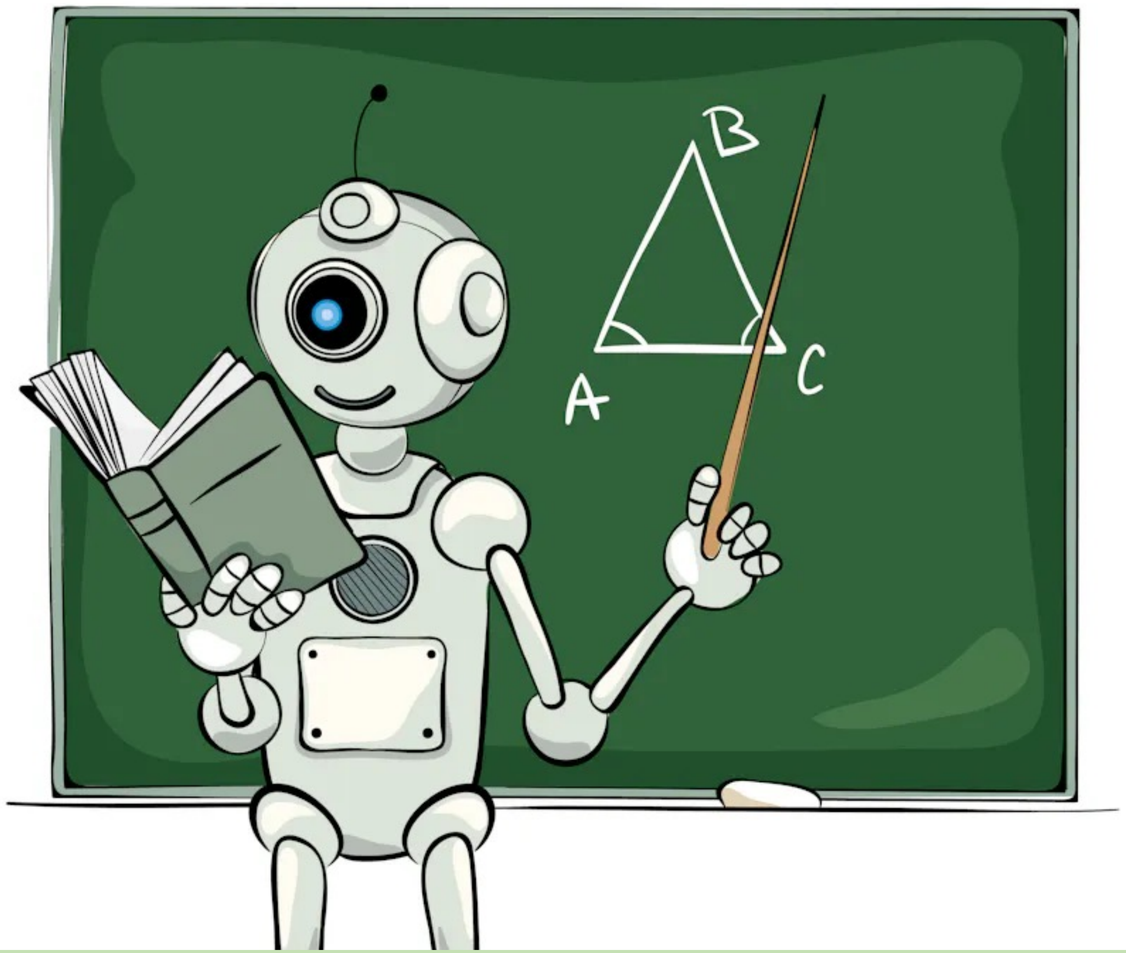
Machine Learning



Lab 10

Project

Introduction

- The project is about the Heart Disease and Stroke Prevention dataset
  - https://www.kaggle.com/datasets/themrityunjaypathak/heart-disease-and-stroke-prevention
- You can find the dataset in the folder **Project** on MS Teams

Goal

- Given information about the public health burden of Cardiovascular Diseases in the United States, we want to define an automatic procedure for **categorizing cardiovascular diseases**

Info about the dataset

- Data Characteristic
  - Multivariate
- Number of instances
  - 42640
- Number of attributes
  - 29
- Missing value
  - Yes

- The target has five categories:
  - 'Heart Failure'
  - 'Coronary Heart Disease'
  - 'Stroke'
  - 'Major Cardiovascular Disease',
  - 'Diseases of the Heart (Heart Disease)',
  - 'Acute Myocardial Infarction (Heart Attack)'

Info about the dataset

- Inside the features, there is an attribute that is highly relative to the class label (provide the same information).
- This attribute should be computed only in conjunction with the class label, so it is an attribute "not available" for unseen data.
- Using it for training the data leads to the so-called "**data leakage**" (or leakage) that happens when your training data contains information about the target, but similar data will not be available when the model is used for prediction. This leads to high performance on the training set (and possibly even the validation data), but the model will perform poorly in production.
- Pay attention: remove the "**Indicator**" attribute (and **other associate attributes**)!

How to create the data split?

Instructions

- Please divide your data into train, test, and val sets:
  - First divide the data into train_tmp and test (**80/20**)
  - Then, divide the train_tmp into train and val (**80/20**)
  - Use the seed equal to **12062024** for the splitting

Instructions

- Analyze the dataset by applying what you learned during the course, trying to achieve your goals
  - You can use any kind of analysis tool that best fits your needs

Instructions

- Analyze the dataset by applying what you learned during the course, trying to achieve your goals
  - You can use any kind of analysis tool that best fits your needs

- Produce:
  - **Python code**
  - Project **Documentation**
  - Project **Presentation**

Instructions

- The maximum score you can achieve in the project is 20
- The project discussion must last 15 minutes at most
- If the project will be approved, you will have the oral proof

Instructions

- **Data Understanding**
  - Analyze your data
- **Data Preparation**
  - Prepare your data
- **Modeling**
  - Create train/test/val sets
  - Select ML models
  - Fit your models
- **Evaluation**
  - Compare your trained models
  - Select the best one

General Info

- The project will last **One year**
  - You can deliver it whenever you are ready
    - Python code and documentation
  - Anyway, **One week** before the exam

- Official notifications will be provided during the exam period