Assignment

Weight 20%

The final submission of the assignment must include:

- The Jupyter Notebook (well-commented) AND
- A Word/PDF document explaining different phases and interpreting the results you got in each step. You may include figures from your Python code.

Data Preparation

Initial Analysis (check course recording on week 11 for more details)

- Run univariate, bivariate, multivariate analysis (you can use propoer plots and comment on the plots),
- Target/Class variable (identify it and use proper plot to see if the dataset is balanced or not, comment on the plot),
- Assign correct data types, appropriate attribute names,
- Identify numerical, categorical attributes,
- Identify and take care of duplicates, errors,
- Impute missing values by e.g., mean, mode, median,
- Outliers using e.g., boxplots.

At the end of this phase, you get Consistent Data Ready for Further Analysis!

Exploratory Data Analysis (EDA)

- Normalize your dataset

Dimensionality Reduction

Discard attributes:

- With many missing values,
- Zero/low variance,
- Highly correlated (remove the one with more missing values), comment propoerly

Principle Component Analysis if still lots of attributes to analyze (PCA)

In this phase you can use results of bivariate analysis.

At the end of this phase, you get data ready for Machine Learning Modelling

Experimental Design

- Split dataset into training and test datasets (70%, 30% respectively),
- If you have an imbalanced dataset, you must take care of it (under-sampling, over-sampling)

Modelling

Develop a machine learning algorithm based on the research problem using the training set:

- Classification vs. regression, decision tree or regression tree, logistic regression,
- Linear vs. non-linear dataset.

Evaluation

Classification

- Confusion matrix,
- Accuracy, recall, precision

Regression

- R squared,
- Root Mean Square Error (RMSE).