COPD Risk Prediction

This report presents a machine learning pipeline for predicting the risk of COPD onset in the ageing population. COPD is a chronic respiratory disease with high global prevalence and mortality. Identifying high-risk individuals can enable preventive interventions and support clinical decision-making. All analyses are carried out in R using data collected through a population health survey.

The work is structured across three main steps. The first focuses on data preprocessing, including missing value handling, outlier detection, class imbalance assessment, and normalization. The second covers model training and validation, while the third addresses feature selection to improve model robustness and interpretability.

Data Preprocessing

This section describes the preprocessing steps applied to the dataset. The dataset includes 3980 adult subjects and 18 predictor variables describing each subject's health status at baseline, and one binary outcome variable indicating whether the subject developed COPD after 10 years.

Data Inspection

The analysis started with an initial assessment of the dataset. Summary statistics and missing value counts were computed for each variable. A few variables showed moderate amounts of missing data, most notably fev1_fvc_ratio and education. The outcome variable was found to be imbalanced, with a large majority of subjects (3603) not developing COPD, and only 377 cases of the disease.

Train-Test Split

The dataset was then split into a training set and a test set, using stratified sampling based on the outcome variable to preserve class proportions. The training set included 3184 subjects, while the test set contained 796. This split ensured 302 positive cases in the training set.

Rule of Thumb

To evaluate whether the training set was sufficient for logistic regression, the number of model coefficients was estimated. After encoding categorical variables with dummy variables, the model included 19 coefficients. With 302 positive cases, the ratio of minority-class cases to coefficients was 15.89, satisfying the recommended threshold of at least 10 observations per coefficient.

Outliers

Outliers in the numerical variables were identified using boxplots (displayed in Figure 1) and removed based on a ±4 standard deviation threshold.

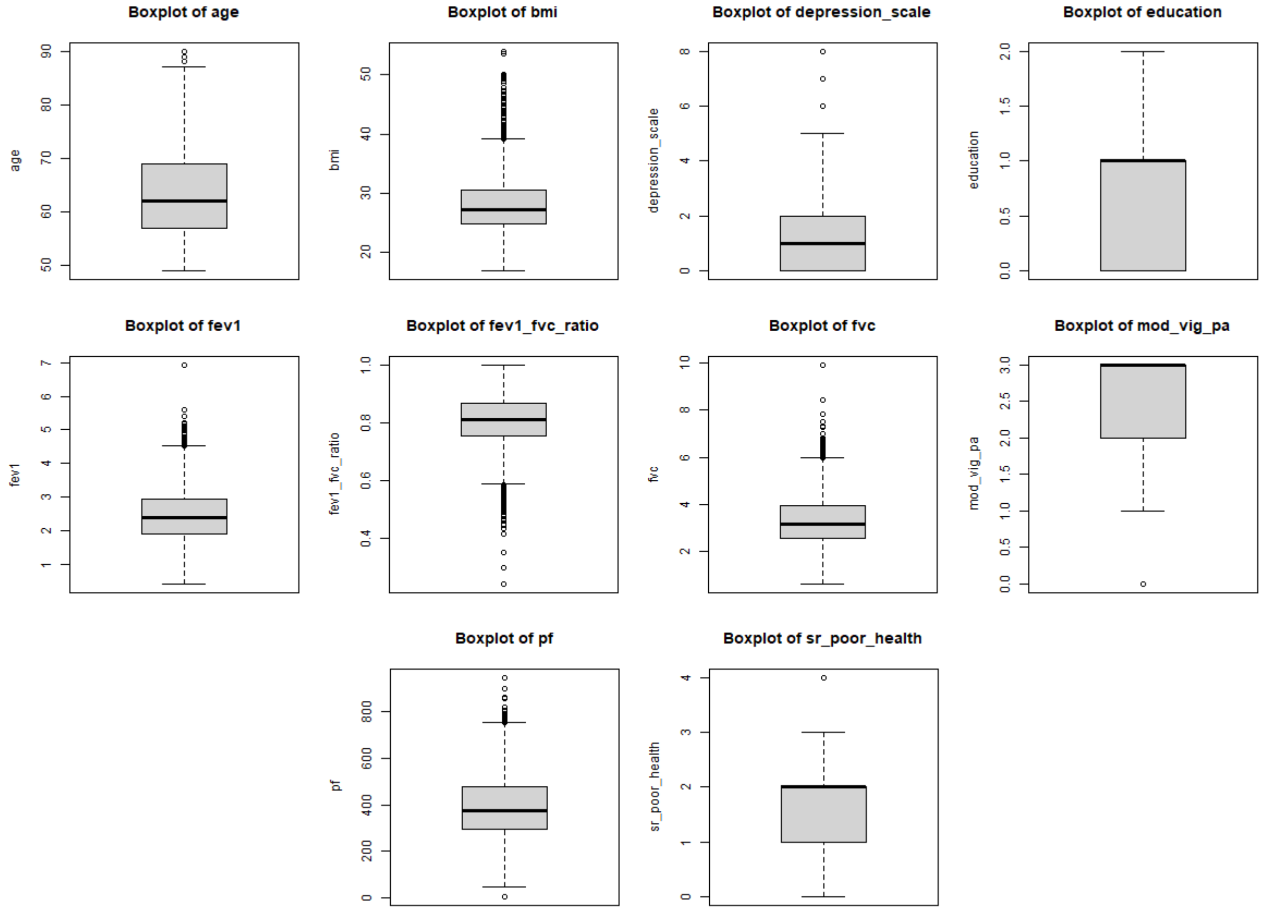

Figure 1. Boxplots of the numerical variables used for outlier detection. Extreme values exceeding ±4 standard deviations from the mean were considered outliers and replaced with missing values.

A small number of extreme values were replaced with missing values: for example, 9 in bmi, 4 in fvc, and 4 in fev1_fvc_ratio.

Collinearity

Correlation between numerical variables was checked numerically and displayed using a correlation matrix (Figure 2).

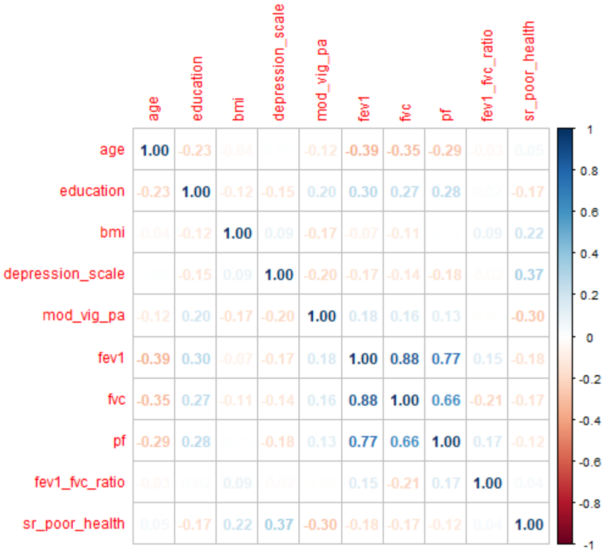

Figure 2. Correlation matrix of numerical variables.

A strong correlation (88.1%) was detected between fev1 and fvc. To avoid redundancy and multicollinearity in the model, fev1 was removed from training and test set.

Missing Values Imputation

Next, missing values were imputed. Numerical variables were filled using the means calculated from the training set. Categorical variables were imputed using the most frequent value (mode). The same values were used to impute the test set. After this step, there were no remaining missing values in either set.

Normalization

All numerical variables were normalized using min-max scaling, based on the minimum and maximum values from the training set. This ensured that all values in the training set fell within the range [0, 1]. As expected, some values in the test set were slightly outside this range, which is a natural result of applying the same scaling parameters to new data.

Training and Testing a GLM

Logistic Regression Full Model

The training set was used to train a logistic regression model for predicting the onset of COPD after 10 years. The glm function in R was applied, specifying a binomial family and using all available predictors.

The output of the model includes 18 predictors (including dummy variables automatically generated for factor variables) and one intercept term. The analysis of model coefficients was based on their estimated effect and associated p-value (significance level: 5%).

Several predictors were found to be significantly associated with the outcome:

- Positive association: gender, smoking (both past and current), short_breath_walking, wheezing, asthma_hx, sr_poor_health. These variables were associated with an increased risk of developing COPD.
- Negative association: mod_vig_pa, fvc, pf, and fev1_fvc_ratio showed a protective effect, reducing the predicted probability of COPD.

Other variables such as age, education, bmi, and depression_scale were not significantly associated with the outcome.

The model achieved an AIC of 1578.1, with a residual deviance of 1540.1 on 3165 degrees of freedom.

Prediction and Model Evaluation on the Test Set

The logistic regression model trained in the previous step was applied to the test set to evaluate its performance. The ability of the model to discriminate between subjects with and without COPD was assessed using the area under the ROC curve (AUC).

The resulting AUC was 0.814, indicating good model performance. The ROC curve is shown in Figure 3a.

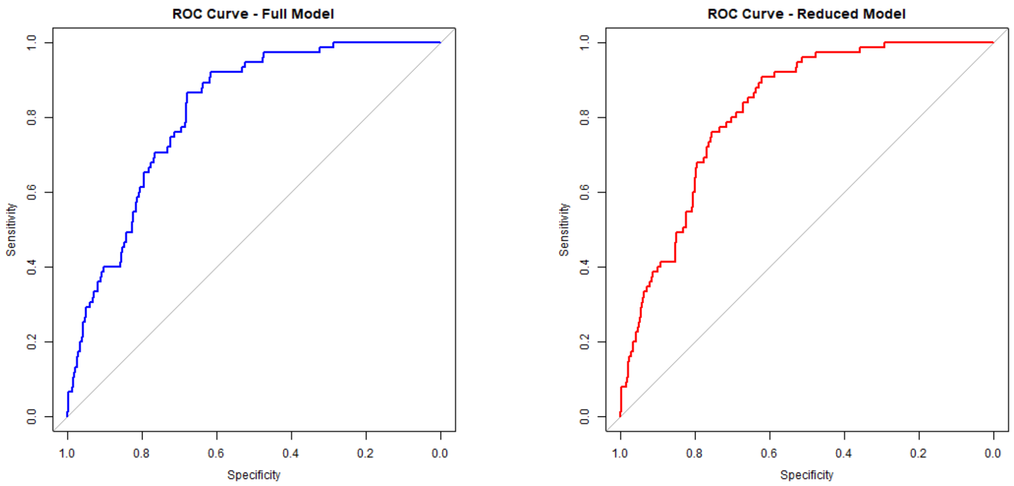

Figure 3. Comparison of ROC curves for the full (a) and reduced (b) logistic regression models on the test set. The reduced model achieved a slightly higher AUC (0.818 vs. 0.814), indicating comparable or better performance with fewer predictors.

Feature selection and model reduction

A backward feature selection procedure was applied to the training set using the step function, starting from the full model. The process iteratively removed variables that did not significantly contribute to model performance, with the goal of reducing the Akaike Information Criterion (AIC).

The following variables were removed during the backward selection process: age, phlegm, wake_breath, depression_scale, bmi, chest_pain, education.

The resulting reduced model included 11 predictors: gender, smoking, mod_vig_pa, short_breath_walking, wheezing, fvc, pf, fev1_fvc_ratio, asthma_hx, sr_poor_health.

The reduced model achieved a deviance of 1544.8 on 3172 degrees of freedom and an AIC of 1568.8, slightly lower than the full model (AIC = 1578.1).

The area under the ROC curve (AUC) on the test set was calculated to assess the performance of the reduced model. The AUC of the reduced model was 0.818 (Figure 3b), slightly higher than the full model (0.814), indicating that the reduction did not compromise model accuracy.

Comparison of selection strategies

To evaluate the robustness of the feature selection, additional procedures were applied: forward selection, stepwise backward, and stepwise forward.

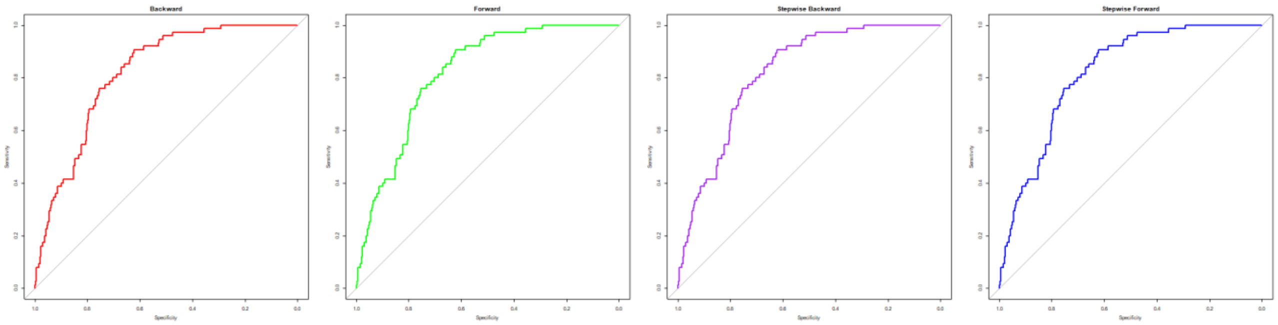

Figure 4. ROC curves for the reduced logistic regression models obtained with different feature selection strategies: (a) Backward, (b) Forward, (c) Stepwise Backward, and (d) Stepwise Forward.

All three methods converged to the same reduced model, selecting the same set of 11 predictors as the backward selection. This confirms that the selected features are consistently informative, and the feature selection process is stable.

Robust Feature Selection

Stability Analysis

To evaluate the robustness of the feature selection procedure, a stability analysis was performed using bootstrap resampling. A total of 50 bootstrap iterations were generated from the training set, stratified by the outcome to preserve class balance. For each iteration, missing values were imputed, numerical variables were normalized, and backward feature selection was applied to the resulting dataset. The variables selected in each iteration were recorded.

At the end of the procedure, the selection frequency of each variable was computed. Variables selected in at least 60% of the bootstrap iterations were retained to train the final logistic regression model. The frequency of selection is shown in Figure 5.

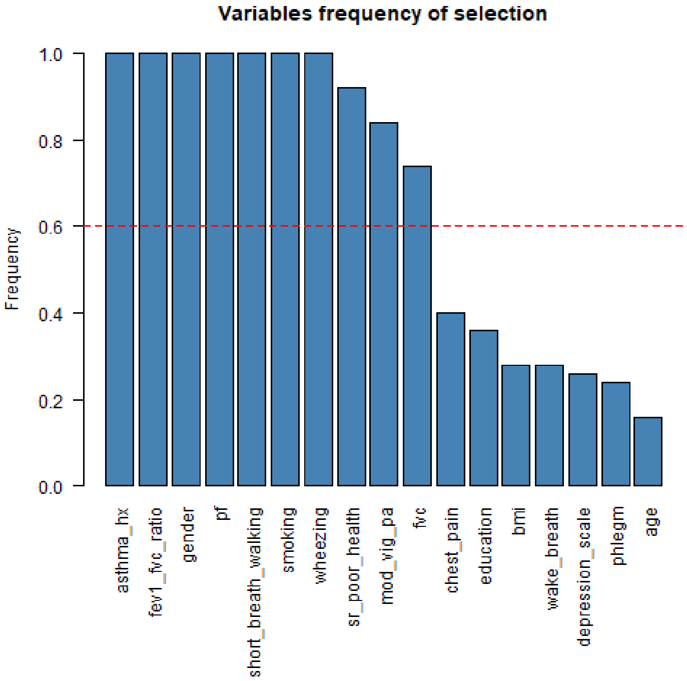

Figure 5. Barplot showing the frequency of selection for each variable across 50 bootstrap iterations. The dashed red line indicates the 60% selection threshold.

The analysis revealed that 11 variables were selected in ≥60% of the bootstrap iterations. The most stable predictors (selected in 100% of the iterations) included asthma_hx, fev1_fvc_ratio, gender, pf, short_breath_walking, smoking, and wheezing. This aligns with the predictors selected by the single-shot backward selection in Lab 3, indicating strong agreement across methods. Conversely, variables such as age, phlegm, depression_scale, and bmi were selected in less than 30% of the bootstrap iterations, suggesting limited contribution to the model's predictive power.

Final Model Evaluation

A logistic regression model was trained using the variables selected in the stability analysis. The final model was applied to the test set, and its performance was evaluated using the area under the ROC curve (AUC). The resulting AUC was 0.818, matching the value obtained with the reduced model in Lab 3. The ROC curve for the final model is shown in.

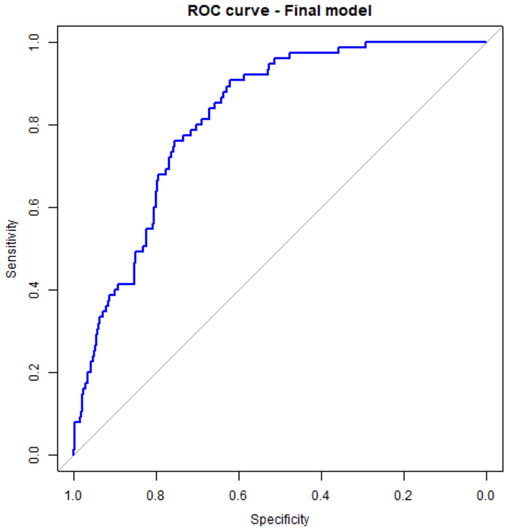

Figure 6. ROC curve of the final model trained on stable variables. The model achieves an AUC of 0.818 on the test set.

This result confirms that the selected subset of predictors is not only stable across bootstrap samples but also effective in maintaining high predictive performance on unseen data.

Comments on the Results

The analysis carried out across the different phases of the project led to the development of a reliable and interpretable model for predicting the onset of COPD. The logistic regression models (both the full and reduced versions) performed well, with an AUC consistently above 0.81 on the test set. This suggests a solid ability to distinguish between individuals who will and will not develop the disease.

The feature selection process proved particularly effective. Through backward elimination, it was possible to reduce the number of predictors without losing predictive power. The reduced model slightly improved performance, showing that many of the excluded variables added little to the model and could be safely removed. This simplification not only helps with interpretation but also reduces the risk of overfitting.

The results of the stability analysis confirmed the robustness of the selected features. Running the selection process across multiple bootstrap samples consistently led to the same core set of predictors. This strengthens confidence in the model’s generalizability and supports its use beyond the current dataset.

It is also worth noting how the model's findings align with clinical expectations. Variables such as smoking status, spirometric measurements, and the presence of respiratory symptoms emerged as important predictors, consistent with known risk factors for COPD. This coherence between data-driven results and medical knowledge adds further credibility to the model.

In summary, the modeling pipeline proved to be well-structured and effective. The final model balances predictive accuracy with simplicity and clinical interpretability, making it a potentially useful tool for identifying individuals at higher risk of developing COPD.