Datasets:

Classification

1. [Health Dataset(diabetes)](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)
2. [Predicting the Future Transaction from Large and Imbalanced Banking Dataset](https://www.kaggle.com/c/santander-customer-transaction-prediction)
3. [Multiple Sclerosis Disease](https://www.kaggle.com/datasets/desalegngeb/conversion-predictors-of-cis-to-multiple-sclerosis/download?datasetVersionNumber=2)
4. [Credit risk classification](https://www.kaggle.com/datasets/ppb00x/credit-risk-customers/download?datasetVersionNumber=3)

Regression:

1. [Wine Quality](https://www.kaggle.com/datasets/rajyellow46/wine-quality/download?datasetVersionNumber=1)
2. [Sleep Efficiency Dataset](https://www.kaggle.com/datasets/equilibriumm/sleep-efficiency/download?datasetVersionNumber=3)
3. [A Waiter's Tips](https://www.kaggle.com/datasets/jsphyg/tipping/download?datasetVersionNumber=1)
4. [Flight Delay from January 2017 - July 2022](https://www.kaggle.com/datasets/jawadkhattak/us-flight-delay-from-january-2017-july-2022/download?datasetVersionNumber=2)

Way of work:

1. Every student has to solve 2 projects
2. Every student will have 10 mins presentation:
3. Every student should follow this protocol:
   * Data loading and preprocessing:
     + Identify features and labels
     + Features:
       - Encoding data Categorical to numerical/ordinal (if needed)
       - Identifying/correcting errors (if needed)
       - Filling/removing missing values (if needed)
       - Removing outliers (if needed)
       - Transformation (Standardization/Normalization)
       - Visualize the features
       - Analyze relationships between features (correlation matrix)
       - Rank the features using PCA and Random forest
       - Use certain features not all the features (up to you to choose the features)
     + Labels
       - Calculate Central/Dispersion measures
       - Check the proportion of the target values in case of classification
       - Rebalance the dataset using resembling techniques (if needed)
     + Split the data into train and test (80-20)
     + Use different base models
     + Use grid search/random search
     + Use ensembling techniques to see if there is any improvement in the overall accuracy.
     + Plot the test set using only two principal components
     + Use the concept of pipeline for your machine learning modeling
     + Communicate your results:
       - Explain the effect of each feature on the modeling process.
       - Report the best model.