# Laboratory Work 2 ML & Data Mining Fall 2022

## General Instructions:

The labs must be uploaded to Canvas as a PDF file. To answer each question, first copy-paste the question itself, then provide the answer bellow. The answer must contain the code that answers each question, the output of the code and most importantly the explanations when these are required.

The PDF files should be saved in the following format: firstname_lastname_LAB2_MLDM

**Assignment:**

Use the dataset Data1.csv to implement the list of classification models below in order to predict whether a tumor is malign or beginning.

The dependent variable in this dataset is Class, it is equal to 2 if the tumor is benign and 4 if it is malign. The remaining columns are features that describe different characteristics of each tumor.

Implement the following models:

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier (with nb_trees = 10)
4. K- Nearest Neighbors (K-NN)
5. Naïve Bayes
6. Support Vector Machine (SVM) – use the 'rbf' kernel

To do so follow the steps bellow:

1. Import the libraries ***(0.25 pt)***
2. Import the dataset ***(0.25 pt)***
3. Split the dataset into Training and Test groups (use 20-80 split, i.e. 20% of data will be used for the Test group and 80% for training). ***(0.5 pt)***
4. Perform feature scaling. (do not scale y – remember y=0/1 so it needs no scaling). ***(0.5 pt)***
5. Train each of the above models and make predictions. Then compare the results by displaying the predicted values of y next to the test values of y in a two-dimensional array. ***(2 pt)***
6. Create and print the Confusion Matrix and the accuracy scores for each model. ***(1 pt)***
7. For each model, use the K-fold cross-validation (use K=10; in python - cv=10). Print the mean of all 10 accuracy scores for each model and their standard deviations. ***(1 pt)***
8. For each model compare the accuracy scores computed using cross-validation in (7) versus when using only one test set in (6). Are the mean accuracy scores from cross-validation higher or lower in comparison to the corresponding scores in (6)? Did you expect them to be higher or lower? Why? ***(1.5 pt)***

9. Choose the best performing model based on the results from performing the k-fold cross-validation. Comment on your choice. ***(0.5 pt)***
10. For the chosen best performing model select two hyperparameters you would like to tune. (*Learn what the chosen hyperparameters do to your model. You can find the description of hyperparameters for your respective model by using the scikit-org API https://scikit- learn.org/stable/modules/classes.html).* Discuss the two hyperparameters and how they impact the model (for example if you choose to tune C in the SVM model, explain that C is responsible for regularization). Choose 2, 3 or 4 different values for each of these hyperparameters to use with grid search. Explain your choice and, if you have any, make some expectations (you could speculate for example which values of the chosen hyperparameters you believe will make the model perform best and why). ***(1.5 pt)***
11. Use grid search to find the best values for your hyperparameters. Print the accuracy of the model and the hyper parameter values of the best model. Do these values confirm your intuition? Does your model have a better accuracy than before tuning? ***(1 pt)***