Machine learning for clinical decision support

Date: 2th of February 2026

Author: Ali Mekni

Role: data analyst

Leveraging Machine Learning Approaches for Lung Disease Diagnosis

Business Assessment and Recommendations Report

Type: Essay

CONTENTS

TABLES

FIGURES

INTRODUCTION

Lung diseases are one of the causes of global mortality, while early detection may improve patient outcomes, traditional diagnostic methods face challenges like limited resources and delayed interpretation. Machine learning models can make the analysis faster and more accurate by detecting relations between features and the target variable.

The goal of this project is to examine 12,000 patient results using classification algorithms to predict lung diseases. The database contains 33 features split into three datatypes 17 numerical features,5 categorical features and 8 binary features.

The report is divided into five chapters: first exploratory data analysis, which looks at distributions and relationships next data preparation, to address processing steps after that model training, which covers algorithm implementation; then model evaluation, which compares each algorithm's performance; and finaly a conclusion, where i summarize results and offers recommendations for the future.

Chapter 1: Exploratory Data Analysis

1. Introduction:

In the first chapter, we conduct a comprehensive exploratory data analysis to understand the foundational structure and patterns within the dataset. The primary objectives are to characterize the target variable, summarize key numerical attributes, and visually investigate the data's distribution and relationships.

2. Target variable analysis:

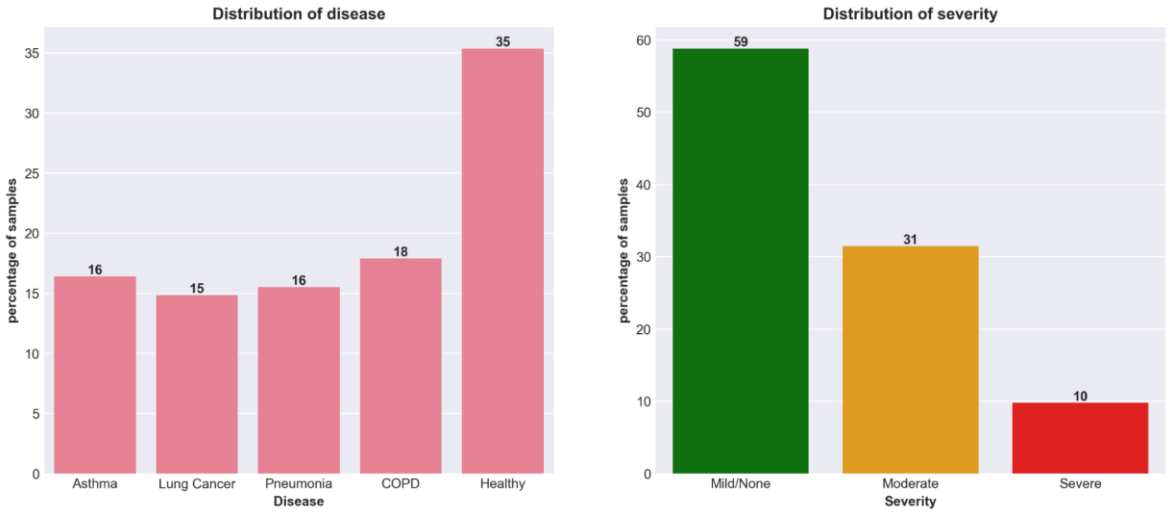

Figure 1 : target variable

The figure shows two potential target variables. The severity feature presents an ordinal distribution while disease is a nominal feature both features are unbalanced datasets but the disease distribution is relatively moderate compared to the severity distribution which is more skewed toward Mild/None class (59%). therefore, we select the disease type as a target variable for this analysis.

3. Statistical summary

	count	mean	median	std	min	25%	50%	75%	max	skewness
age	12000.0	58.07	58.00	14.74	18.00	48.00	58.00	68.00	90.00	-0.07
bmi	12000.0	27.01	27.00	4.96	16.00	23.50	27.00	30.40	45.60	0.08
pack_years	12000.0	14.77	10.30	14.85	0.00	2.70	10.30	22.80	81.80	1.16
pm25_exposure	12000.0	15.11	15.00	6.69	3.00	10.30	15.00	19.70	41.30	0.22
weight_loss_kg	12000.0	1.62	1.40	2.42	-5.00	0.00	1.40	3.00	13.30	0.50
spo2	12000.0	95.96	96.00	2.38	87.20	94.40	96.00	97.70	100.00	-0.27
respiratory_rate	12000.0	17.97	18.00	2.99	10.00	15.90	18.00	20.00	28.70	-0.00
crp_mg_L	12000.0	20.79	10.30	30.61	0.00	1.50	10.30	22.90	218.60	2.40
wbc_10e9_L	12000.0	8.10	7.80	2.61	2.00	6.30	7.80	9.50	20.70	0.69
fev1_fvc	12000.0	0.74	0.75	0.10	0.32	0.68	0.75	0.81	0.95	-0.40
fev1_pct_pred	12000.0	83.81	85.10	18.77	20.00	72.20	85.10	96.50	140.00	-0.33
fvc_pct_pred	12000.0	94.92	94.90	12.05	49.60	86.80	94.90	103.10	140.00	-0.01
dlco_pct_pred	12000.0	81.51	82.40	17.63	20.00	70.80	82.40	93.50	139.50	-0.36
ct_nodule_size_mm	12000.0	4.64	3.00	5.93	0.00	1.00	3.00	5.40	47.80	2.51
ct_emphysema_pct	12000.0	8.28	6.30	8.29	0.00	2.50	6.30	10.90	51.60	1.75
sixmwd_m	12000.0	403.81	405.00	109.52	-29.00	329.00	405.00	478.00	816.00	-0.06
hospital_visits_last_year	12000.0	0.62	0.00	0.74	0.00	0.00	0.00	1.00	5.00	1.13

Figure 2: statistical description

the table shows the various statistical metrics for numerical features generally if the median is less than the mean this indicates a positively skewed distribution in contrast if the median is approximately equal to mean this indicate a symmetric distribution, to enhance precision and interpretability a skewness column has been added to the table.

4. Histogram and box plot visualizations

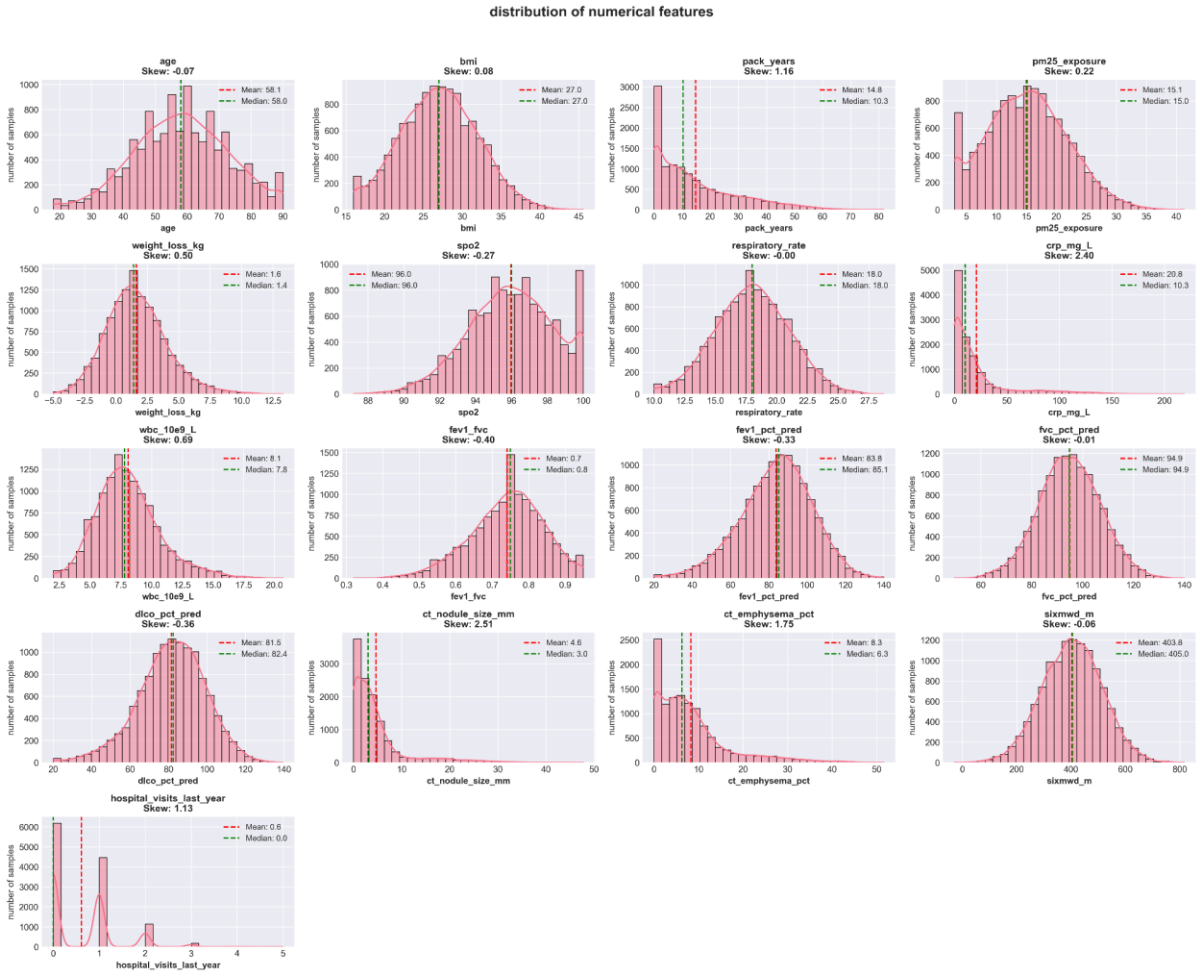

Figure 3: numerical features histograms

The histogram reveals the distributions of the numerical features. While many variables such as age and BMI approximate a normal distribution, others are right-skewed that resemble exponential or power-law-like distributions. Notably, the variable hospital_visits_last_year displays a discrete right-skewed distribution with a high frequency of zero values, which may suggest a Poisson-like process; however, formal goodness-of-fit tests would be required to confirm this hypothesis. Our goal is not to identify the exact distribution, but rather to visually detect which are skewed and which are symmetric.

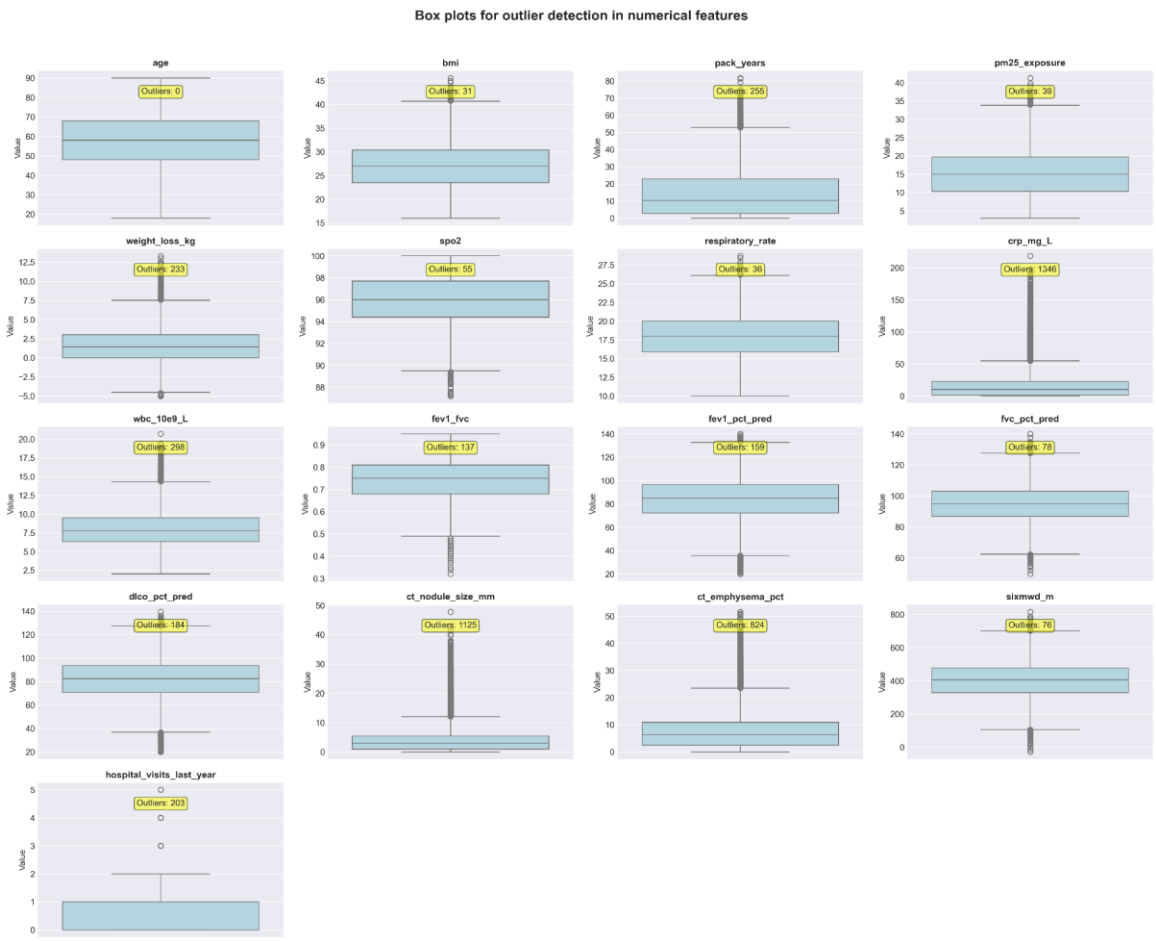

Figure 4: outliers visalizations

The box plot visualization reveals that all numerical features show outliers, with the exception of age. However, a critical question arises: are these values truly “outliers” in the statistical sense? the answer is yes, but in clinical context unhealthy samples are often show extreme values like high CRP or low SpO_2 that lie outside the typical range observed in healthy individuals. in the next chapter (data preparation) we will introduce techniques to deal with it.

5. Correlation results with target variable:

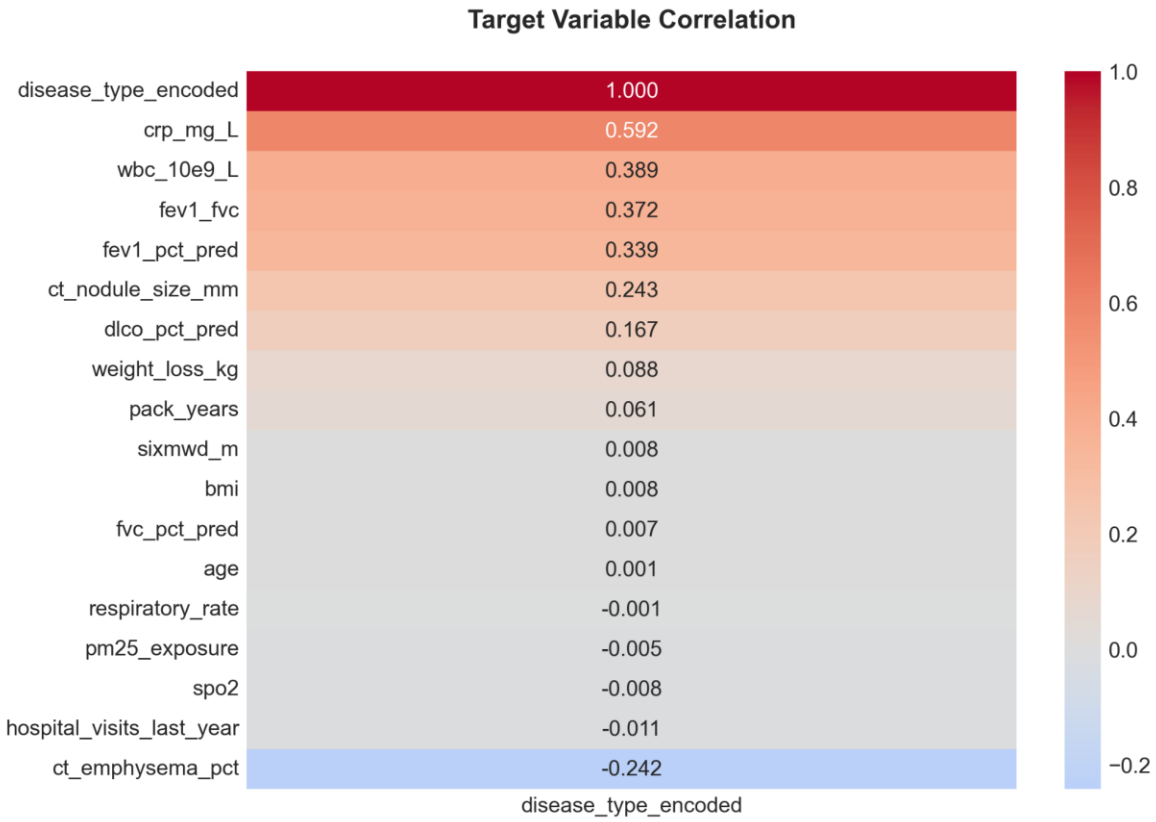

Figure 5: target varibale correlation

The figure shows the correlation results between numerical features and encoded target variable, cpr level is the strongest positive linear relation with 0. 592. This analysis is important, especially for models that assume linear relationships, such as logistic regression.

Chapter 2 Data Preparation

1. Introduction:

In this chapter, we describe the data preparation process which transforms raw data into a structured format suitable for training. Key steps are handling missing values, encoding categorical features and scaling numerical features.

2. Missing values inspection:

	family_history
Missing Values	7176.0
Percentage	59.8

Figure 6: missing values table

The table shows that approximately 60% of family history values are missing. Despite the high portion we can assume that patients are unaware of their family medical history, then it is better to keep the column and impute the values with a new category like 'unknown' rather than remove it entirely.

3. Encoding Methods:

Feature Name	Type	Encoding Method	Reasons
Disease Type	Target (Nominal)	Label Encoding	One dimension array No ordinal assumption
Severity	Input (Ordinal)	Manual Mapping Ordinal Encoding	One dimension array ordinal assumption
Categorical features	Input (Nominal)	One-Hot Encoding	two-dimension array No ordinal assumption

4. Feature Scaling:

Method	Numerical features	Reasons
RobustScaler	Symmetric features	Preserves the shape of the distribution Robost to outliers
PowerTransformer **(Yeo johnson)**	Skewed features	Transform data to Gaussian-like distribution Robust to outlier Handles both negative and positive values

Table 1: feature scaling

5. Before and after scaling visualizations:

a. Skewed features :

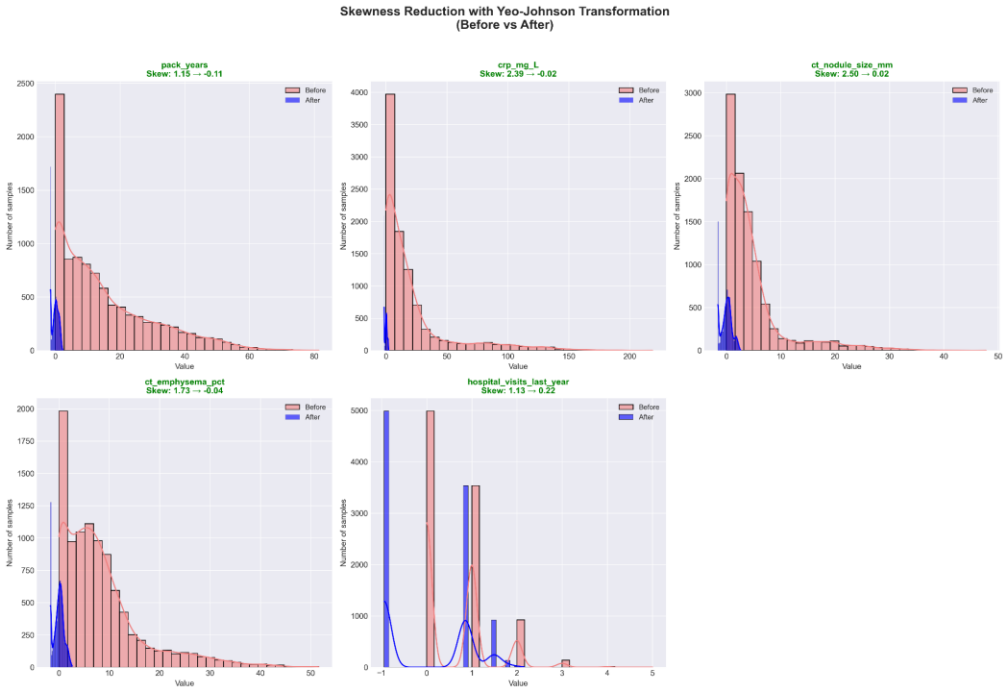

Figure 7: skewed reduction distributions before and after

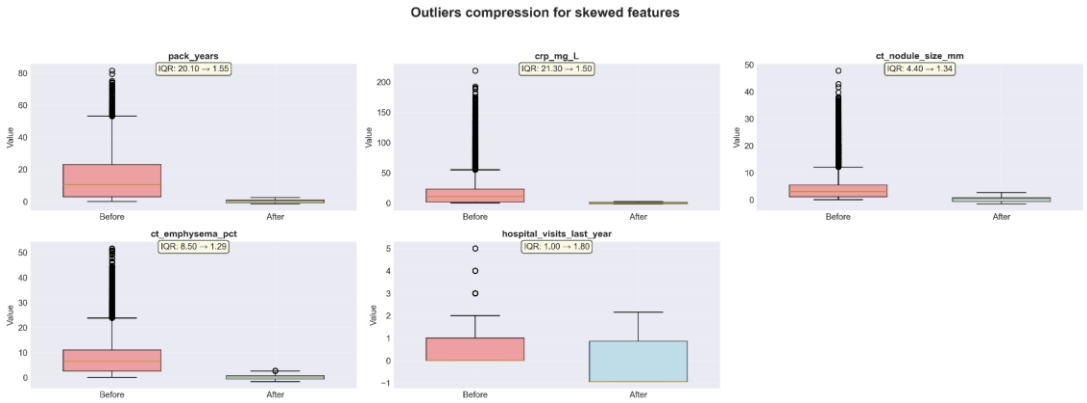

Figure 8: Outliers compression for skewed features

Powertranformer is a nonlinear transformation that changes the distribution to look more like a normal bell curve. A symmetric distribution has skewness values between -0.5 and 0.5 which is our case across all features additionally the transformation changes the spread of the data which results to a modified interquartile Range (IQR). A perfect normal distribution will be approximately 1.349 in our case all features display values close to this.

b. Symmetric features :

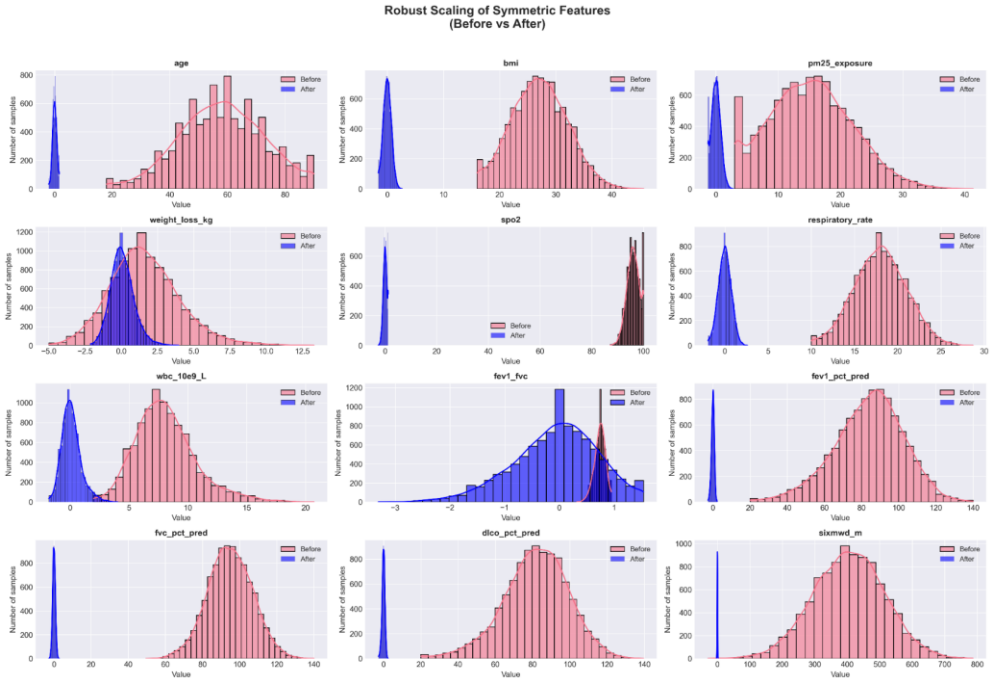

Figure 9: robost scaling for symmetric features before and after

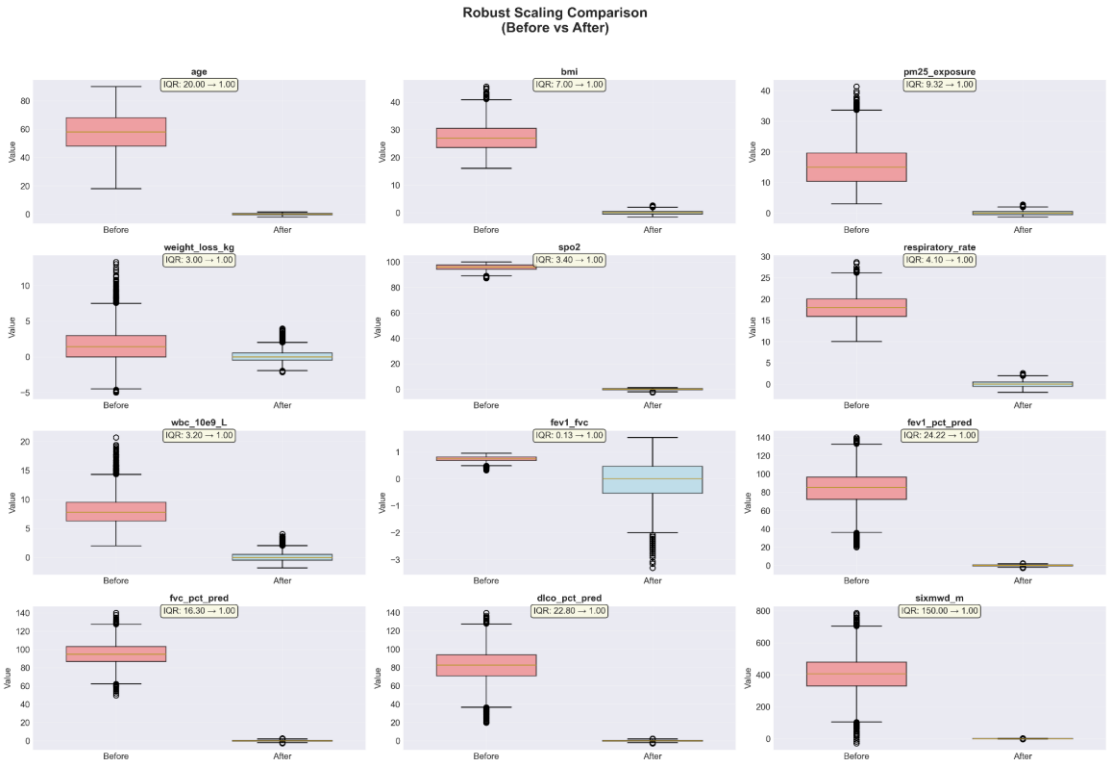

Figure 10:Outliers compression for symmeric features

Robostscaler is a linear transformation that preserve the distribution shape but it takes the original distribution and squeeze it so that the IQR is exactly 1. IQR will be reduced if the IQR was greater than 1 and increased if it was smaller than 1 which is our case applies across all the features

Chapter 3 Model Training

1. Introduction:

In this chapter, we discuss the model training process which is teaching an algorithm to find patterns. Key steps are handling missing values, encoding categorical features and scaling numerical features.

2. Train-Test Split Strategy:

The split of the data depends on the size of the dataset. If the data is relatively small, allocating 20% to 30% for testing is sufficient to build a reliable model. In contrast, if the data is relatively huge (for instance, around 1 million samples), even 1% is sufficient for the test set. (Muraina, 2022)

3. Classification Algorithms and relation to the datasets :

Algorithm	Sensitivity to distribution	Sensitivity to outliers	Default hyperparameters
Logistic regression	Yes	Yes	penalty: l2 C: 1.0 solver: lbfgs
Support vector machine	Yes	Yes	kernel: rbf C: 1.0 gamma: scale
K nearest neighbord	Yes	Yes	n_neighbors: 5 weights: uniform
Decision tree	No	No	max_depth: None min_samples_split: 2
Random forest	No	No	n_estimators: 100 max_depth: None

Table 2: classification algorithm and relation

Among the trained models, Logistic Regression, Support vector machine, and k nearest neighbors are known to be sensitive to distribution shape and outliers. To address this, all numerical features were preprocessed using distribution-aware scaling. In contrast, tree-based models (Decision Tree, Random Forest) are inherently robust to these issues. (Pinheiro et al., 2025)

Chapter 4 Model evaluation and visualization:

1. Introduction:

In this chapter we assess the machine learning model's performance to ensure it generalizes well to new data. This face includes models' evaluation, visual diagnostics and hyperparameter tuning for the best model.

2. Model evaluation:

a. Performance metrics:

	Model	Accuracy	Precision (macro)	Recall (macro)	F1-score (macro)	ROC AUC (macro)
3	Random Forest	0.938	0.948	0.933	0.940	0.992
1	SVM	0.938	0.946	0.931	0.938	0.991
0	Logistic Regression	0.935	0.942	0.930	0.935	0.992
2	Decision Tree	0.888	0.890	0.888	0.889	0.929
4	KNN	0.843	0.884	0.811	0.838	0.949

Figure 11: performance metrics

The evaluation of five classification algorithms using macro-averaged metrics show strong performance of Random Forest model which achieves the highest F1-score (0.940) and precision (0.948). Surprisingly logistic regression and support vector machine performs better than decision tree although K nearest neighbors remains the worst model even after data scaling therefore, we choose the random forest algorithm for fine tuning.

b. Confusion Matrix:

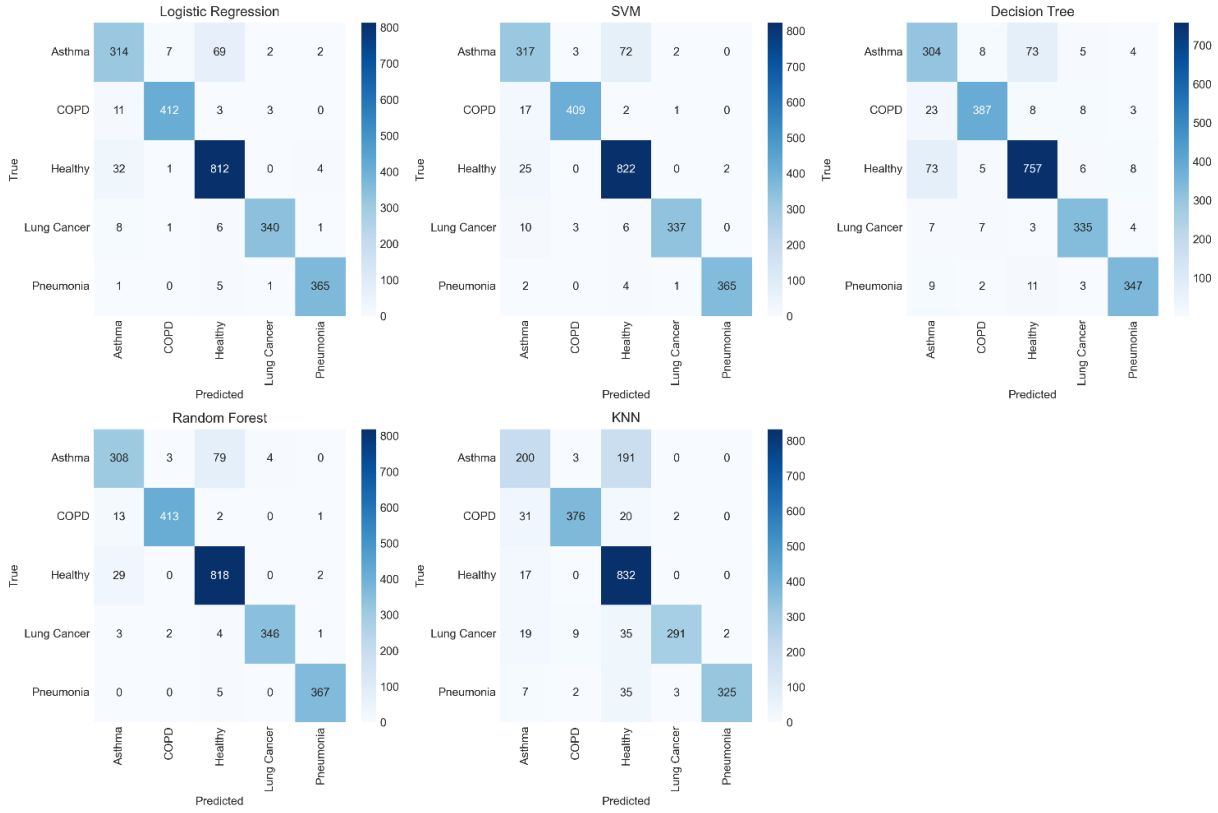

Figure 12: Confusion matrix

The confusion matrix is another performance metric that calculates misclassification counts for each class. For instance, in a K-Nearest Neighbors model applied to health analysis, while achieving 200 correct predictions, it misclassified 191 healthy samples. This represents poor performance and poses a dangerous outcome for healthcare applications

c. Multicalss Roc curves:

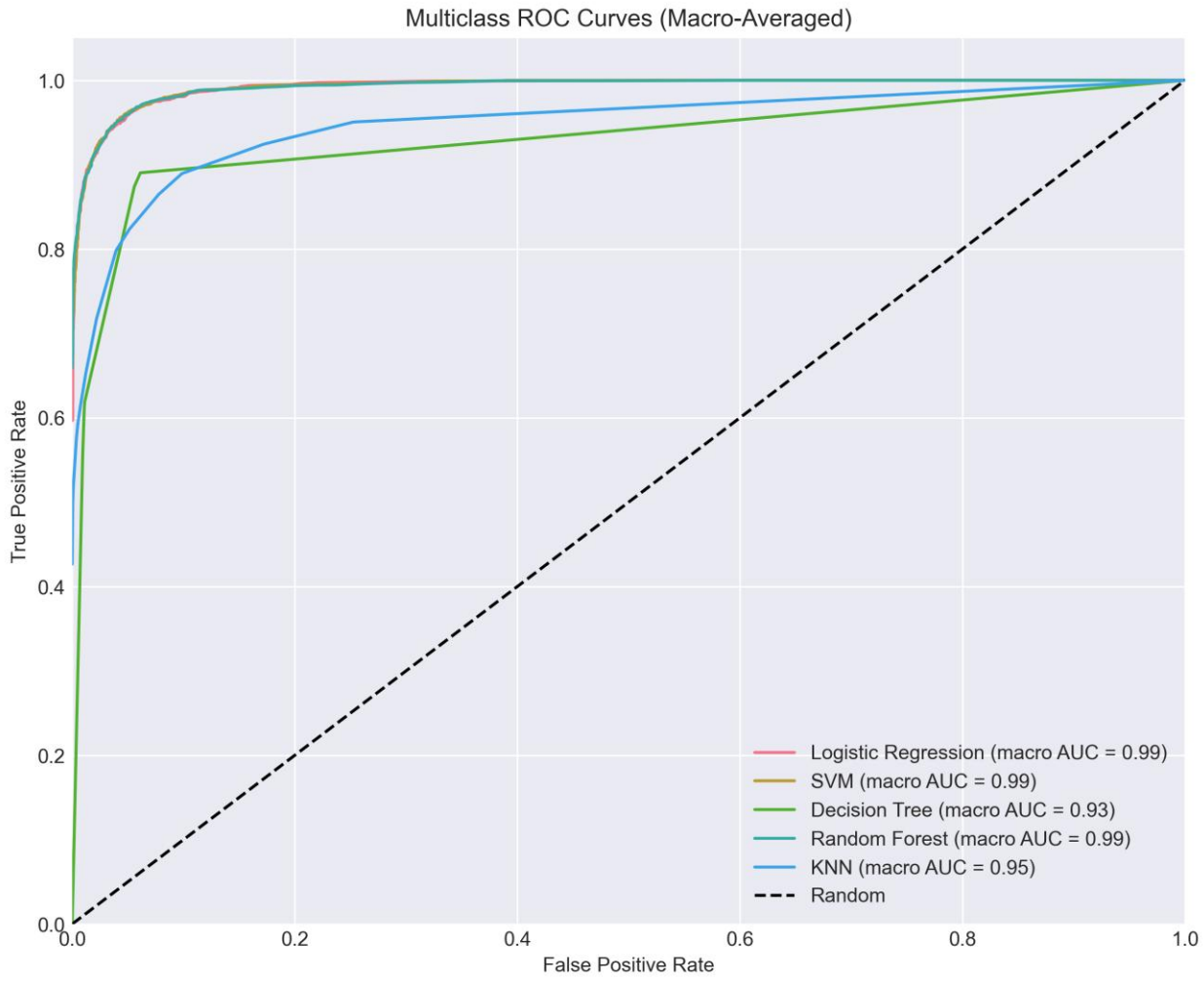

Figure 13: multiclass ROC curve

This ROC plot confirms that Logistic Regression, SVM, and Random Forest are the top-performing models, achieving near-perfect macro-averaged AUC (0.99). KNN performs well (0.95), while Decision Tree lags behind (0.93).

3. Model performance after tuning

	Metric	Before Tuning	After Tuning
0	Accuracy	0.938	0.942
1	Precision (macro)	0.948	0.953
2	Recall (macro)	0.933	0.935
3	F1-score (macro)	0.940	0.942

Figure 14: before and after tuning

The table compares key machine learning evaluation metrics before and after model tuning.we notice a slight improvement across all metrics. The tuning enhanced precision without affecting recall.This suggests the adjustments likely reduced false positives while maintaining the model's ability to detect true positives.

4. Feature important SHAP method:

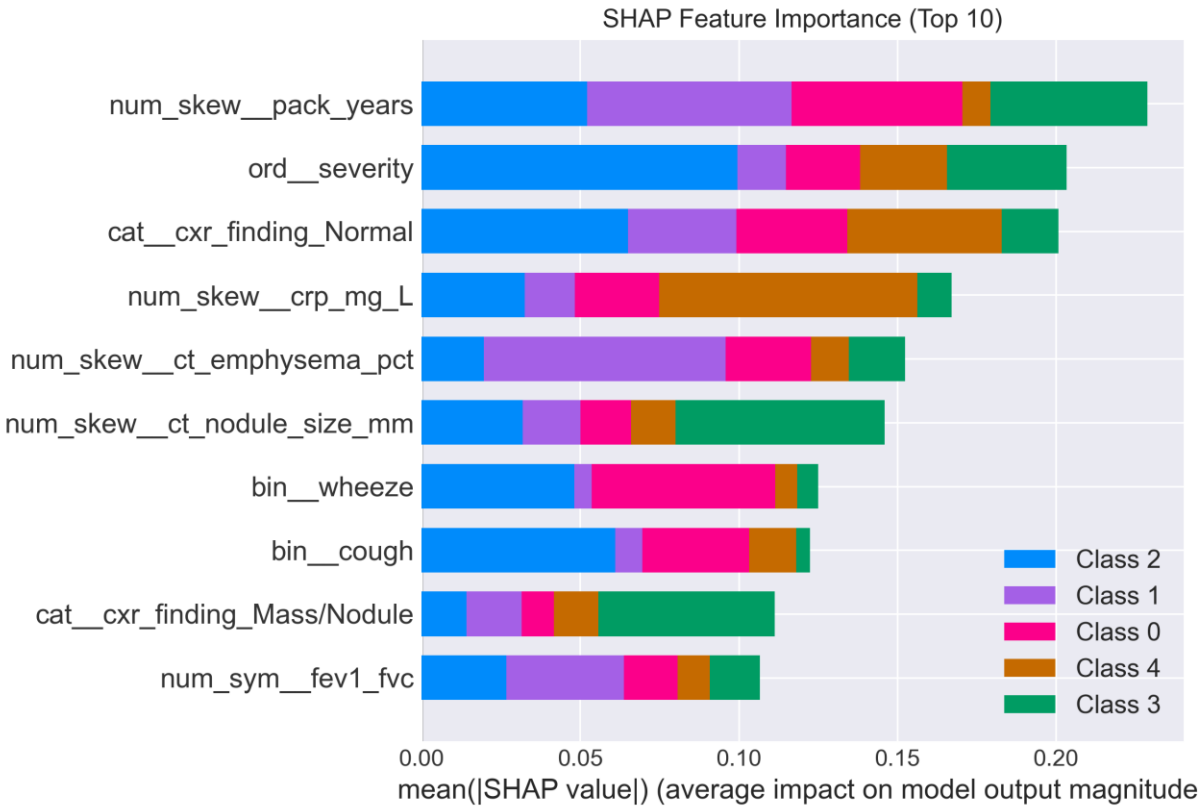

Figure 15: Sharp feature importance

The SHAP feature importance where the `num_skew_pack_years` being the most influential feature, followed by `ord_severity`, `num_skew_crp_mg_L`, and radiological features like `num_skew_ct_nodule_size_mm`. Categorical chest X-ray findings—especially “Normal” and “Mass/Nodule”—also contribute significantly. Feature impacts vary across the five predicted classes, indicating class-specific decision patterns; for example, nodule-related features strongly influence Class 3. Overall, the model relies on both skewed numerical biomarkers and qualitative imaging labels to stratify outcomes.

Chapter 5: Conclusion and future work

1. Introduction:

This chapter discusses key limitations of the dataset. we outline future directions, including deep learning and multimodal integration, to improve real-world clinical utility.

2. Study limitation:

Despite the comprehensive nature several must be acknowledge:

Missing data: The family_history variable has 7,176 missing values (59.8% of records), severely limiting analysis of hereditary factors—a major gap in understanding genetic predisposition.

Cross sectional Design: The study's static snapshot prevents causal inference or modeling of disease progression. Longitudinal tracking is needed, especially for severity transitions (e.g., Mild → Moderate → Severe)

Binary symproms representation: Symptoms (cough, dyspnea, etc.) are encoded as presence/absence (0/1), omitting severity, duration, and frequency—key clinical dimensions that affect diagnosis and management.

Features encoding constraints : Categorical variables like occupation and cxr_finding were one-hot encoded, potentially introducing high dimensionality and sparsity. Some features (e.g., disease_type) were ordinal-encoded for SHAP analysis, which may not reflect true relationships

External validity: Models were trained on a single dataset; external validation on independent cohorts (different regions, healthcare systems) is essential before clinical deployment

Clinical context gaps The dataset lacks treatment history, medication use, and specific diagnostic criteria (e.g., spirometry thresholds for COPD), limiting direct translation to clinical decision-making.

3. Recommendations for Future Research

Longitudinal Cohort Studies: Implement prospective studies tracking patients over time to model disease progression, remission, and transition between severity levels.

Advanced Modeling Approaches:Develop ensemble methods (e.g., stacking Logistic Regression + Random Forest) to improve robustness.Apply deep learning architectures: CNNs for CXR/CT images, RNNs for temporal symptom patterns, and multimodal fusion models.Incorporate survival analysis for time-to-event outcomes (e.g., time to hospitalization, time to cancer diagnosis).

Enhanced Feature Engineering:Create composite indices (e.g., Smoking Risk Score = pack-years × duration × cessation status).Extract interaction terms (e.g., smoking_status × occupation, ct_nodule_size_mm × ct_emphysema_pct).Derive temporal features from repeated measurements if longitudinal data becomes available.

Explainable AI Integration: Deploy SHAP and LIME for real-time model explanation in clinical settings, with clinician-friendly interfaces showing feature contributions to individual predictions.

4. Conclusion:

The findings from this analysis provide a robust foundation for advancing lung disease research and improving clinical care. By addressing the identified limitations and implementing the recommended strategies, healthcare systems can significantly enhance early detection, risk stratification, and personalized treatment approaches for patients with lung diseases. The integration of these insights into clinical practice has the potential to improve patient outcomes, reduce healthcare costs, and enhance quality of life for millions of individuals affected by respiratory conditions worldwide.

Critically, this work demonstrates that data-driven clinical intelligence when grounded in rigorous statistical validation and clinical interpretability can transform reactive care into proactive, precision medicine. Future success hinges on collaboration between data scientists, clinicians, and public health experts, ensuring that AI tools augment not replace clinical judgment.

BIBLIOGRAPHY

Muraina, I.O. (2022) ‘Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts’, 7th International Mardin Artuklu Scientific Research Conference, pp. 502-503

Pinheiro, J.M.H. et al. (2025) ‘The impact of feature scaling in machine learning: effects on regression and classification tasks’, IEEE Access, pp. 1-12. [Available at: https://arxiv.org/html/2506.08274v5].