Data Preparation Course

IEEE CIS - Eng. Kasit 101

IEEE Computational Intelligence Society
University of Jordan Chapter

Course Syllabus

1. **Introduction to Data Preparation**
2. **Reading & Loading Data**
   - EDA Process
   - Intro. Numpy, Pandas, matplotlib, and Seabon
   - Correlation analysis & Statistical Analysis & Skewness
3. **Data Cleaning**
   - Handling Missing Data
   - Categorical Data
   - Outliers Analysis
4. **Feature Engineering**
   - Feature Extraction
   - Feature Selection
   - PCA
5. **Splitting the Data into**
   - Training
   - Validation
   - Testing
6. **Tips & Tricks**
7. **Final Project**

What is Data Preparation

- Data Preparation is the process of cleaning, transforming, and converting raw data to make it useful for analysis.
- Data Preparation is a crucial step in the data science workflow.

Why is Data Preparation Necessary

- Data is rarely clean and ready for analysis out of the box.
- Data often contains missing values, duplicates, and outliers that can impact the accuracy of our analysis.
- Data Preparation ensures that we have clean and reliable data for analysis.

Overview of the Data Preparation Process

The Data Preparation Process consists of the following steps:

- Reading & Loading Data
- Exploratory Data Analysis (EDA)
- Data Cleaning
- Feature Engineering
- Splitting the Data

Reading & Loading Data

a. EDA process

b. Intro. Numpy, Pandas, Matplotlib, and Seaborn

c. Extract meaningful information from the data

d. Correlation analysis and feature importance

e. Statistical analysis and summary

f. Skewness

EDA Process (Exploratory Data Analysis)

- Once we have loaded the data, we need to perform Exploratory Data Analysis (EDA) to gain insights into the data.
- EDA involves visualizing and analyzing the data to understand its properties and relationships.

Introduction to Numpy, Pandas, Matplotlib

- **Numpy** is a Python library for performing numerical computations on arrays and matrices.
- **Pandas** is a Python library for data manipulation and analysis.
- **Matplotlib** is a Python library for creating visualizations.

Let’s practice with some **CODE**