# Machine learning
# Fall 2023
# HW1

---

## PartA - theory questions

1. How can using the Random Forest algorithm solve the problem of overfitting?

2. Considering the values in the table below:

a) Find the entropy of hunting a buffalo by condition of the unknown creature being a crocodile or not.
b) If we know the creature is a crocodile, how much information do we gain about hunting the buffalo?
c) Find the entropy of the creature not being a crocodile if the buffalo is not hunted.

| | Be eaten | Not to be eaten |
|---|---|---|
| Crocodile | 38/100 | 26/100 |
| Not a crocodile | 14/100 | 22/100 |

## PartB - practical questions

1. In this exercise, we aim to implement a decision tree classifier from scratch for multi-class classification.

a) Train this model on the given dataset and report the results of trained models with different hyperparameters. You should choose the best value for each of the hyperparameters with a plot and analysis.
b) Using available libraries, train Random Forest and Gradient Boosting models on this dataset.Then, train your models with 75%, 50%, 25%, and 100% of the data separately, and draw Learning Curve plots for each of them using Scikit-learn. Finally, write your analysis of the obtained results.

> Notes:
> 1. Note that alongside the implementation provided in the notebook, you can have any additional implementation; however, you must also provide complete documentation for it. Otherwise, you will not receive any score. A small but useful change could be to change the input of the function from data (X, Y) to their indices in the original dataset (indexes).
> 2. Make sure to use training, validation, and test sets for evaluating your model. Otherwise, you won't receive any score for the analytical part.

2.The dataset provided to you includes the following features:

- Disease: The name of the disease or medical condition.
- Fever: Indicates whether the patient has a fever (Yes/No).
- Cough: Indicates whether the patient has a cough (Yes/No).
- Fatigue: Indicates whether the patient experiences fatigue (Yes/No).
- Difficulty Breathing: Indicates whether the patient has difficulty breathing (Yes/No).
- Age: The age of the patient in years.
- Gender: The gender of the patient (Male/Female).
- Blood Pressure: The blood pressure level of the patient (Normal/High).
- Cholesterol Level: The cholesterol level of the patient (Normal/High).
- Outcome Variable: The outcome variable indicating the result of the diagnosis or assessment for the specific disease (Positive/Negative).

Based on the medical symptoms in the Disease column, whether the diagnosis of the disease is correct or incorrect is stated in the Outcome Variable column. Other features include gender, age, blood pressure, and cholesterol level.

First, perform the necessary preprocessing on the dataset. Then, on the available dataset:

a) Implement K-Nearest Neighbors from scratch.
b) Report the accuracy of your model using Scikit-learn functions, and explain what it means and its practical significance.
c) Plot the ROC curve and write your analysis of its performance for the KNN model results.