# Midterm Project

## Diagnosis of Cancer Using Blood Microbiome Data

Recently it was discovered that DNA samples of microbes in human blood might be indicators of several tumor types in human body. The study [1] indicating this is published in `https://www.nature.com/articles/s41586-020-2095-1`
You can find more information and press coverage here, if you are interested:
- `https://www.genengnews.com/news/machine-learning-models-detect-early-stage-cancers-using-cancer-associ`

- `https://www.genomeweb.com/sequencing/microbiome-based-cancer-dx-emerges-cancer-genome-atlas-reanalysis.XrK_X3UzYVk`

[1] Poore, Gregory D., et al. "Microbiome analyses of blood and tissues suggest cancer diagnostic approach." Nature 579.7800 (2020): 567-574.

### Data

In this assignment, data gathered for study are provided to you. You can download the data from
`https://drive.google.com/file/d/15evTOZTYuopoBnolYWOPy2P_VF6wnlFm/view?usp=sharing`

We have blood sample data of 355 people with 4 most common cancer types: Colon cancer, breast cancer, lung cancer, and prostate cancer.
You are given a label file, *labels.csv*, indicating the sample names, and the disease type of each person with the corresponding sample name. The data are stored in *data.csv*. Again, each row has the sample name of the corresponding person, and the remaining are the number of DNA fragments belonging to each microorganism type (virus or bacteria). 1836 different microorganisms appear as features.

NOTE: SINCE EACH DATA COMPONENT IS GIVEN AS COUNTS, DIVIDING EACH COUNT TO THE SUMMATION OF ALL 1836 COUNTS FOR EACH SAMPLE MIGHT GIVE BETTER PERFORMANCE. oTHER DATA NORMALIZATION SCHEMES CAN BE USED AS WELL.

### Goal

With this project it is expected to have the highest possible correct classification scores (see performance measures below). 4 different classification (each cancer vs. others) 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 $25^{th}$, 2024, 23:59 O'clock. The result reports will be sent to e-mail: nalbantoglu@odev.erciyes.edu.tr
Good Luck!