Task 1:

Due date for submission: 25/4/2021

The aim of this task it to develop a classifier for estimating if a patient is infected by Covid -19 by using regular blood tests.

Data: The data can be downloaded from the Kaggle website (only the file dataset.xlsx from the "Bloot Test dataset" folder is needed for this task):
https://www.kaggle.com/mridulmittal/virtualcoronadetection-test

The target attribute is SARS-Cov-2 exam result (negative or positive)

Steps:

1. Read the following paper that describe how decision forest can be used for training a classifier:

   Marcos Antonio Alves, Giulia Zanon de Castro, Bruno Alberto Soares Oliveira, Leonardo Augusto Ferreira, Jaime Arturo Ramírez, Rodrigo Silva, Frederico Gadelha Guimarães, Explaining Machine Learning based Diagnosis of COVID-19 from Routine Blood Tests with Decision Trees and Criteria Graphs, Computers in Biology and Medicine, 2021, 104335, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.104335
   (https://www.sciencedirect.com/science/article/pii/S0010482521001293)
2. Write the code for reproducing (as much as possible) the predictive performance results of the methods: Logistic Regression, XGBoost and Random Forest presented Table 3 in the above paper. The code can use any existing Python/R packages including the XGBoost package and scikit-learn. Make sure to:
   a. Perform the pre-processing described in Section 5.2
   b. Perform hyper-parameter optimization and evaluate the models using nested cross-validation procedure as described in section 5.3
3. Suggest at least 5 new features that can be created from the raw features (e.g. ratio between two existing features) with the aim to improve the predictive performance. Please report the new results.
4. Evaluate the CatBoost and LightGBM in addition to Xgboost and report their predictive performance
5. Watch the video about SHapley Additive exPlanations (SHAP) method for about Model Explainability in the medical domain. The video is available in the link:
   https://drive.google.com/file/d/1Y3uByS0aPp_zvwhErtwXN-vohyNZ8RuW/view
6. Illustrate the feature importance using SHAP as described in Fig 7.0 – Please provide the figures for RF, Xgboost, Catboost and LightGBM. Refer to: https://github.com/slundberg/shap for an example.

   Please submit:

   1. Your python/R code for running all the above mentioned steps
   2. A word document for reporting the evaluations, feature importance by SHAP and for describing your suggestions for new generated features.