Lung Cancer Prediction

6. Analysis Report

Jishen Harilal

5/29/25

Logistic Regression

Table of Contents

Introduction

This report looks at using a classification model to predict lung cancer based on a person's demographics, lifestyle, and symptoms. The *More Accurate Lung Cancer Dataset* from Kaggle was chosen for this analysis.

The dataset is small, clean, and has meaningful features, making it a good fit for a proof-of-concept classification task.

The target variable, *LUNG_CANCER*, is binary, making it suitable for supervised learning methods like logistic regression. Most features are binary variables, with one numerical feature (*AGE*), all linked to lung cancer risk. This report covers the data exploration, model training, evaluation, and key findings, with suggestions for how the results can be used.

Dataset Overview

The dataset (lcs.csv) contains 1157 records and 16 features:

Variables

Numerical Independent Variable:

- *AGE*

Dichotomous Independent Variables:

- *GENDER*
- *SMOKING*
- *YELLOW_FINGERS*
- *ANXIETY*
- *PEER_PRESSURE*
- *CHRONIC_DISEASE*
- *FATIGUE*
- *ALLERGY*
- *WHEEZING*
- *ALCOHOL_CONSUMING*
- *COUGHING*
- *SHORTNESS_OF_BREATH*
- *SWALLOWING_DIFFICULTY*
- *CHEST_PAIN*

Dichotomous Target Variable:

- *LUNG_CANCER*

From a domain perspective, the independent variables make logical sense as predictors for lung cancer. For example:

- Smoking is a well-known for increasing the risk of lung cancer.
- Most of the features included in the dataset are common symptoms experienced by individuals with respiratory issues, closely linked to lung health, and thus could serve as indicators for lung cancer.
- Demographic factors such as age and gender could reveal broader trends in lung cancer contraction.

Initial Findings:

- No missing values were found.
- Some fields need to be properly encoded using binary encoding.

Exploratory Data Analysis (EDA)

Statistical Analysis

	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE
count	1157.000000	1157.000000	1157.000000	1157.000000	1157.000000	1157.000000	1157.000000
mean	50.750216	1.317200	1.339672	1.504754	1.332757	1.322385	1.458946
std	17.183339	0.465587	0.473803	0.500194	0.471404	0.467592	0.498527
min	20.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
25%	35.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
50%	54.000000	1.000000	1.000000	2.000000	1.000000	1.000000	1.000000
75%	64.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
max	87.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000

ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN
1157.000000	1157.000000	1157.000000	1157.000000	1157.000000	1157.000000	1157.000000
1.535869	1.336214	1.597234	1.509939	1.337943	1.292135	1.315471
0.498927	0.472618	0.490666	0.500117	0.473214	0.454941	0.464904
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
2.000000	1.000000	2.000000	2.000000	1.000000	1.000000	1.000000
2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000

Figure 1: Statistical summary

The summary in Figure 1, shows that the dataset looks clean and complete and that every column contains the same number of entries and that there are no missing values.

- Age ranges from 20-87, showing that there are no unrealistic outliers or errors in the age data. The average age is around 51 years.
- The rest of the variables are categorical and will be further analysed.

Overall, the data appeared reliable and well-suited for modelling, no signs of irregular or missing values.

Data Preparation

The dataset underwent several cleaning steps to prepare it for analysis:

- **Binary Encoding:**
  All binary variables were properly encoded for model use.
  - *GENDER* was originally recorded as 'M' or 'F' and has been converted to '1' for male and '0' for female.
  - *LUNG_CANCER* (the target variable) was recorded as 'YES' or 'NO' and has been encoded to '1' for yes and '0' for no.
  - All other categorical (binary) variables, originally recorded as '1' or '2', were standardised to '0' (no) and '1' (yes) to align with binary expectations in most machine learning libraries.
- **Deduplication:**
  The dataset initially contained 1,157 records, but a significant number of duplicates (246) were identified and removed. After deduplication, 911 unique records remained for analysis.

These steps ensured the data was consistently formatted, free of duplicate entries, and ready for further exploration and model development.

Distribution Analysis

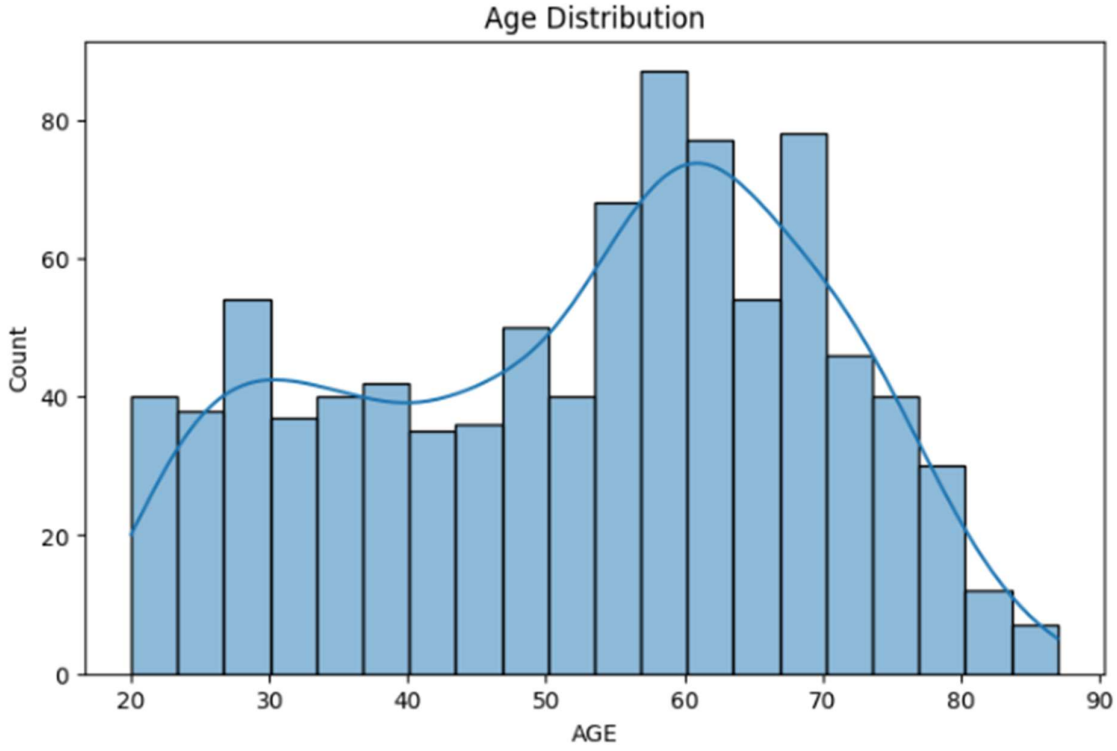

Figure 2: Age Distribution.

Figure 2 illustrates the age distribution in the dataset. It shows that individuals around the ages of 60 and 70 seem to be represented nearly twice as often as most younger age groups, with a sharp falloff in population after the age of 70, which is a realistic expectation when considering the average life span.

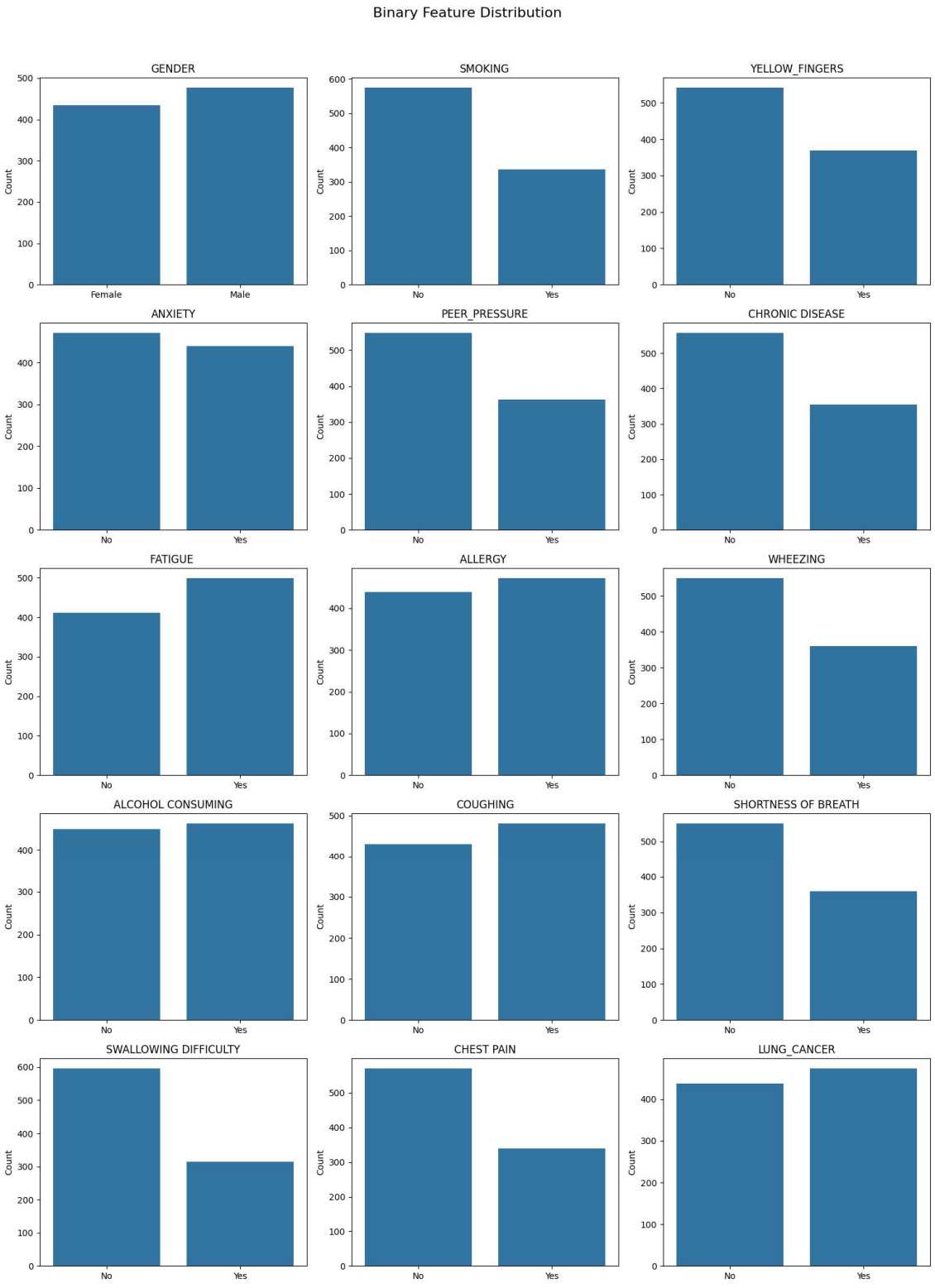

Figure 3: Binary Feature Distribution

Figure 3 shows the distribution/ balance of the binary features present in the dataset.

It is important that all categories maintain some level of balance so that no group is over or under-represented causing a biased outcome.

- We can see that most features contain a balanced number of individuals in both categories, with others only slightly imbalanced, a realistic expectation given the number of symptoms being reported.

- Symptoms that relate to smoking show a similar balance, which makes perfect sense, since those that smoke are more likely to experience those symptoms, over those who do not. This is also potentially an early indicator of multicollinearity present in the dataset, having to do with certain symptoms being related, for example, it is logical to assume that people who have a wheezing condition experience shortness of breath and chest pain.

Target Class Balance

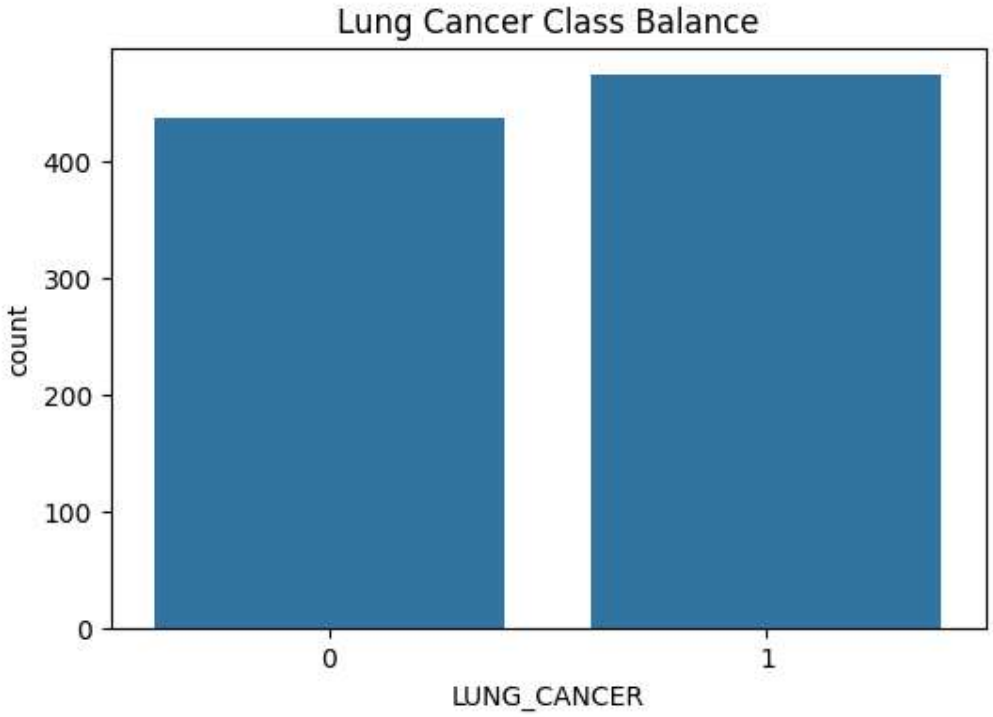

Figure 4: LUNG_CANCER (Target) Class Balance

Figure 4 shows the balance of the classes for the *LUNG_CANCER* variable – the target variable.

- The two classes seem to be almost evenly balanced, which is a good sign, and means that class imbalance will not play a role in affecting model evaluation. If the classes were highly imbalanced, say 80-20, then the model could get away with 80% accuracy by repeatedly predicting just one of the classes, making the metric less meaningful.

Feature Selection

Feature Correlation Analysis

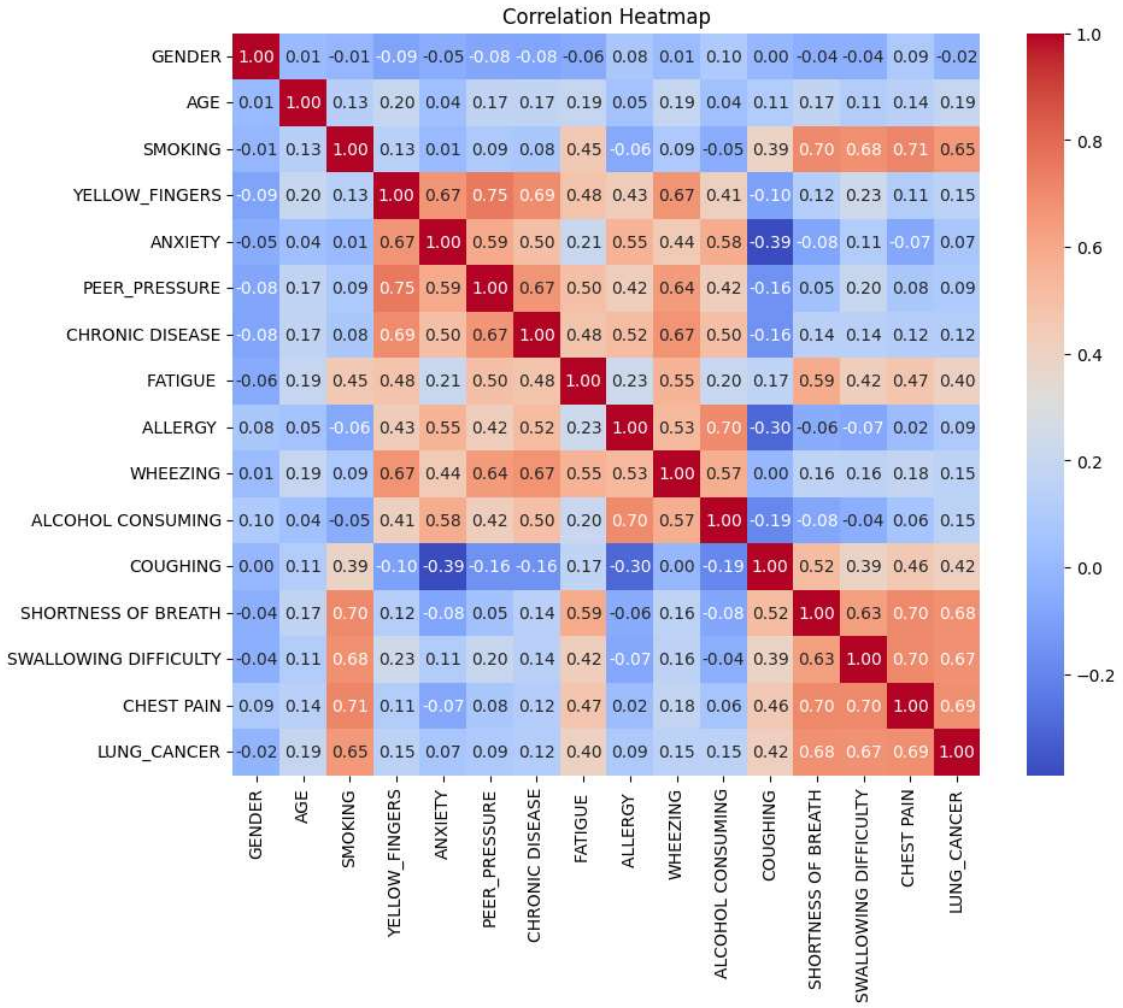

Figure 5: Correlation Heatmap

The heatmap presented in Figure 5 highlights key patterns in how features relate to lung cancer.

- **Strong correlations** were identified between lung cancer and clinical symptoms such as chest pain, shortness of breath, swallowing difficulty, and smoking. These variables are likely to serve as primary predictors and are expected to contribute significantly to the model's performance.
- **Moderate correlations** were observed with features like coughing and fatigue. While not as central as the top predictors, they still offer valuable additional predictive power.

- **Weak or minimal correlations** were found for age, wheezing, yellow fingers, alcohol consumption, chronic disease, peer pressure, allergy, and anxiety. These variables may have limited direct impact but can still provide some background context.
- **No meaningful correlation** was detected for gender, indicating it does not play a notable role in differentiating lung cancer cases in this dataset.

Based on these insights, the feature selection process will focus on variables with stronger predictive relationships to improve model performance and accuracy.

Backward Elimination Process

```
Removed 'PEER_PRESSURE' with p-value 0.8255
Removed 'YELLOW_FINGERS' with p-value 0.4179
Removed 'CHRONIC DISEASE' with p-value 0.5090
Removed 'FATIGUE ' with p-value 0.2774
Removed 'AGE' with p-value 0.1145
Removed 'GENDER' with p-value 0.0766
Final selected features: ['const', 'SMOKING', 'ANXIETY', 'ALLERGY ', 'WHEEZING', 'ALCOHOL CONSUMING',
'COUGHING', 'SHORTNESS OF BREATH', 'SWALLOWING DIFFICULTY', 'CHEST PAIN']
                           Logit Regression Results
==============================================================================
Dep. Variable:            LUNG_CANCER   No. Observations:                  911
Model:                          Logit   Df Residuals:                      901
Method:                           MLE   Df Model:                            9
Date:                Thu, 29 May 2025   Pseudo R-squ.:                  0.6834
Time:                        15:03:05   Log-Likelihood:                -199.71
converged:                       True   LL-Null:                       -630.71
Covariance Type:            nonrobust   LLR p-value:                9.566e-180
============================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                     -5.0958      0.524     -9.722      0.000      -6.123      -4.068
SMOKING                    1.8837      0.487      3.870      0.000       0.930       2.838
ANXIETY                    0.9622      0.464      2.074      0.038       0.053       1.872
ALLERGY                    1.1531      0.423      2.728      0.006       0.325       1.981
WHEEZING                  -1.8921      0.314     -6.025      0.000      -2.508      -1.277
ALCOHOL CONSUMING          2.7210      0.412      6.598      0.000       1.913       3.529
COUGHING                   2.2651      0.443      5.113      0.000       1.397       3.133
SHORTNESS OF BREATH        3.0985      0.486      6.381      0.000       2.147       4.050
SWALLOWING DIFFICULTY      4.0354      0.596      6.768      0.000       2.867       5.204
CHEST PAIN                 2.0577      0.508      4.053      0.000       1.063       3.053
============================================================================================
```

Figure 6: Logit Regression and Backward Elimination Results

Backward Elimination using p-values was used to iteratively remove insignificant features. The results of this process are shown in Figure 6.

- Logit regression was used to iteratively identify features with p-values > 0.05 and remove them, resulting in the removal of features: *PEER_PRESSURE, YELLOW_FINGERS, CHRONIC DISEASE, FATIGUE, AGE, GENDER*

- Final Logit regression results show no features with p-values > 0.05

Model Training Process

Model Training Summary

A logistic regression model was developed to predict lung cancer using nine key features: smoking status, anxiety, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain. These variables were selected based on their statistical significance identified during the backward elimination feature selection process.

The dataset was divided into a training set (70%) and a test set (30%) to allow for fair and reliable model evaluation. The logistic regression was configured with an increased iteration limit (max_iter = 1000) to ensure the model properly converged during training.

This setup offers a clear and interpretable method for classifying lung cancer presence, focusing on the most meaningful predictors to achieve robust and explainable results.

Model Evaluation

Evaluation Metrics Results

Metric	*Result*
Accuracy	0.9343
Precision	0.9771
Recall	0.8951
F1-score	0.9343
ROC AUC	0.9850

Several key metrics were used to evaluate the performance of the logistic regression model.

- Accuracy (93.4%): This indicates that the model correctly predicts lung cancer presence or absence in most cases overall. It reflects the proportion of total correct predictions.
- Precision (97.7%): The model is highly precise, meaning when it predicts a positive lung cancer case, it is almost always correct.
- Recall (89.5%): The model successfully captures most of the true positive cases, ensuring that most of the actual lung cancer cases are correctly identified.
- F1-Score (93.4%): This 'harmonic' mean of precision and recall shows the model balances both well, offering strong overall predictive reliability.
- ROC AUC (98.5%): Will be further discussed in the next section.

These metrics suggest the model is highly effective for lung cancer classification, with both strong precision and sensitivity, making it suitable for applications where both accuracy and patient safety matter.

ROC Curve Analysis

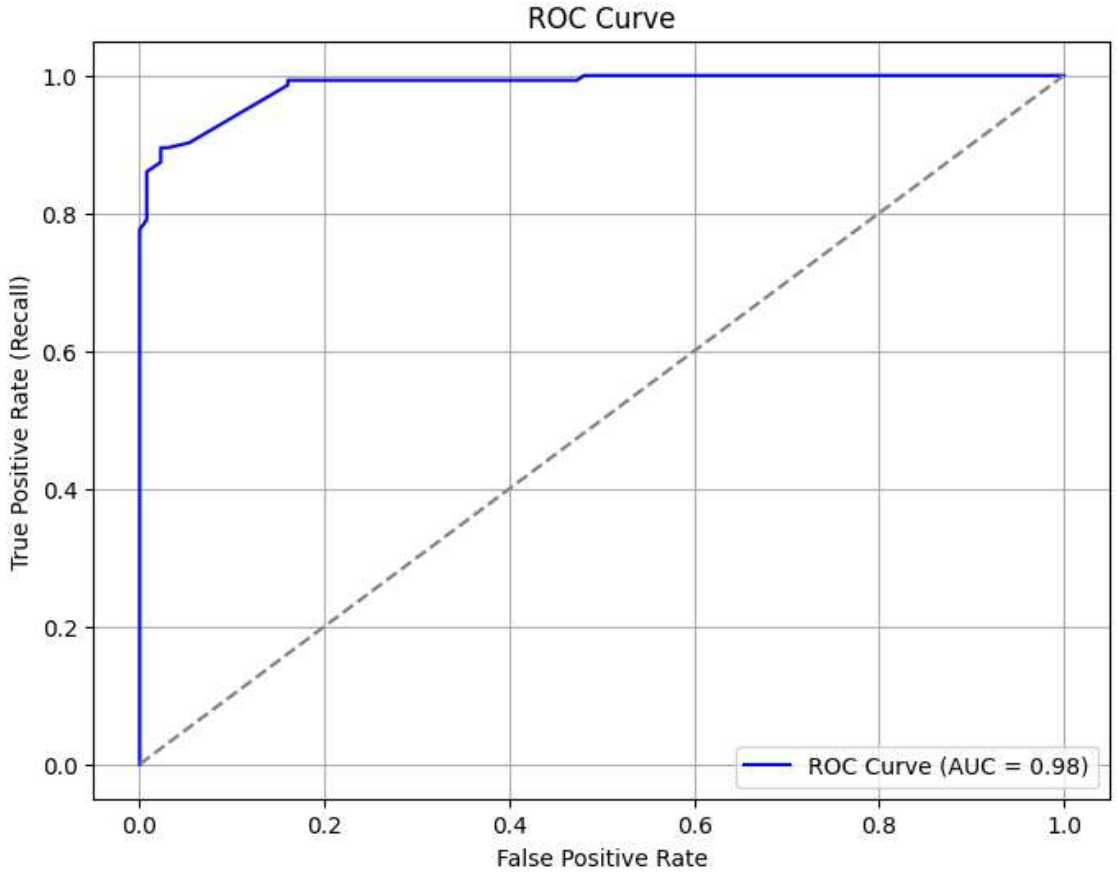

Figure 7: ROC Curve plot

The ROC curve for the logistic regression model shows excellent performance, with an AUC score of 0.98. This indicates the model has a very strong ability to distinguish between patients with lung cancer and those without. A score close to 1.0 suggests that across all decision thresholds, the model consistently ranks positive cases higher than negative ones.

In practical terms, this means the model is highly effective at separating high-risk patients from low-risk ones, supporting reliable classification even when adjusting the sensitivity or specificity balance.

Confusion Matrix Analysis

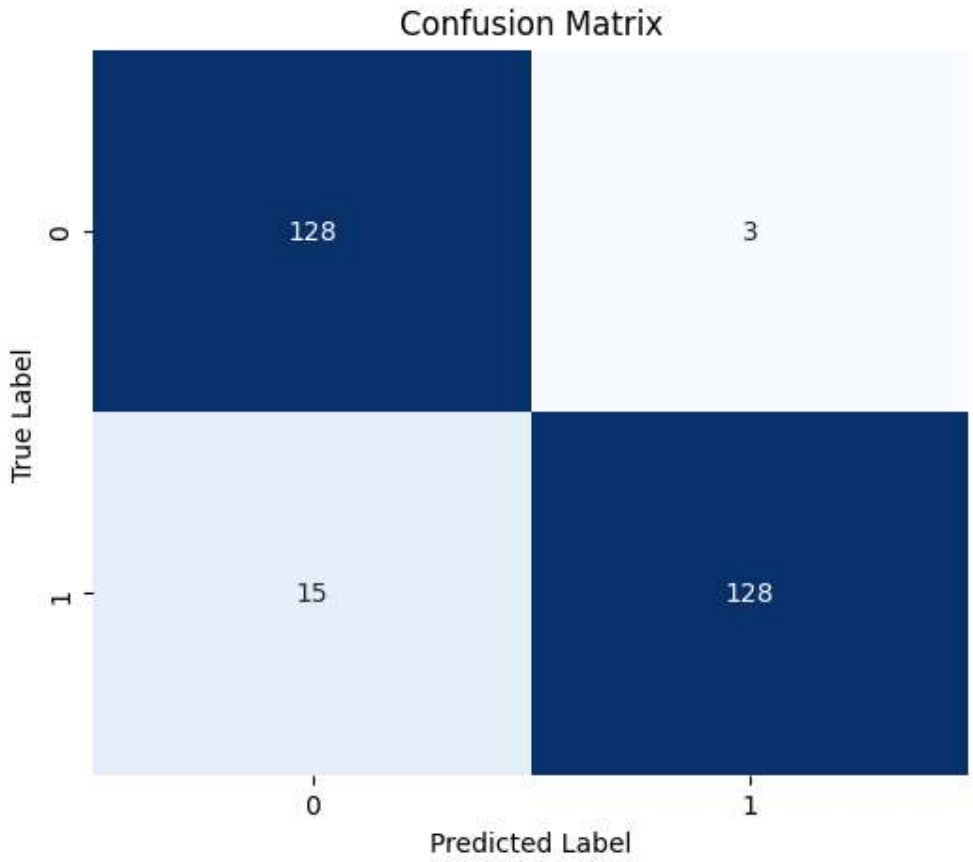

Figure 8: Confusion Matrix

The confusion matrix shown in Figure 8 shows that the model is very good at identifying non-cancer cases (very few false alarms).

It's also strong at detecting cancer, though it misses some cases (15 false negatives), which is important in a medical context because missing positive cases can have serious consequences.

Overall, the confusion matrix reflects a well-performing classifier with high accuracy, precision, and recall.

Consideration for Model Retraining

Based on the current evaluation metrics, retraining the logistic regression model is not considered necessary at this stage. The model achieves high performance across key indicators, including an accuracy of over 93%, precision near 98%, recall close to 90%, an F1-score matching the overall accuracy, and an excellent ROC AUC score of 0.98.

The confusion matrix further supports this, showing minimal false positives (only 3 cases) and a strong ability to correctly identify both positive (cancer) and negative (non-cancer) cases. While there are some false negatives (15 cases), the overall balance between precision and recall indicates that the model maintains reliable sensitivity and specificity.

Given these strong outcomes, the current model generalises well to unseen data, and no immediate performance concerns justify retraining. Retraining may only become necessary in the future if additional data is collected, the underlying data patterns change, or the business context shifts, requiring the model to adapt to new conditions.

Recommendations

The logistic regression model offers several opportunities to improve decision-making around cancer benefit applications:

- **Risk-Based Benefit Prioritisation:** Strong predictors like smoking status, chest pain, and shortness of breath can help medical schemes flag higher-risk applicants for faster processing or enhanced review, ensuring benefits are directed where most needed.
- **Preventive Health Programs:** The model's identification of key lifestyle and symptom factors (e.g., smoking, alcohol consumption) points to opportunities for targeted health campaigns or smoking cessation initiatives to reduce future cancer claims.
- **Decision Support Tools:** Given the model's simplicity and interpretability, it can be integrated into internal decision tools to assist case managers in making consistent, data-driven evaluations when reviewing benefit applications.
- **Resource Allocation:** By focusing on the most predictive variables, schemes can better allocate clinical assessment resources, focusing attention on applicants with higher symptom burdens or risk profiles.
- **Future Data Strategy:** Insights from the model can guide which additional data points (e.g., medical history, environmental exposures) to collect in future to further improve prediction accuracy and benefit design.

Limitations and Future Work

While the logistic regression model shows strong performance, several limitations should be noted:

- **Dataset Size and Scope:** The dataset is relatively small (~900 records after cleaning) and sourced from a specific context, which may limit the model's generalisability to broader or more diverse populations.
- **Feature Range:** The available features focus mostly on self-reported symptoms and basic demographics, without including deeper clinical or genetic information that could improve prediction power.
- **Potential Biases:** The dataset may carry inherent biases, such as underreporting or unbalanced representation across groups, which could affect the model's fairness and accuracy.
- **Binary Classification Constraint:** The model currently predicts only presence or absence of lung cancer. Future work could explore more nuanced outputs, such as risk probabilities or severity levels, to better support clinical decision-making.
- **Limited Model Comparison:** While logistic regression offers simplicity and interpretability, other machine learning approaches (e.g., random forests, gradient boosting) could be explored in future work to potentially enhance performance.

Future work should focus on expanding the dataset with more diverse and detailed records, exploring additional predictive variables, testing alternative modelling techniques, and conducting external validation to ensure the model's robustness across different patient populations.

End