# Laboratory Work 2

# ML & Data Mining

# Fall 2021

**Q1.**
Watch the video on Support Vector Machines:
https://www.youtube.com/watch?v=efR1C6CvhmE

**Q2.**
Use the Social_Network_Ads.csv file to build an SVM model.

The dataset contains data about the age and salary of customers of a company that produces cars. Recently the company released a new SUV and organized a marketing campaign. The dependent variable (Purchased) is a categorical variable that takes the value of 1 if a customer purchased the new SUV after the campaign and zero otherwise.
Navigate scikit learn library's API here: https://scikit-learn.org/stable/modules/classes.html to find the description of the SVC classifier in the svm module of the library. Use the SVC classifier to predict whether one will purchase the new SUV or not. Follow the following structure in your approach:

1. Import the libraries
2. Import Dataset
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).
4. Perform feature scaling. (do not scale y – remember y=0/1 so it needs no scaling).
5. Train the SVC classifier (for the classifier specify random_state=0 to ensure that everybody gets the same results. Also choose the 'linear' kernel -- you must specify this choice, otherwise the default kernel is 'rbf').
6. Make predictions and compare the results by displaying the predicted values of y next to the test values of y in a two-dimensional array.
7. Create and print the Confusion Matrix and the accuracy score of the model and interpret the two.
8. Use the K-fold cross-validation method (use K=10; in python - cv=10) to ensure that the accuracy score calculated only on one test set was not just a lucky occurrence. Print the best accuracy score and the standard deviation of the scores computed. Comment on the result, compare to the score computed when using only one test set in (7).
9. Implement the grid search method to tune the following two hyper parameters:

**C** - is a parameter of regularization. One can control the possibility of over-fitting by specifying different values of C. Try the following values for C = 0.25, 0.5, 0.75, 1.

**Kernel** – is the kernel used to manipulate the data. Try the following two kernels: 'linear', 'rbf'.

10. Print the accuracy score of the best performing model found by the grid search. Print the parameters of the best model.

**Q.3.**

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)

In your implementation follow the first 8 steps specified in Q2, that means, for each model create the confusion matrix and implement the k-fold cross validation.

**Q.4.**

Choose the best performing model based on the results from performing the k-fold cross-validation. Discuss your choice.

**Q. 5.**

For the chosen best performing model select two hyper parameters you would like to tune using the grid search method. Learn what the two chosen hyper parameters do to your model. Look into the description of hyper parameters for your respective model by using the scikit-org API https://scikit-learn.org/stable/modules/classes.html. Discuss here the two hyper parameters you chose and how they impact the model. Choose between 2 and 4 different values for each of these hyper parameters to test with grid search. Explain your choice and, if you have any, make some expectations (you could speculate for example which values of the chosen hyper parameters you believe will make the model perform best and why).

Use grid search to find the best model. Print the accuracy of this model and the hyper parameter values of the best model. Do these values confirm your intuition?