Midterm Project

Risk Analysis of Hospitalized Covid-19 Patients

One of the problems with Covid-19 pandemic, which is a global health threat, is the inability to predict the disease progress of hospitalized Covid patients using the clinical data. It is known that while some patients only suffer form mild infections, others' conditions worsen and need intensive care support and some patients eventually need intubation. However, this progression cannot be predicted during the early phases using clinical data gathered from the patients. A predictive ability of these risks at the early stages of the disease would be very valuable for the management of the disease.
In this project, it is aimed to evaluate the performance of potential predictive models utilizing the early-disease clinical data and guessing if a patient would need intensive care support or intubation. A published clinical dataset is provided for this task.
For this purpose, using the data features (laboratory and clinical values of the patients) the state of these patients (intensive care support, intubation) will be predicted separately. Random forest and Gradient Boosted Trees algorithms will be used as the predictive models.

Data

In this assignment, data gathered for study are provided to you. You can download the data from
`https://drive.google.com/file/d/1xD48p6EnWp9J1Yd0p-L3DeYZFqSJEYft`

In the data file *covid19_dataset.csv* the data are provided in tabular format (as csv file) and each row belongs to a patient (total 1439 patients). Among the columns, one column contains the patients IDs (column name: 'ID'), one columns contains the intensive care support information (column name: 'INTENSIVE CARE', YES: intensive care support received, NO: intensive care support not needed), one columns contains the intubation information (column name: 'INTUBATION', YES: patient intubated, NO: patient did not need intubation). The remaining 34 columns contains the laboratory results and clinical values. These 34 values should be used as the features of the machine learning models to be developed.
NOTE: MISSING VALUES CAN BE REPLACED BY AVERAGES OR THEY MIGHT BE NEGLECTED. 'INTUBATION' SHOULD NOT BE USED AS AN INPUT FEATURE IN THE PREDICTION OF 'INTENSIVE CARE' CLASSIFICATION, AND 'INTENSIVE CARE' SHOULD NOT BE USED AS AN INPUT FEATURE IN THE PREDICTION OF 'INTUBATION' CLASSIFICATION.

Goal

With this project, it is expected to have the highest possible correct classification scores (see performance measures below) both for intensive care support and intubation separately. 2 different classifications are going to be performed.

Classification Algorithms

The classifications are going to be performed in Random Forest and Gradient Boosted Trees. The performance of these two algorithms are going to be compared. You are free to use **ANY** programming language/platform. As the Gradient boosted tree algorithms, you can use any of these three if you wish:

- XGBoost: `https://xgboost.readthedocs.io/en/latest/index.html`
- LightGBM: `https://lightgbm.readthedocs.io/en/latest/`
- CatBoost: `https://catboost.ai/`

Performance Measures

Sensitivity and *specificity* is requested as the output of the program performance.
Sensitivity: $= \frac{correct\ number\ of\ prediction\ of\ the\ first\ class}{totalnumberofelementsinthefirstclass}$
Specificity: $= \frac{correct\ number\ of\ prediction\ of\ the\ second\ class}{totalnumberofelementsinthesecondclass}$

NOTE: THE PERFORMANCES WILL BE MEASURED EITHER WITH SPLITTING THE DATA INTO TEST-TRAIN DATASETS, OR USING CROSS-VALIDATION.

Evaluation

The results of four classification tasks, on their performances with Random Forests and Gradient Boosted trees are expected.
DEADLINE IS MAY 21st, 2021. The result reports will be sent to e-mail: nalbantoglu@odev.erciyes.edu.tr
Good Luck!