this flexibility is particularly beneficial as the dataset contains many variables which may interact in complex ways to influence the likelihood of developing lung cancer. Additionally, decision trees provide a clear and interpretable model, making them suitable for medical settings where transparency in decision-making is important.

Preprocessing:

During the preprocessing step, the dataset is standardized, and is then divided into two subsets: a training set and a testing set. This division is essential to evaluate the performance of the decision tree model. The 80-20 split, where 80% of the data is used for training and 20% for

testing, is used as it is a good balance between having enough data to train the model effectively and having enough data to assess its performance accurately.

Result:

The performance of the decision tree model in identifying individuals with diabetes is characterized by a confusion matrix and decision tree diagram. The confusion matrix in ***Figure 14*** shows the decision tree algorithm can correctly classify two-hundred individuals into their correct severity level. The accuracy of the prediction yields 100%, demonstrating how efficient decision trees are at classifying patients. Moreover, from the decision tree diagram in ***Figure 13***, we can infer that passive smoking is likely the main symptom leading to lung cancer, with coughing of blood possibly being a critical symptom. Likewise, wheezing may be correlated with passive smoking, and obesity might be an influencing factor for wheezing.

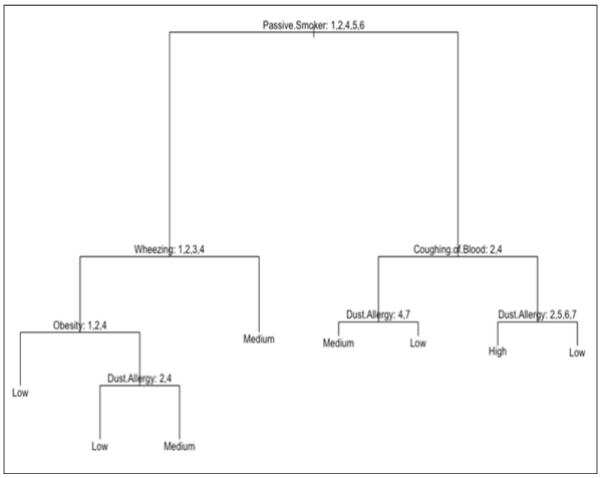


Figure 14. Decision Tree Model Results

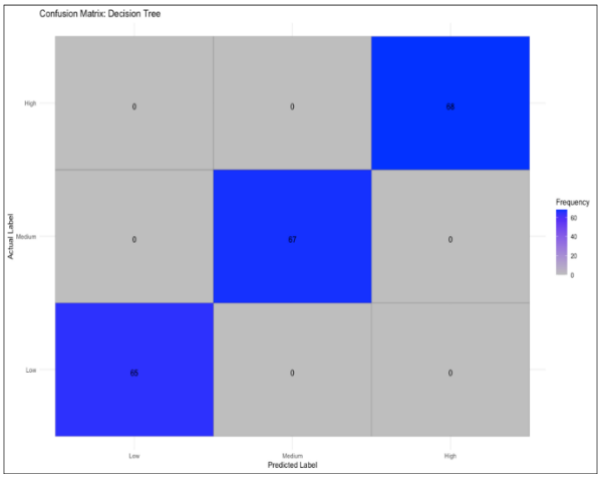


Figure 15. Confusion Matrix of Decision Tree Model

3.4 Random Forest

Hypothesis:

Our hypothesis is that the random forest model will effectively classify the levels of lung cancer by capturing complex interactions and non-linear relationships among the various predictors within the dataset.

Model Explanation:

Random forests are a learning method that builds multiple decision trees during training and outputs the mode of the variable or the mean prediction of the individual trees [8]. This approach makes random forests robust to outliers and noise, reducing the risk of overfitting, especially in datasets with many features like our medical dataset [8]. Additionally, random forests can provide estimates of feature importance, allowing us to identify the most significant predictors out of potentially many correlated variables [8]. In the context of our lung cancer dataset, this

can help us understand which risk factors or symptoms are most indicative of having lung cancer, providing valuable insights for further research or interventions.

Preprocessing:

During the preprocessing step, the dataset is standardized, and is then divided into two subsets: a training set and a testing set. This division is essential to evaluate the performance of the random forest model. The 80-20 split, where 80% of the data is used for training and 20% for testing, is used as it is a good balance between having enough data to train the model effectively and having enough data to assess its performance accurately.

Result:

From the confusion matrix of our Random Forest model in ***Figure 16***, it can easily be determined that it classified each corresponding cancer level with 100% accuracy. The MeanDecreaseAccuracy plot as revealed in ***Figure 15***, implies which variable would lead to the greatest decrease in classification accuracy, if removed from the dataset. Observing the plot, we can see that wheezing, obesity, and passive smoking appear to be the three most important and influential predictors in the dataset leading to an accurate classification, where wheezing would result in nearly 40% decrease in accuracy. Additionally, we can see that passive smoking had the highest importance score, indicating that it's very influential in the model's decisions. Obesity and wheezing again appear as important features which are consistent with the other plot.

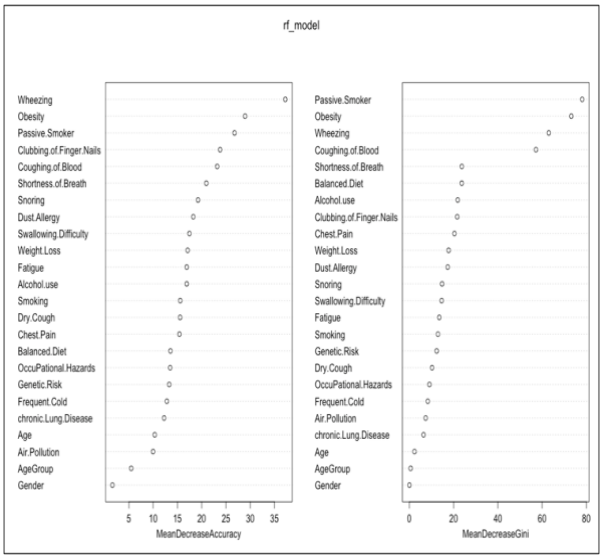


Figure 16. Variable Importance Plot from Random Forest

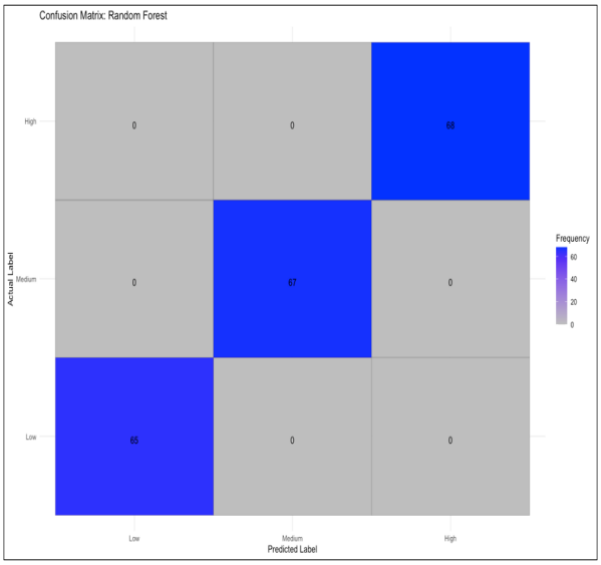


Figure 17. Confusion Matrix of Random Forest Model

4.0 Result Analysis and Discussion

Table 3. Performance Comparison of All Four Statistical Methods

Model	Accuracy	Kappa
K-Nearest Neighbors (KNN)	100%	1.0
Logistic Regression	100%	1.0
Decision Tree	100%	1.0
Random Forest	100%	1.0

The results show that accuracy, measured by the proportion of correct classifications within the total number of data points is 100%, indicating that all predictions made were correct. The table also reveals the Kappa Statistic, which measures the level of agreement between the models' classifications and the true level. The Kappa Statistic for each model produced 1.0, indicating perfect agreement. Both accuracy and Kappa Statistics achieving perfect scores for all models in the real world is nearly improbable, and the limitations of these results will be discussed in the following section.

4.1 Limitations

As aforementioned in the previous section, both the accuracy and Kappa statistic had perfect scores for all our models, indicating that the dataset had been modified or was created artificially to attain 100% accuracy. Therefore, it is difficult to draw conclusions from this dataset as meaningful and applicable in the real world. However, other limitations include the analysis relying on a single dataset and the fact that the sample size of a thousand individuals may not be representative of the entire population, both of which may limit the generalizability of the findings to other populations or settings.

4.2 Solutions

In response to the identified limitations, several strategic measures can be implemented to enhance the reliability and applicability of the models. To begin, rigorous validation procedures should be employed to verify the integrity and representativeness of the dataset. Cross-validation techniques can be utilized to assess the robustness of the models and ensure their effectiveness across different subsets of the data. Furthermore, external validation using independent datasets is imperative to validate the generalizability of the models beyond the training data. Evaluating model performance on unseen data ensures the reliability and robustness of the findings in real-world scenarios. Moreover, expanding the sample size of the dataset can reinforce the analysis and enhance the precision of the estimates. A larger sample size allows for more reliable inferences and strengthens the validity of the conclusions drawn from the data. These strategic measures can improve the reliability, validity, and applicability of the models in practical settings.

4.3 Lessons Learned

Regarding the coding process, our group learned the importance of choosing high-quality data datasets, as our results indicated that our datasets were modified to attain 100% accuracy, making it doubtful that our trained models would be useful for other test datasets. We also went through the process of organizing and distributing work amongst a group of strangers (as we had not met before the project), improving our communication and collaboration skills.

5.0 REFERENCES

[1] Canadian Cancer Society. (2023, November). *Lung and bronchus cancer statistics*. Retrieved April 19, 2024, from https://cancer.ca/en/cancer-information/cancer-types/lung/statistics

[2] World Health Organization: WHO. (2023, June 26). *Lung cancer*. World Health Organization. Retrieved April 19, 2024, from https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[3] Thanoon, M. A., Zulkifley, M. A., Zainuri, M. A. A. M., & Abdani, S. R. (2023, August 8). *A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images*. National Center for Biotechnology Information. Retrieved April 19, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10453592/

[4] *Lung Cancer Prediction*. (2022, November 14). Kaggle. Retrieved April 19, 2024, from https://www.kaggle.com/datasets/thedevastator/cancer-patients-and-air-pollution-a-new-link

[5] IBM. (n.d.). *What is the KNN algorithm?* Retrieved April 19, 2024, from https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN)%20algorithm%20is%20a%20non,used%20in%20machine%20learning%20today.

[6] Lawton, G., Burns, E., & Rosencrance, L. (2022, January). *logistic regression*. Tech Target. Retrieved April 19, 2024, from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression#:~:text=Logistic%20regression%20is%20a%20statistical,or%20more%20existing%20independent%20variables.

[7] IBM. (n.d.-a). *What is a decision tree?* Retrieved April 19, 2024, from https://www.ibm.com/topics/decision-trees