LUNGS CANCER PREDICTION USING MACHINE LEARNING MODELS

Student ID: D3622546

Student Name: Amoo Adura

Total Word count = 2100 without

reference and table of content

Table of Contents

Abstract

This study explores the application of machine learning models for predicting lung cancer risk levels. Leveraging logistic regression, Gaussian Naïve Bayes, and Multinomial Naïve Bayes algorithms, the research evaluates their performance in accurately categorizing patients into high, medium, and low-risk groups. Initial assessments showcase strong predictive capabilities, with logistic regression emerging as the top performer. Further model tuning enhances accuracy, particularly in logistic regression, achieving perfect classification. K-fold cross-validation confirms the robustness of logistic regression and Gaussian Naïve Bayes, while Multinomial Naïve Bayes exhibits moderate performance. These results highlight how machine learning can help with lung cancer earlier identification and risk assessment, potentially improving patient outcomes.

Keyword: Prediction models, Lung Cancer, Logistic Regression, Gaussian Naïve Bayes and Multinomial Naïve Bayes.

1.0 Introduction

Lung cancer remains a major global health challenge and is the second greatest cause of mortality globally. Over 2.2 million fresh cases were diagnosed in 2020 only. Primary risk factors include tobacco smoking, though significant percentages of non-smokers also get lung cancer as a result of breathing in second-hand smoking, being near radon and asbestos,and genetic factors. The absence of symptoms in early stages makes lung cancer particularly perilous, underscoring the importance of proactive screening for at-risk populations (Kocarnik *et al.*, 2022).

Recent advancements in Information Technology have significantly enhanced the capabilities for early diagnosis. Machine Learning (ML), with its diverse methodologies such as supervised, unsupervised, and reinforcement learning, plays a crucial role in this advancement. These technologies are adept at analyzing complex datasets, enabling the categorization of patient data with high precision, thus improving predictive outcomes. For example, studies have demonstrated that ML algorithms like Voting Classifiers, Random Forest, Neural Networks, and Support Vector Machines can be particularly effective in predicting lung cancer, with a Voting Classifier achieving an accuracy rate of 99.5% (Thallam *et al.*, 2020).

Moreover, the integration of machine learning with image processing techniques has proven to be highly effective in the classification and prediction of lung cancer. Advanced preprocessing methods like geometric filtering enhance the quality of medical images, while segmentation techniques such as the K-means method facilitate the precise identification of regions of interest. This synergy between ML and image processing is pivotal in developing systems that achieve high accuracy in lung cancer detection (Nageswaran *et al.*,).

Despite these advances, challenges remain, particularly concerning the accuracy and reliability of ML models. Studies have noted that many ML-based studies fail to address issues such as missing data, and lack of normalization and standardization, which can introduce bias and affect the outcomes of the studies. This is especially problematic in non-image-based datasets, which are less common than those for other cancers such as breast cancer (Altuhaifa, Win and Su, 2023) .

Further research has explored the effectiveness of ensemble classifiers that combine several ML techniques to enhance diagnostic accuracy, particularly in imaging tests like CT scans. These classifiers, which might include methods like SVM, KNN, MLP, and Decision Trees, show promise in achieving high diagnostic accuracy, approaching that of more traditional methods like the Random Forest classifier (Shanbhag *et al.*, 2022). Continued development and evaluation of these ensemble methods could potentially offer even more effective tools for the early identification and treatment of lung cancer, as evidenced by various studies that compare the performance metrics of different ML algorithms (Jenipher and Radhika, 2020) .

This ongoing research and development in ML for lung cancer detection underscore the significant potential of technology to improve outcomes in cancer diagnosis and treatment, paving the way for innovations that could substantially reduce mortality rates associated with the disease.

2.0 Methods

The methodology can be separated into several stages and the machine learning framework can be described in Figure 1.

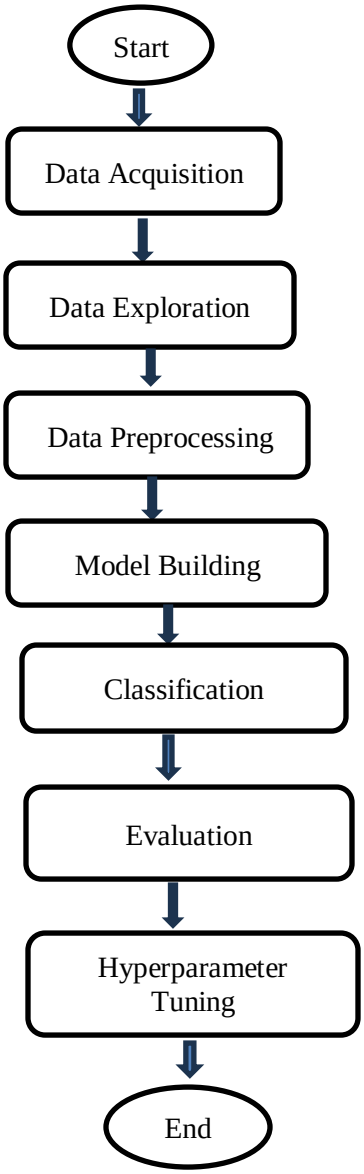

Figure 1. proposed machine learning methodology for lung cancer predictions

2.1 Data Collection

The study's use of lung cancer dataset came from the open-access Kaggle repository (https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer). There are 26 characteristics in the dataset and 1000 observations. The data type compromised of 1 categorical data and 25 numerical data. The target variable in the dataset is denoted as 'Level' and is represented as a multi-class category with 2 denoted as 'High risk', 1 as 'medium Risk' and 0 as 'low risk' of having lung cancer respectively.

	index	Patient Id	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails
0	0	P1	33	1	2	4	5	4	3	2	...	3	4	2	2	3	1
1	1	P10	17	1	3	1	5	3	4	2	...	1	3	7	8	6	2
2	2	P100	35	1	4	5	6	5	5	4	...	8	7	9	2	1	4
3	3	P1000	37	1	7	7	7	7	6	7	...	4	2	3	1	4	5
4	4	P101	46	1	6	8	7	7	7	6	...	3	2	4	1	4	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	995	P995	44	1	6	7	7	7	7	6	...	5	3	2	7	8	2
996	996	P996	37	2	6	8	7	7	7	6	...	9	6	5	7	2	4
997	997	P997	25	2	4	5	6	5	5	4	...	8	7	9	2	1	4
998	998	P998	18	2	6	8	7	7	7	6	...	3	2	4	1	4	2
999	999	P999	47	1	6	5	6	5	5	4	...	8	7	9	2	1	4

1000 rows × 26 columns

Table 1. Showing the dataset observation

2.2 Data Exploration

Prior to partitioning the data into training and testing sets for machine learning model fitting, meticulous data examination is imperative. This process enhances comprehension of data quality and alerts to potential issues. Data exploration facilitates summarization, visualization, and a more comprehensive quantitative understanding of the dataset.

- Statistical distribution and visual of the data

Table 2 illustrates the statistical distribution of the data.

	index	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	...	Coughing of Blood	
count	1000.000000	1000.000000	1000.000000	1000.0000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	1000.000000	...	1000.000000	100
mean	499.500000	37.174000	1.402000	3.8400	4.563000	5.165000	4.840000	4.580000	4.380000	4.491000	...	4.859000	
std	288.819436	12.005493	0.490547	2.0304	2.620477	1.980833	2.107805	2.126999	1.848518	2.135528	...	2.427965	
min	0.000000	14.000000	1.000000	1.0000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	
25%	249.750000	27.750000	1.000000	2.0000	2.000000	4.000000	3.000000	2.000000	3.000000	2.000000	...	3.000000	
50%	499.500000	36.000000	1.000000	3.0000	5.000000	6.000000	5.000000	5.000000	4.000000	4.000000	...	4.000000	
75%	749.250000	45.000000	2.000000	6.0000	7.000000	7.000000	7.000000	7.000000	6.000000	7.000000	...	7.000000	
max	999.000000	73.000000	2.000000	8.0000	8.000000	8.000000	8.000000	7.000000	7.000000	7.000000	...	9.000000	

8 rows × 24 columns

Table 2. Descriptive statistic of the numerical features of the dataset

The utilization of the seaborn visualization tool in Figure 2 enables a comprehensive analysis of the correlation between the target values and the categorical values. The results suggest a higher propensity for lung cancer development among smokers compared to non-smokers.

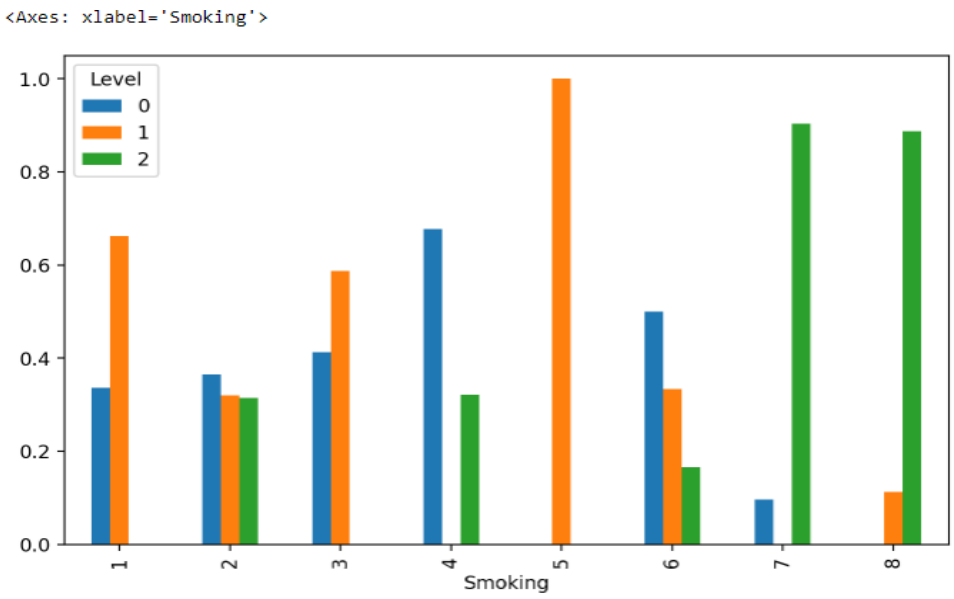

Figure 2. Distribution of Smoking and Target Variable using Count Plot

Figure 3 illustrates the process that is undertaken to acquire insights and ascertain the presence of outliers. While the removal of outliers enhances prediction accuracy and efficiency, within this study, the singular outlier pertains to age, potentially exerting an influence on the dataset.

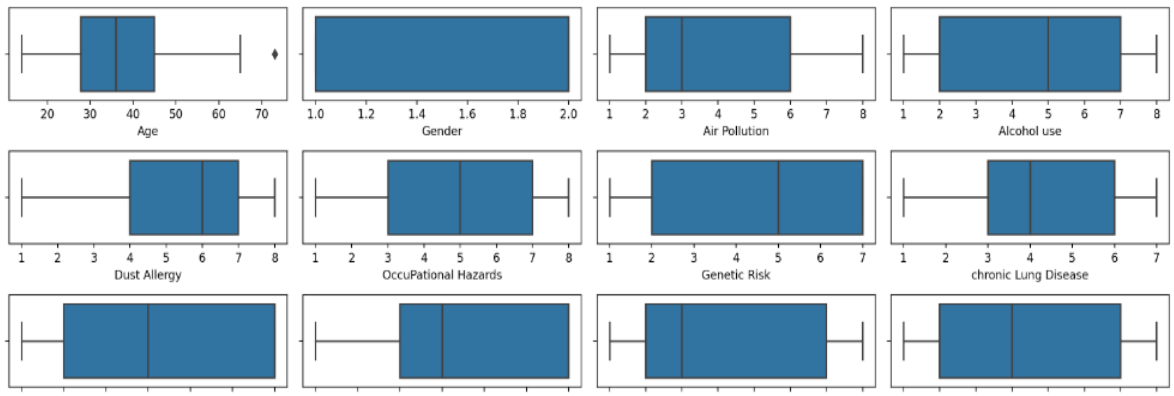

Figure 3. Viewing Outliers.

- Correlation

The heatmap visually represents the correlation matrix, providing valuable insights into the relationships between variables. By displaying the correlation coefficients as color gradients, it offers a comprehensive overview of how each variable is related to others in the dataset. This analytical tool aids in identifying patterns, dependencies, and potential multicollinearity, thereby informing subsequent data analysis and modeling decisions.

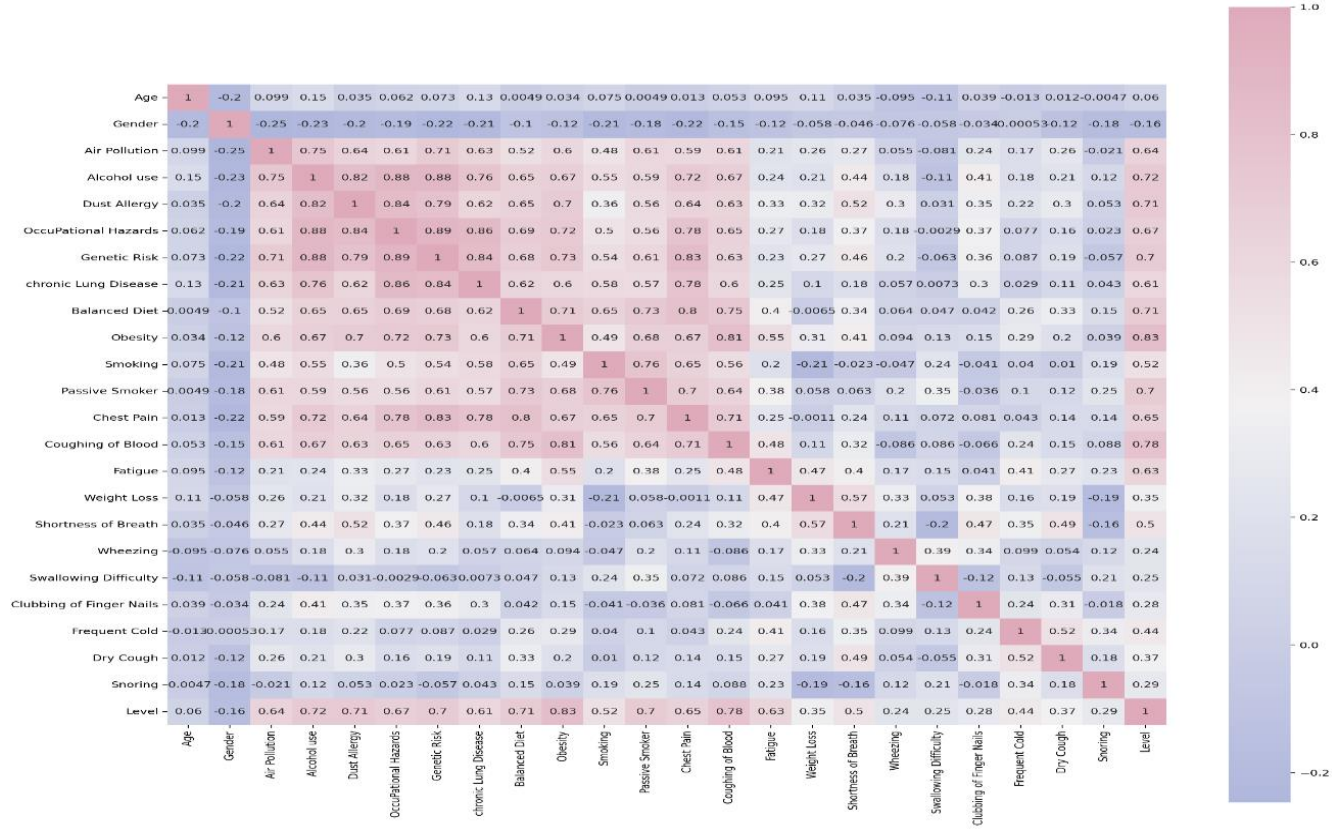

Figure 4. Correlation Matrix visualization.

2.3 Data Preprocessing

Firstly, the dataset was added to the python IDE , and necessary libraries are imported into the pandas frame of data to facilitate tabular viewing. Statistical distribution insights, such as mean, median, and standard deviation, were obtained through the describe() method. Null value detection was conducted using the Info() function, revealing an absence of null values. Categorical values were numerically encoded to enable seamless classification.

To enhance model accuracy, unique features such as index and patient ID were dropped as they could skew model predictions. An 80-20 split was employed, allocating 80% for the purpose of training the machine learning model and setting aside 20% for examination of its performance on unseen data.

During data analysis, extreme values were scrutinized, with only outliers detected in the Age feature, deemed non-influential to the dataset as illustrated in Table 2. Categorical data within the target variable were converted to numerical format.

A check on the distribution of the target variable indicated a balanced distribution, obviating the need for balancing techniques like SMOTE. Normalization was performed using the MIN-MAX scalar. It's imperative that standardization is fitted solely on the training dataset to prevent data leakage into the testing process. Subsequently, the test dataset was standardized using the parameters derived from the training dataset.

Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty	Clubbing of Finger Nails	Frequent Cold	Dry Cough	Snoring	Level
5	4	3	2	2	4	...	3	4	2	2	3	1	2	3	4	0
5	3	4	2	2	2	...	1	3	7	8	6	2	1	7	2	1
6	5	5	4	6	7	...	8	7	9	2	1	4	6	7	2	2
7	7	6	7	7	7	...	4	2	3	1	4	5	6	7	5	2
7	7	7	6	7	7	...	3	2	4	1	4	2	4	2	3	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7	7	7	6	7	7	...	5	3	2	7	8	2	4	5	3	2
7	7	7	6	7	7	...	9	6	5	7	2	4	3	1	4	2
6	5	5	4	6	7	...	8	7	9	2	1	4	6	7	2	2
7	7	7	6	7	7	...	3	2	4	1	4	2	4	2	3	2
6	5	5	4	6	7	...	8	7	9	2	1	4	6	7	2	2

Table 3. Conversion of categorical to numerical variable

	Age	Gender	Air Pollution	Alcohol use	Dust Allergy	OccuPational Hazards	Genetic Risk	chronic Lung Disease	Balanced Diet	Obesity	...	Fatigue	Weight Loss	Shortness of Breath	Wheezing	Swallowing Difficulty
0	33	1	2	4	5	4	3	2	2	4	...	3	4	2	2	3
1	17	1	3	1	5	3	4	2	2	2	...	1	3	7	8	6
2	35	1	4	5	6	5	5	4	6	7	...	8	7	9	2	1
3	37	1	7	7	7	7	6	7	7	7	...	4	2	3	1	4
4	46	1	6	8	7	7	7	6	7	7	...	3	2	4	1	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	44	1	6	7	7	7	7	6	7	7	...	5	3	2	7	8
996	37	2	6	8	7	7	7	6	7	7	...	9	6	5	7	2
997	25	2	4	5	6	5	5	4	6	7	...	8	7	9	2	1
998	18	2	6	8	7	7	7	6	7	7		3	2	4	1	4

Table 4. Drooping Features

2.4 Machine Learning Methods Applied

This study applied classification algorithms to a multi-class problem; high risk, medium risk and low risk of having lung cancer respectively. Several machine learning model, namely, logistic regression, gaussian naïve bayes, and multinomial naïve bayes was used in this study. Categorical features were encoded, and numerical features were standardized. These techniques were implemented in Python using the Sci-kit learn library.

- Logistic Regression Classifier

Logistic Regression is a statistical technique primarily employed in binary classification problems, but it also use in multi-class which aims to predict a multiple output variable (e.g., 0 , 1, and 2 high, medium and low respectively) based on given input features. At its core, Logistic Regression utilizes the logistic function to model the connection between the outcome and the input variables.

The logistic regression model computes the log-odds of the likelihood that the positive outcome occurs, which is then transformed using the logistic function to yield the probability itself. This probability serves as the basis for class assignment.

The hyperparameter tuning for the Logistic Regression model is perform and found that the best combination of hyperparameters is as follows: the regularization strength (C) is set to 10, the penalty term is 'l1', and the solver used for optimization is 'liblinear'.

- Gaussian Naïve Bayes Classifier

The probability classification method Gaussian Naive Bayes (GNB) uses the "naive" notion that characteristics are unaffected by the class label and assumes that they have a Gaussian (normal) distribution. GNB demonstrates robust performance even with limited training data, as it quick by estimating the model estimates (mean and variance) individually for every characteristic and class, particularly advantageous for substantial datasets.

Hyperparameter tuning for GNB is typically unnecessary due to its minimal number of hyperparameters. However, techniques like cross-validation are employed for parameter estimation. The cross-validation scores for GNB are as follows: [0.90625 0.9125 0.85 0.8875 0.925].

- Multinomial Naïve Bayes Classifier

Multinomial Naive Bayes (MNB) is a classification model suitable for text categorization tasks, where features represent word counts or frequencies. It assumes a multinomial distribution of features and makes the 'naive' assumption of feature independence given the class label. Despite its simplicity, MNB can achieve strong performance in text classification tasks, particularly with large and sparse datasets typical in natural language processing (NLP).

Hyperparameter tuning for MNB is straightforward due to its limited number of hyperparameters. Typically, tuning involves parameters such as the smoothing factor (alpha), which aids in handling unseen features. In the context of hyperparameter tuning, the best parameters found for MNB are {'alpha': 0.1, 'fit_prior': True}.

2.5 Performance Evaluation

With K-fold cross-validation, the three algorithms' efficiency was evaluated, and employing the area under the curve (AUC) metric. AUC serves as a robust measure for classification performance, capturing the models' ability to discriminate between randomly selected positive and negative instances. It quantifies the balance or imbalance between the true positive rate and the false positive rate, offering a comprehensive view of model performance. A higher AUC score indicates superior discriminatory power and better overall performance, making it a valuable tool for comparing and explaining the effectiveness of different models.

One popular method in machine learning for evaluating models is K-fold cross-validation. Its primary purpose is to evaluate a prediction model's generalisation performance to a separate dataset. The process involves partitioning the the initial dataset into k roughly equal-sized groups, or "folds," After that, the model is trained k times, utilising the extra fold for validation and k-1 folds for training every single time.

3.0 Results and Discussion

Following model fitting, predictions were made on both the training and testing datasets. Subsequently, the models underwent evaluation based on metrics such as accuracy, area under the curve (AUC), and cross-validation.

3.1 Model Performance Evaluation

Table 5: Classification Metrics Result

Models	Accuracy	Precision	Recall
Logistic Regression	0.97	0.98	0.97
Gaussian Naïve Bayes	0.88	0.88	0.88
Multinomial Naïve Bayes	0.70	0.71	0.70

Table 6: Hyper-parameter Tuning Result

Model	Accuracy
Logistic Regression	1.00
Gaussian Naïve Bayes	0.89
Multinomial Naïve Bayes	0.70

Table 7: K-fold Cross Validation Result

Model	Accuracy
Logistic Regression	1.00
Gaussian Naïve Bayes	0.89
Multinomial Naïve Bayes	0.70

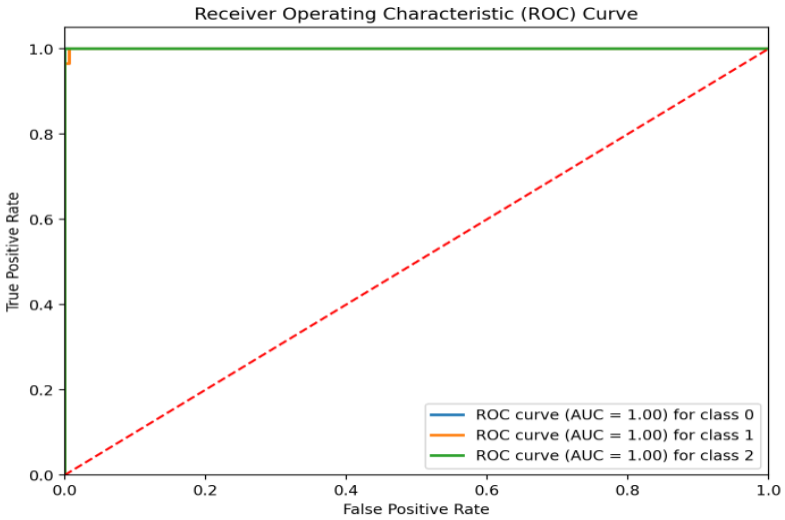

Figure 8: ROC of Logistic Regression

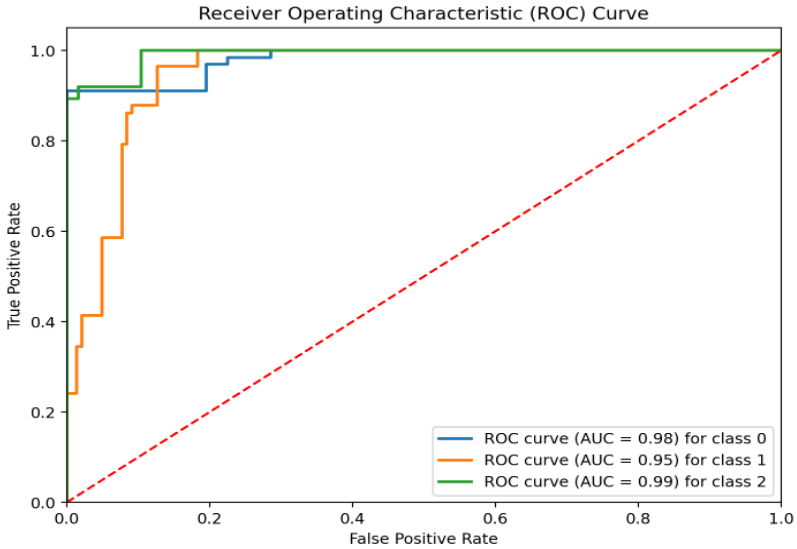

Figure 9: ROC of Gaussian Naïve Bayes

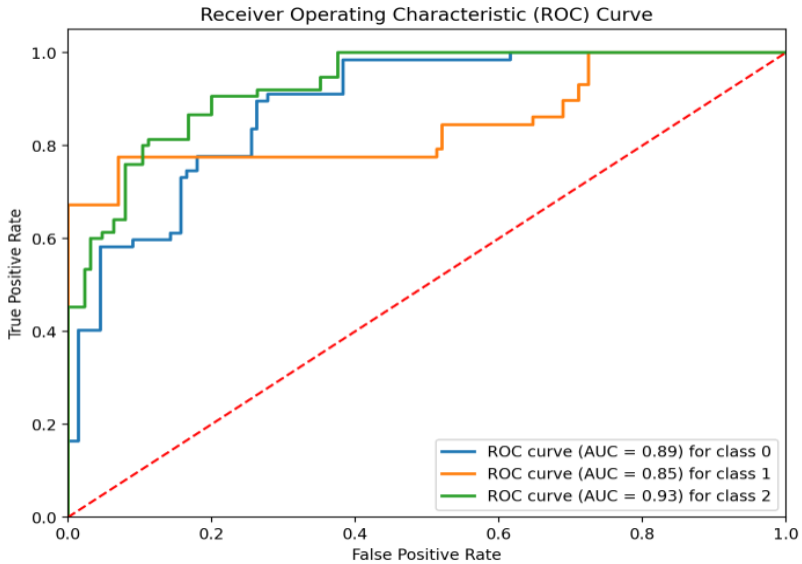

Figure 10: ROC of Multinomial Naïve Bayes

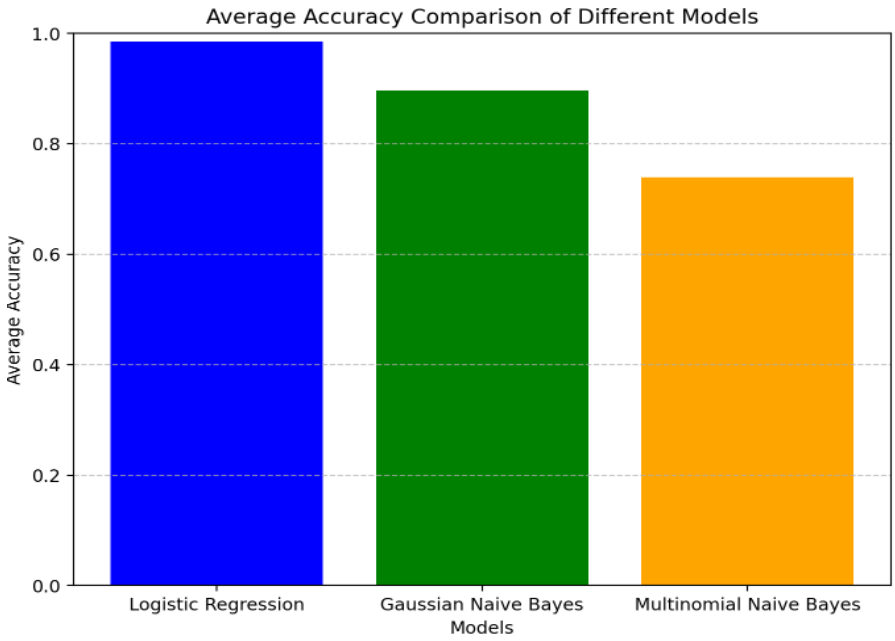

Figure 11: K-fold Cross Validation

The model results indicate strong performance across all models in predicting patients' risk levels for lung cancer- high, medium, and low. Logistic regression stands out with exceptional accuracy (97%), precision (98%), and recall (97%), followed closely by Gaussian Naïve Bayes with an accuracy of 88%. However, Multinomial Naïve Bayes shows lower precision (71%), suggesting less accuracy in identifying patients with lung cancer. Further tuning of the models was conducted to improve performance. Post-tuning, Logistic regression achieved a perfect accuracy of 100%, followed by Gaussian Naïve Bayes at 89%, and Multinomial Naïve Bayes at 70%. In K-fold cross-validation, Logistic regression and Gaussian Naïve Bayes continued to perform strongly, both achieving accuracies of 100% and 89%, respectively, while Multinomial Naïve Bayes exhibited moderate performance at 70%.

4.0 Conclusion

In conclusion, the application of machine learning models, including logistic regression, Multinomial Naïve Bayes (MNB), and Gaussian Naïve Bayes (GNB), shows promising results in predicting lung cancer risk levels. Logistic regression emerges as the top performer, exhibiting exceptional accuracy, precision, and recall rates. Despite initial discrepancies, further model tuning enhances overall performance across all algorithms. Notably, GNB and MNB demonstrate respectable accuracies, with GNB outperforming MNB in most metrics. These results highlight how machine learning methods can help in early diagnosis and risk assessment of lung cancer, thereby contributing to improved patient outcomes and healthcare decision-making.

References

Altuhaifa, F.A., Win, K.T. and Su, G. (2023) 'Predicting lung cancer survival based on clinical data using machine learning: A review', *Computers in Biology and Medicine,* , pp. 107338.

Jenipher, V.N. and Radhika, S. (2020) 'A study on early prediction of lung cancer using machine learning techniques', *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS).* IEEE

Kocarnik, J.M. *et al.* (2022) 'Cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life years for 29 cancer groups from 2010 to 2019: a systematic analysis for the global burden of disease study 2019', *JAMA Oncology,* 8(3), pp. 420-444.

Nageswaran, S. *et al.* '[Retracted] Lung Cancer Classification and Prediction Using Machine Learning and Image Processing', *BioMed Research International,* 2022

Shanbhag, G.A. *et al.* (2022) 'Prediction of lung cancer using ensemble classifiers', *Journal of Physics: Conference Series.* IOP Publishing

Thallam, C. *et al.* (2020) 'Early stage lung cancer prediction using various machine learning techniques', *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA).* IEEE