Respiratory Disease Classification using Machine Learning

I. Project Overview

This project will determine strong machine learning models to categorize respiratory diseases using clinical, biological, imaging, and environmental data. The aim is to find the most useful predictive model with a high level of methodological rigor, reproducibility, and suitable treatment of real-life healthcare data.

The dataset consists of heterogeneous measurements, including laboratory measurements, CT-based measurements, symptoms, comorbidities, and exposure variables of the patient. Since there are various types and sizes of features, a hierarchical approach to preprocessing was adopted to achieve the best model performance and remain interpretable.

II. Models Implemented

The project evaluates multiple classification algorithms to ensure a comprehensive comparison across linear, distance-based, and non-linear approaches:

- Logistic Regression — interpretable baseline model
- Support Vector Machine — effective in high-dimensional spaces
- K-Nearest Neighbors — captures local similarity between patients
- Decision Tree — provides transparent decision rules
- Random Forest — improves predictive performance through ensemble learning