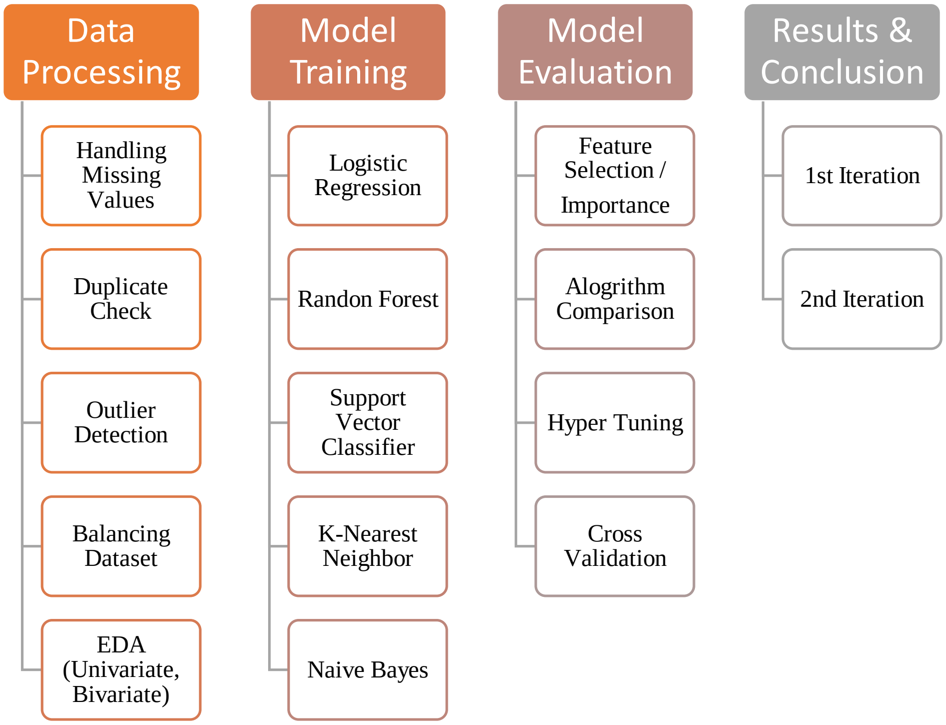
Approach



To begin with, initial preprocessing and cleaning of the raw data was performed. This process included data cleaning, handling missing values, checking duplicates, outlier detection and balancing dataset. In order to understand the relationship between variables (symptoms) and their impact on predicting the occurrence of a disease, techniques such as univariate and bivariate analysis were used as part of Exploratory Data Analysis (EDA). The goal of EDA is to uncover patterns, trends, or distributions within the dataset by visualizing the data through plots, charts, or summary statistics.

In order to determine the models' accuracy and reliability, it is essential to evaluate their performance. Therefore, MCC and Brier Score were used as standard metrics for binary classification.

Once the data was explored and preprocessed, modeling was conducted using SKlearn package.

Feature Selection aims to choose a subset of input variables by removing features that provide no predictive information. The feature selection method is split into three categories: 1) filters; 2) wrappers; and 3) embedded methods.

For the purpose of this study the following Feature Selection techniques were used: 1) filter (chi-square, ANOVA); 2) wrapper (forward selection, backward elimination, recursive feature elimination); 3) embedded (Decision Tree based).

Next step was Feature Importance. Feature importance quantifies the importance of each feature compared to the other features in the model. It helps identify the most influential features relative to the others. It represents the contribution of each feature in the models decision making process.

Five algorithms were used for modeling: Logistic Regression, Random Forest, Support Vector Classifier, K-Nearest Neighbor and Gaussian Naïve Bayes algorithm. The modelling process was repeated twice, once with default parameters and the second after hyper-tuning the parameters.

After modelling, cross validation strategies such as Leave one out, Stratified K Fold, and Holdout Method using Logistic Regression and Support Vector are explored in this paper. Lastly, the results and any recommendations are discussed later on in this paper.