A Comparative Study on Lung Cancer Prediction using Naïve Bayes and Logistic Regression

Alyssa Alexandra Lee[1], Aren Deza[1], and Joseph Victor Sumbong[1]

University of The Philippines Visayas, Miag-ao, Iloilo City

1 Introduction

1.1 Background of the Study

Lung cancer is a disease characterized by a malignant growth or tumor formed from the uncontrollable division of abnormal cells, specifically within the lungs [2]. It is known to be the most common type of cancer and is also the main cause of cancer death globally. The chance to get and develop this disease increases with a myriad of factors such as tobacco smoking, exposure to chemicals in the workplace, long-term exposure to air pollution, exposure to radon gas, having a family history of lung cancer, etc.

In the Philippines, it is common to encounter a person smoking as you walk down the street. According to MacroTrends [6], the prevalence of Filipino smokers aged 15 and above on a daily or non-daily basis is 22.90% in the year 2020. The population density in the Philippines is also greatest in the National Capital Region with 21,765 persons per square kilometer [7]. This region also accounts for 13,484,462 persons of the 109,033,245 or 12.37% total population of the country. In NCR, the most densely populated city was Manila with 73,920 persons per square kilometer. According to IQAir [5], in 2019, Manila had an average of 18.2 µg/m³ of PM2.5 or particulate matter under 2.5 micrometers which poses health concerns. This recording placed them under the "Moderate" category of the World Health Organization (WHO). However, with the lockdown being lifted and traffic becoming more of a problem, the air quality is forecasted to worsen. In a press release by the Department of Health in 2022 [3], Lung cancer is the 2nd leading type of cancer in the country and is the leading cause of mortality in all cancer types.

If it is identified and diagnosed in its earliest stages, lung cancer can fortunately be treated. However, a patient suffering from lung cancer can be asymptomatic while the disease is in its earlier stages of progression, which can make it difficult to detect. For this reason, having the ability to quickly and reliably predict whether a patient may have or is at risk of contracting lung cancer can be invaluable in reducing their chances of dying from the disease. Therefore, the researchers aim to employ machine learning technologies to identify significant factors among lung cancer patients and construct a model that can be used to

predict and evaluate an individual's risk of contracting the disease while choosing the best machine learning algorithm that the researchers have learned.

There have been studies such as "A Study On Prediction Of Lung Cancer Using Machine Learning Algorithms" by Ahmad et al. [4] which compared 3 machine learning algorithms in their ability to predict lung cancer, however, their dataset was biomedical images and they had an image processing step. Other similar studies also used a collection of labeled images for their dataset. A similar study that did not utilize an image dataset is "A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms" by Radhika et al. [8] It utilized Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and Decision Tree machine learning algorithms as well as a different dataset than the researchers will use. This paper differs from the aforementioned study due to its different dataset as well as the machine learning algorithms that will be utilized and compared.

1.2   Statement of the Problem

The researchers will implement and compare Machine learning models that can identify an individual's likelihood of developing lung cancer. Specifically, the researchers will:

1. train a Logistic Regression model to predict likelihood of lung cancer;
2. train a Naive Bayes model to predict likelihood of lung cancer;
3. compare Naive Bayes and Logistic Regression's evaluation scores;
4. determine the top 3 factors for people with high likelihood of contracting lung cancer;
5. determine the top 3 factors for people with medium likelihood of contracting lung cancer; and
6. determine the top 3 factors for people with low likelihood of contracting lung cancer.

1.3   Hypothesis

The researchers will not be able to implement a Machine learning model that can identify an individual's likelihood of developing lung cancer.

1.4   Scope and Delimitation

The scope of this paper is to determine only lung cancer likelihood using the dataset of the user, "The Devastator", from the website, Kaggle. It will also only implement and compare Naive Bayes and Logistic Regression as its Machine Learning models. The researchers will make use of the following tools: (1) Jupyter Notebook, (2) Google Docs, (3) Github, and (4) TeXworks. The LaTeX packages used are the following: (1) color, (2) graphicx, (3) amsmath, (4) listings, (5) pdflscape, (6) geometry, and (7) natbib. The Python packages

that will be employed are: (1) Pandas, (2) NumPy, (3) scikit-learn, and (4) Matplotlib which will be used in implementing the models.

The paper will not use any other dataset aside from the one mentioned above. It will also not predict whether a patient has lung cancer or not, but rather only the patient's likelihood of developing lung cancer. Additionally, it will not predict the likelihood of other types of cancer. It will only use the following features: (1) age; (2) gender; (3) level of air pollution exposure; (4) level of alcohol use; (5) level of dust allergy; (6) level of occupational hazards; (7) level of genetic risk; (8) level of chronic lung disease; (9) level of balanced diet; (10) level of obesity; (11) level of smoking; (12) level of passive smoker; (13) level of chest pain; (14) level of coughing of blood; (15) level of fatigue; (16) level of weight loss; (17) level of shortness of breath; (18) level of wheezing; (19) level of swallowing difficulty; (20) and level of clubbing of finger nails.

2   Methodology

2.1   Dataset

The dataset contains information on patients with lung cancer. This includes background information such as their age, gender, exposure to air pollution, alcohol use, occupational hazards, genetic risk, balanced diet, obesity, history of smoking or passive smoking, and presence of dust allergies and chronic lung disease. It also provides data of their lung cancer symptoms based on severity, which includes chest pain, coughing of blood, fatigue, weight loss, shortness of breath, wheezing, difficulty swallowing, clubbing of finger nails, snoring, and the level of lung cancer. This dataset was retrieved from the website, Kaggle.

To use this dataset for the purpose of this study, the researchers excluded the columns of *Patient Id, Frequent Cold, Snoring,* and *Dry Cough.* The remaining columns excluding Level are designated as the features that are of interest, while the Level column values are used as the true label. The values of the features were then standardized. The manipulated dataset was then split with a 1:3 ratio, resulting in 75% of the entries being the training set and 25% being the testing set.

2.2   Implementation of Machine Learning Models

The machine learning models to be used in this paper are Logistic Regression and Naive Bayes. The scikit-learn library was used to create these models, specifically, the LogisticRegression function for the creation of the Logistic Regression Model which is not only binary but is able to predict multiple outcomes, and the GaussianNB function for the creation of the Naive Bayes model which can also predict multiple outcomes. They were both trained using the same training set as mentioned in the Dataset portion of this paper.

2.3   Metrics of Evaluation

Each of the created model's performance was measured based on their accuracy, precision, recall, f1-score, and confusion matrix for both their training and testing sets. The evaluation was done using sklearn's accuracy_score, classification_report, and confusion_matrix function, while its ConfusionMatrixDisplay function was used to show the confusion matrix.
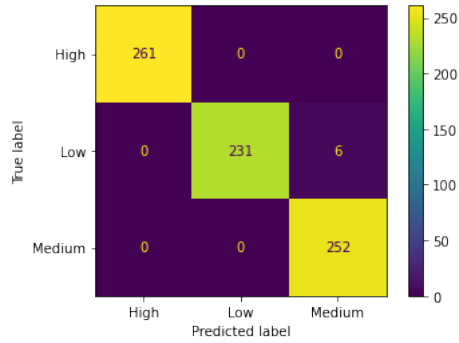
3   Results

3.1   Logistic Regression

Table 1. Logistic Regression Training Set

	precision	recall	f1-score	support
High	1.00	1.00	1.00	261
Low	1.00	0.97	0.99	237
Medium	0.98	1.00	0.99	252
accuracy			0.99	750
macro avg	0.99	0.99	0.99	750
weighted avg	0.99	0.99	0.99	750

Within the training set, the Logistic Regression model possessed an accuracy of 99.2%, with the macro average of precision, recall, and f1-score all being 99%.

Fig. 1. Confusion matrix for Linear Regression on training set
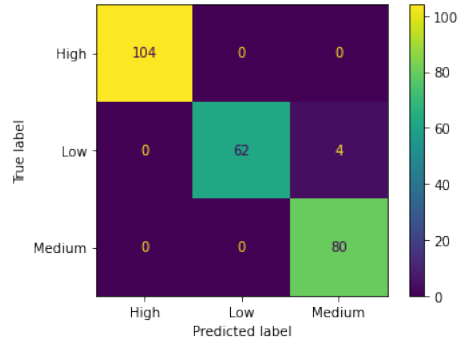

The confusion matrix generated for the training set indicated that the model correctly identified all patients who had High and Low risks of lung cancer, and only mislabeling 6 patients who had a Medium risk of lung cancer as low risk.

Table 2. Logistic Regression Testing Set

	precision	recall	f1-score	support
High	1.00	1.00	1.00	104
Low	1.00	0.94	0.97	66
Medium	0.95	1.00	0.98	80
accuracy			0.98	250
macro avg	0.98	0.98	0.98	250
weighted avg	0.98	0.98	0.98	250

Meanwhile, when used on the testing set, the model had an accuracy of 98.4%, with macro average of precision, recall, and f1-score all being 98%.

Fig. 2. Confusion matrix for Linear Regression on testing set

Its generated confusion matrix displays similar results to the training set's confusion matrix, where the model correctly identified patients with both High and Low risks of lung cancer, and only mislabeled 4 patients with Medium risk of lung cancer as low risk.

Based on the results obtained from using the model in the training and testing datasets, it can be asserted that the model is highly accurate in its predictions, and can be reliably used to identify how likely it is for a given patient to contract lung cancer.

3.2   Naive Bayes

Table 3. Naive Bayes Training Set

	precision	recall	f1-score	support
High	0.82	0.95	0.88	261
Low	0.88	0.92	0.89	237
Medium	0.86	0.69	0.76	252
accuracy			0.85	750
macro avg	0.85	0.85	0.85	750
weighted avg	0.85	0.85	0.85	750

In the training set, the Naive Bayes model had 84.93% accuracy with macro average precision, recall, f1-score of 85%.

**Fig. 3.** Confusion matrix for Naive Bayes on training set
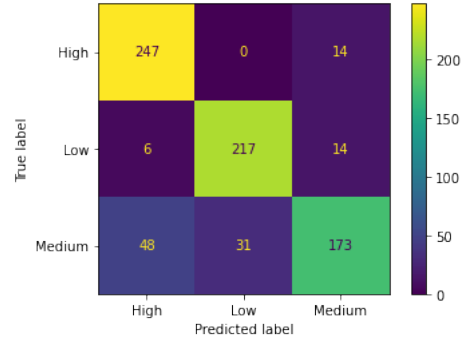
The confusion matrix in the training set showed that the majority of the results laid on the hypotenuse which showed that it was fairly accurate in its predictions. The highest error it had in its prediction was predicting 48 Medium Risk being labeled as High Risk.

Table 4. Naive Bayes Testing Set

	precision	recall	f1-score	support
High	0.86	0.94	0.90	104
Low	0.85	0.85	0.85	66
Medium	0.83	0.72	0.77	80
accuracy			0.85	250
macro avg	0.85	0.84	0.84	250
weighted avg	0.85	0.85	0.85	250

In the testing set, the model had an accuracy of 84.8%, with macro average precision being 85%, and both recall and f1-score being 84%.

Fig. 4. Confusion matrix for Naive Bayes on testing set
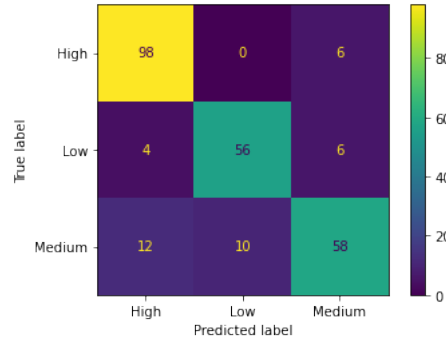
The confusion matrix generated showed that it had correctly identified the majority of each class. However, the model produced some misidentifications. All classes where the highest mistake was 12 Medium risk were incorrectly labeled as High by the model.

These results show that the Naive Bayes model is fairly accurate, and has the capacity to be a good model for predicting the likelihood with which a patient may contract lung cancer.

3.3 Data Visualization

As seen in the figures presented here, Logistic Regression gained a higher score on all metrics of evaluation on both train and testing sets. While Naive Bayes has performed well on its metrics, the results showed that Logistic Regression is the better algorithm in identifying an individual's likelihood of developing lung cancer.

Fig. 5. Comparison Between Logistic Regression and Naive Bayes in the Training Set
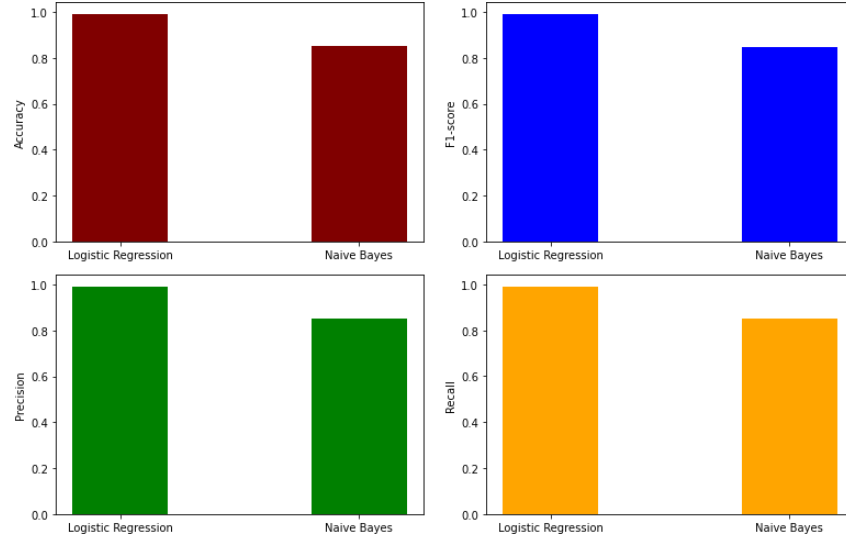
Fig. 6. Comparison Between Logistic Regression and Naive Bayes in the Testing Set
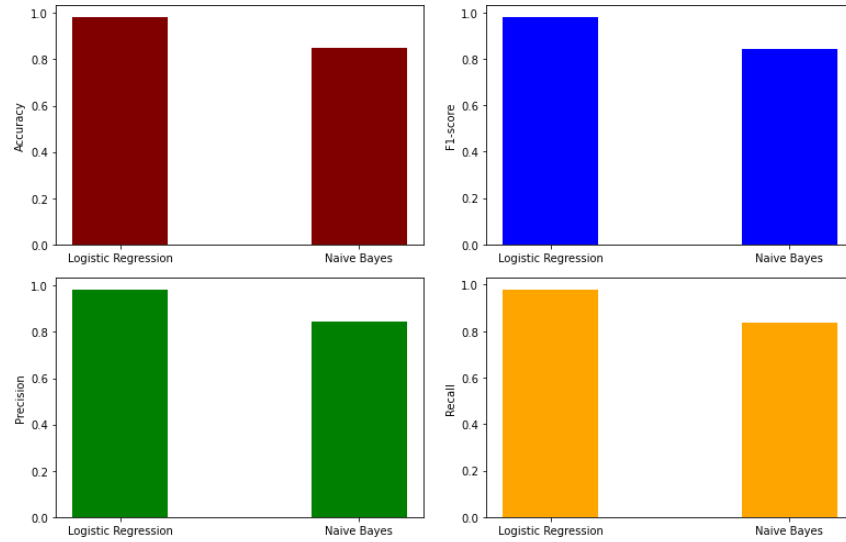
### 3.4   Logistic Regression Feature Analysis

Since Logistic Regression was seen as the better model to predict the likelihood of an individual to develop lung cancer, the weights for each feature for each class were analyzed and the resulting figures were made.

Fig. 7. Correlation of Factors and High Risk
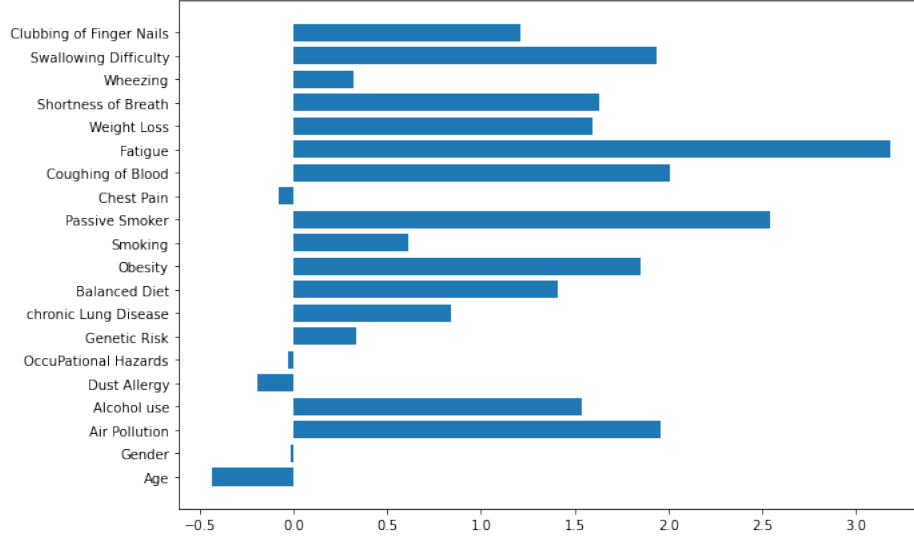
In high risk patients (Figure 7), it can be observed that the top 3 factors with the strongest correlation include **Fatigue** (3.179506), **Passive Smoking** (2.541472), and **Exposure to Air Pollution** (1.960435). Each of these 3 features possess very strong positive correlations, with Fatigue presenting itself as the best possible predictor for a high risk of lung cancer. Findings also indicate that factors such as Chest pain, Occupational hazards, Dust allergies, and Gender possess weak negative correlations with having a high risk of lung cancer, and Age possesses a medium negative correlation with having a high risk of lung cancer.

In low risk patients (Figure 8), it can be observed that the top 3 factors with the highest correlation include **Fatigue** (-3.414484), **Clubbing of Finger Nails** (-2.858945), and **Wheezing** (-2.696070), each possessing strong negative correlations. The findings also show that factors such as Gender and Chest Pain possess weak positive correlations with having a low risk of lung cancer.

In medium risk patients (Figure 9), it can be observed that the top 3 factors with the highest correlation include **Fatigue** (0.234977), **Passive Smoking** (0.825177), and **Coughing of Blood** (0.498174), with each of these factors possessing strong positive correlations with a medium risk of lung cancer. It is also worth noting that factors such as Age, Dust Allergies, and Chest Pain possess weak negative correlations with a medium risk of lung cancer. Based on what has been displayed by the findings, Fatigue has consistently displayed itself as a suitable predictor when determining a patient's risk of contracting lung cancer, displaying strong positive correlations for high and medium risk patients, and strong negative correlations for patients with low risk.

Fig. 8. Correlation of Factors and Low Risk
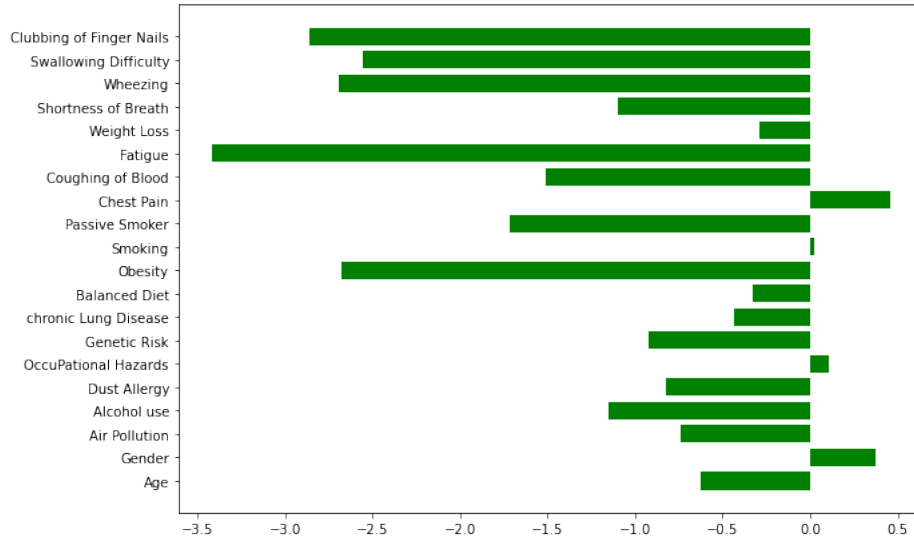
Fig. 9. Correlation of Factors and Medium Risk
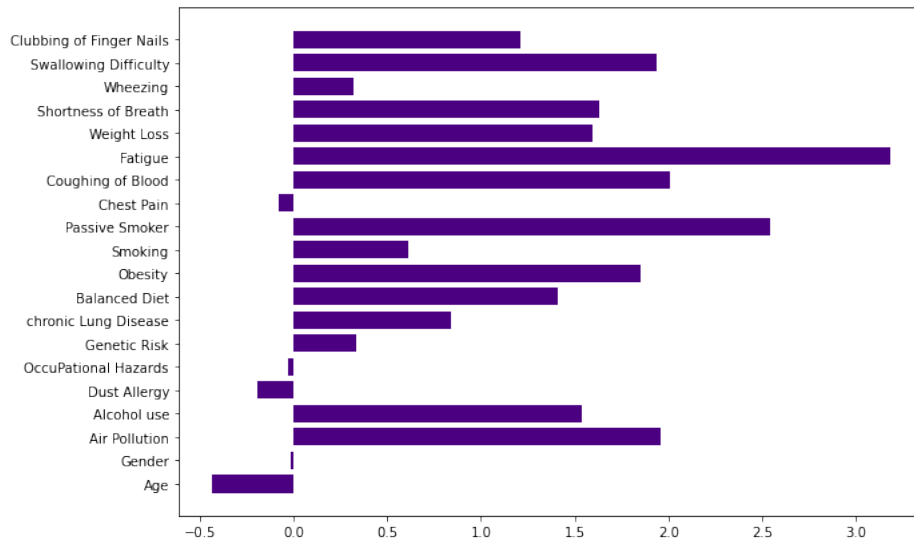
4   Conclusion

This study was made due to the desire of the researchers to implement and compare Machine learning models that can identify an individual's likelihood of

developing lung cancer. Specifically, they sought to: train a Logistic Regression model to predict likelihood of lung cancer; train a Naive Bayes model to predict likelihood of lung cancer; compare Naive Bayes and Logistic Regression's evaluation scores; determine the top 3 factors for people with high likelihood of contracting lung cancer; determine the top 3 factors for people with medium likelihood of contracting lung cancer; and determine the top 3 factors for people with low likelihood of contracting lung cancer. The researchers used a modified dataset of information on patients with lung cancer for the training and testing of the machine learning models.

At the end of the study the researchers were able to implement a Machine learning model that can identify an individual's likelihood of developing lung cancer. Specifically, they were able to train both a Logistic Regression model and Naive Bayes model with a 98.4% and 84.8% accuracy score respectively in the testing set which indicates that both sets are fairly reliable in predicting an individual's likelihood of developing lung cancer.

Logistic Regression proved to be the better model in predicting the level of lung cancer in both training and testing sets. While this high accuracy of 98.4% in the testing set may initially seem unrealistic, the precision, recall, and f1-scores also exhibited high results (98% macro average for all mentioned measures in the testing set), showing it was capable of identifying the classes correctly which outperformed Naive Bayes in all metrics (accuracy of 84.8% and 85% macro average of precision, recall, and f1-score in the testing set). The imbalance is also ruled out by the fact that the dataset had a close amount of each class. These results are supported by how Logistic Regression sees much practical use as a prediction algorithm in machine learning, and is most ideal when applied to these types of problems [1]. Therefore, it can be safely concluded that Logistic Regression models are more effective at predicting a person's likelihood of developing lung cancer compared to Naive Bayes models.

Furthermore, the feature analysis showed that the top 3 factors that can be used to predict whether a patient has a high risk of contracting lung cancer are fatigue (3.179506), passive smoking (2.541472), and exposure to air pollution (1.960435). Meanwhile, the top 3 factors that can be utilized to predict whether a patient has a low risk of contracting lung cancer are fatigue (-3.414484), clubbing of finger nails (-2.858945), and wheezing (-2.696070). To predict whether a patient has a medium risk of contracting lung cancer, the top 3 factors that can be used are fatigue (0.234977), passive smoking (-0.825177), and coughing of blood (-0.498174). Based on the findings, Fatigue has consistently displayed itself as a suitable predictor when determining a patient's risk of contracting lung cancer, displaying strong positive correlations for high and medium risk patients, and strong negative correlations for patients with low risk

Given similar data found in the dataset, the model built in the study can accurately predict an individual's likelihood of developing lung cancer Even without the model to make predictions, the study provided information about an individual experiencing a level of fatigue, exposure to air pollution, and passively smoke has an increased likelihood in developing lung cancer.

4.1   Recommendations

Based on the findings and limitations of the study, the following recommendations for improvement are made:

1. Test other machine learning models such as SVM and Decision Trees to identify an individual's likelihood in developing lung cancer for more comparison.
2. Use the excluded columns of the dataset as features in training and testing the model.
3. Use other similar but larger datasets with more features but not too much in training and testing the models.
4. Delve into studying the features that showed strong positive correlation with likelihood of developing lung cancer.

Bibliography

[1] Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179.

[2] Cancer Research UK (2019). Lung cancer.

[3] Department of Health (2021). Doh leads national lung cancer awareness month, free medicines available nationwide.

[4] Gupta, A., Zuha, Z., Ahmad, I., and Ansari, Z. (2022). A study on prediction of lung cancer using machine learning algorithms.

[5] IQAir (2022). Manila air quality index (aqi) and philippines air pollution.

[6] MacroTrends LLC (2022). Philippines smoking rate 2000-2022.

[7] Philippine Statistics Authority (2021). Highlights of the population density of the philippines 2020 census of population and housing (2020 cph).

[8] Radhika, P., Nair, R. A., and Veena, G. (2019). A comparative study of lung cancer detection using machine learning algorithms. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–4. IEEE.