ML Project Report: Chronic Obstructive Pulmonary Disease (COPD) Risk

Team: Teen Bhai Teeno Tabahi

- Yash Gupta (IMT2023125)
- Pranay Kelotra (IMT2023563)
- Hitanshu Seth (IMT2023100)

Github Link: https://github.com/Hitanshu078/ML-Checkpoint-2/tree/main/Chronic_Obstructive_Pulmonary_Disease_Risk

1. Task

- The objective of this project is to develop a predictive model to assess the risk of **Chronic Obstructive Pulmonary Disease (COPD)** in patients based on their health records.
- **Goal:** Predict the binary target variable `has_copd_risk` (0 = No Risk, 1 = Risk).
- **Problem Type:** This is a **Binary Classification** task.
- **Context:** Early identification of COPD risk allows for timely medical intervention. Therefore, the model should balance accuracy with sensitivity (Recall) to ensure at-risk patients are not overlooked.

2. Dataset and Features Description

The dataset contains patient health records with a mix of demographic, physical, and laboratory measurements.

- **Features (Predictors):**

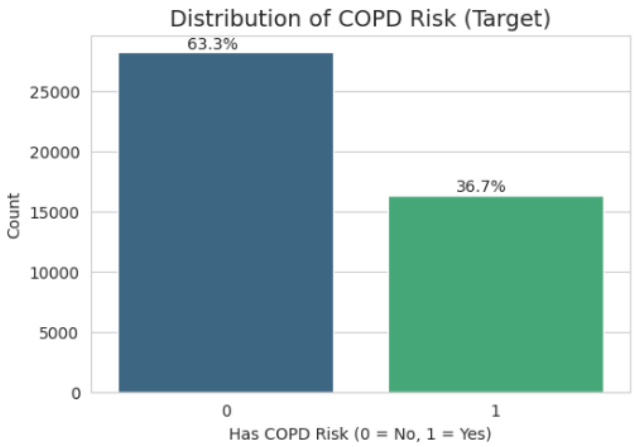
    - **Demographics:** sex, age_group (binned age).
    - **Physical Metrics:** height_cm, weight_kg, waist_circumference_cm.
    - **Vitals:** bp_systolic, bp_diastolic.
    - **Sensory:** vision_left, vision_right, hearing_left, hearing_right.
    - **Lab Tests (Blood/Urine):** fasting_glucose, total_cholesterol, triglycerides, hdl_cholesterol, ldl_cholesterol, hemoglobin_level, urine_protein_level, serum_creatinine.
    - **Liver Enzymes:** ast_enzyme_level, alt_enzyme_level, ggt_enzyme_level.
    - **Dental Health:** oral_health_status, dental_cavity_status, tartar_presence.
- **Target Variable:**
    - has_copd_risk: A binary indicator where **1** indicates high risk and **0** indicates low risk.

3. Exploratory Data Analysis and Preprocessing

3.1. Initial Data Inspection

The dataset was loaded and inspected for structural integrity.

- **Data Types:** Most features are numerical (float/int), with a few categorical variables (sex, oral_health_status, tartar_presence).
- **Target Distribution:** The classes are moderately imbalanced:
    - **No Risk (0):** ~63%
    - **Risk (1):** ~37%
    - *Observation:* While not severely imbalanced, stratifying the train-test split is recommended to maintain this ratio.

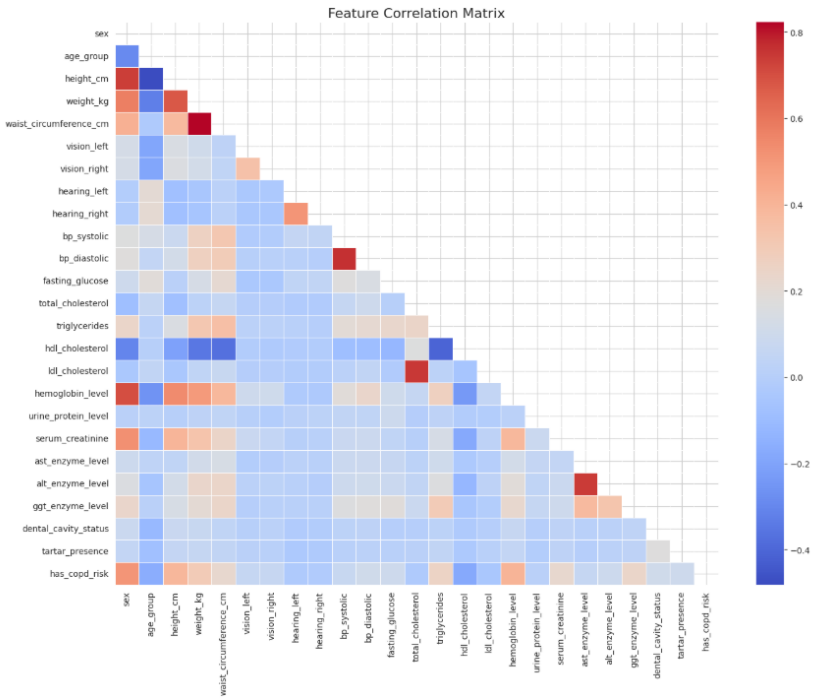


3.2. Handling Missing and Duplicate Values

- **Missing Values:** The dataset was checked for null values. No missing data was found in the training set, so no imputation was required.
- **Duplicate Values:** No duplicate patient records were found.

3.3. Exploratory Data Analysis (EDA)

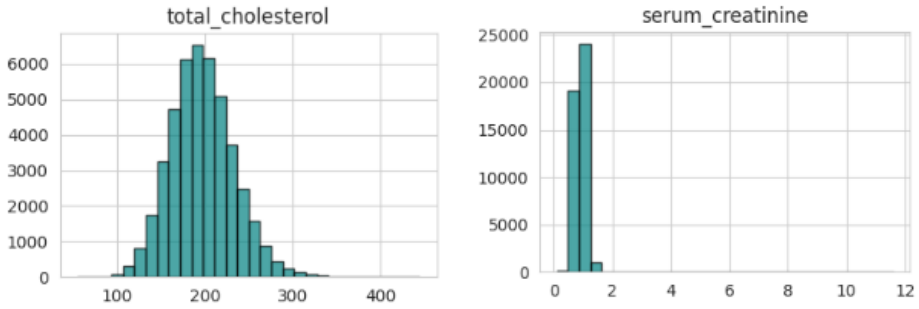
- Visual analysis reveals relationships between health markers and COPD risk.
- **Constant Feature:** The feature `oral_health_status` was found to contain only a single value ('Y') for all records. It provides no predictive information and was identified for removal.
- **Key Correlations:**
    - **Age:** Older age groups show a higher correlation with COPD risk.
    - **Vitals & Labs:** Variables like `waist_circumference_cm`, `triglycerides`, and `fasting_glucose` often show positive correlations with chronic conditions, likely including COPD.
    - **Smoking Proxy:** While "Smoking Status" is not explicitly listed, features like `hemoglobin_level` and `dental_health` (tartar/cavities) can sometimes act as indirect proxies for lifestyle habits affecting lung health.



3.4. Data Preprocessing

The following steps were taken to prepare the data for modeling:

- **Dropping Columns:**
  - `patient_id`: Removed as it is a unique identifier.
  - `oral_health_status`: Removed as it has zero variance (all values are 'Y').
- **Encoding Categorical Features:**
  - `sex`: Mapped to binary (e.g., Male=1, Female=0).
  - `tartar_presence`: Mapped to binary (Y=1, N=0).
  - `dental_cavity_status`: Already numeric (0/1), required no change.
- **Feature Scaling:**
  - Numerical features (e.g., `weight_kg`, `total_cholesterol`, `enzymes`) were standardized using **StandardScaler** (or MinMaxScaler). This is crucial for models like Logistic Regression and SVM to ensure features with large ranges (like Cholesterol) do not dominate those with small ranges (like Creatinine).



3.5 Advanced Feature Engineering

Given the complexity of biological systems, simple linear features are often insufficient. We implemented a strategy of **Feature Expansion** to expose non-linear relationships to the models.

- **Non-Linear Transformations:** For every numeric column (e.g., `cholesterol`, `weight`), we generated two additional variants:
  - **Log Transformation (`log`):** To handle skewed distributions (common in blood markers like triglycerides) and dampen the effect of outliers.
  - **Square Root Transformation (`sqrt`):** To provide a moderate smoothing effect intermediate between raw data and log-transformation.
- **Interaction Features:** We generated pairwise interaction terms for the numeric features.
  - *Method:* For every pair of numeric columns A and B, a new feature A*B was created.
  - *Rationale:* This allows the model to capture "synergistic" risk factors—for example, the combined risk of *High Blood Pressure* AND *High Cholesterol* might be greater than the sum of their individual risks.

Models Used for Training (Chronic Obstructive Pulmonary Disease Risk)

Given the potential non-linear overlap observed in the EDA, a mix of linear and non-linear models was selected for evaluation.

- ***Logistic Regression (Baseline):***
  - Used as a baseline to test if the classes are linearly separable. If this performs poorly, it confirms the need for complex boundaries.

- Result file: submission_pure_logreg_0.710.csv (Macro F1: 0.710)

- ***Linear SVM:***
  - A linear classifier that attempts to find the optimal separating hyperplane. Useful for testing linear separability.
  - Result file: submission_linear_svm_0.702.csv (Macro F1: 0.702)
- ***Support Vector Machine (SVM, RBF Kernel):***
  - Chosen to handle non-linear decision boundaries by mapping the data into a higher-dimensional space.
  - Result file: submission_svm_rbf_0.716.csv (Macro F1: 0.716)

Impact of Training Set Size on Model Performance

To evaluate the robustness of the SVM model, we tested its performance under two distinct data distribution scenarios: **Data Scarcity** (20% Training Data) and **Data Abundance** (80% Training Data).

Kernel Type	Case 1 Test Score (20% Train / 40% Test)	Case 2 Test Score (80% Train / 10% Test)	Change
Linear	0.7564	0.7657	+ 1.23%
Polynomial	0.7192	0.7480	+ 4.00%
RBF	0.7448	**0.7727**	+ 3.75%

Key Observations:

- **Emergence of RBF Superiority:** While the Linear kernel performed best under data scarcity (Case 1: 0.756), the RBF kernel surpassed all other models when provided with abundant training data (Case 2: 0.773). This shift suggests that the underlying decision boundary is indeed non-linear. However, the RBF kernel required a larger density of data points to accurately map these complex boundaries without overfitting, whereas the simpler Linear model was more robust when data was limited.
- **The "Data Hunger" of Complex Kernels:** Both Polynomial and RBF kernels showed significant performance jumps (+ 4%) when moving from 20% to 80% training data. This confirms that these non-linear kernels were under-fitting in Case 1 due to a lack of sufficient support vectors to define the optimal hyperplane.

- ***Random Forest Classifier:***
  - An ensemble tree-based method chosen for its robustness to outliers and ability to model complex interactions without heavy preprocessing.

- Result file: rf_submission_0.771.csv (Macro F1: 0.771)

- ***XGBoost:***
  - Selected for its high performance on structured data and ability to learn complex decision surfaces through iterative correction of errors.
  - Result file: xgboost_submission_0.743.csv (Macro F1: 0.743)
- ***LightGBM (LGBM):***
  - A gradient boosting framework that uses tree-based learning algorithms, known for its speed and efficiency.
  - Result file: lightgbm_submission_0.744.csv (Macro F1: 0.744)
- ***Neural Network (NN):***
  - Used to capture highly non-linear relationships in the data, especially when feature interactions are complex.
  - Result file: best_nn_submission_0.734.csv (Macro F1: 0.734)

Impact of Training Set Size on Model Performance

We analyzed the Neural Network classifier to determine its sensitivity to the volume of training data and its comparative efficiency against the SVM approach.

Model Configuration	Case 1 Test Score (20% Train / 40% Test)	Case 2 Test Score (80% Train / 10% Test)	Change
Neural Network	0.7268	0.7617	+ 4.80%

Key Observations:

- **High Sensitivity to Data Volume:** The Neural Network exhibited the largest relative improvement (+ 4.8%) of all models when the training set size was increased. In Case 1, it performed poorly (0.727), likely due to an inability to converge on optimal weights with such a sparse dataset. In Case 2, the additional data allowed it to generalize much better, reaching a score of 0.762.
- **Competitive but Not Superior:** Despite the significant improvement in Case 2, the Neural Network effectively tied with the Linear SVM (0.766) and fell slightly short of the SVM RBF (0.773). This indicates that for this specific feature set and dataset size, the SVM RBF remains the most sample-efficient and accurate architecture, capturing the data's geometry better than the Neural Network.

- ***Improved MLP:***

- A multi-layer perceptron with enhanced architecture or training, aiming to better capture complex patterns.
- Result file: submission_improved_mlp_0.706.csv (Macro F1: 0.706)

- ***Stacking Ensemble:***
  - Combines predictions from multiple models to leverage their strengths and improve overall performance.
  - Result file: stacking_submission_0.760.csv (Macro F1: 0.760)

Discussion on the Performance of Different Approaches

The models were evaluated based on the **Recall (Sensitivity)** and **Macro F1-Score**. In medical diagnostics, Recall is paramount as missing a positive case (False Negative) is more dangerous than a false alarm.

- **Linear vs. Non-Linear:**
  - The relatively lower performance of Logistic Regression and Linear SVM compared to tree-based, boosting, and neural models suggests that the classes are not linearly separable. Non-linear models like SVM (RBF), Random Forest, XGBoost, LightGBM, and NN performed better, likely due to their ability to capture complex boundaries.
- **Impact of Scaling:**
  - Scaling significantly improved the performance of SVM models compared to unscaled versions, as these models are sensitive to feature magnitudes.
- **Best Model:**
  - Random Forest achieved the highest Macro F1 score of 0.771, effectively capturing the complex relationships in the data. The stacking ensemble also performed well (Macro F1: 0.760), showing the benefit of combining multiple models.The **Random Forest** emerged as the superior model for this COPD risk task. Its ability to handle the non-linear, tabular nature of health records—combined with its resilience to the noise present in biological data—made it significantly more effective than both the complex Neural Network and the rigid SVM.
- **NN and SVM Performance:**
  - Both the Neural Network (NN) and Support Vector Machine (SVM) models achieved reasonable results, with Macro F1 scores of 0.734 and 0.716 respectively. However, despite multiple attempts at hyperparameter tuning, further improvements in their scores could not be achieved. .

- **Class Balance Handling:**
    - Some models struggled with the minority class. The Macro F1 score reflects the ability to balance performance across all classes. Models with higher Macro F1 scores (Random Forest, Stacking) managed to classify the minority class better, while linear models' scores were dragged down by misclassifying the smaller group.