Introductory Notes

Machine Learning Process

- **Data Pre-Processing**
  - Import the data
  - Clean the Data
  - Split into training and test sets
- **Modelling**
  - Build the model
  - Train the model
  - Make predictions
- **Evaluation**
  - Calculate performance metrics
  - Make a verdict

Training and Testing Data sets

- Usually you would take 20% of the total dataset and allocate it for testing the model. Meanwhile the remaining 80% would be used for training a machine learning model. Once model is created, we use the 20% data set to test the models by making predictions and use them to evaluate the model.

Feature Scaling

- Feature Scaling is only applied to **Columns**
- Two types of Feature Scaling:
  - Normalization
    - Process of taking the Minimum of the column and subtracting it from every single value in that column and than dividing by the difference of the Maximum and Minimum.
    - $X' = (X - X_{Min})$ / (X_{MAX} - X_{MIN})
    - Every single value in a column is adjusted this way, resulting in a new column where values **[0;1]**.
  - Standardization
    - Process is taking the average μ of all the values in the column, and subtracting each value by the average and dividing it by the standard deviation σ

    - $X' = (X - \mu) / \sigma$
    - Every single value in a column is adjusted this way, resulting in a new column where values are **[-3;+3]**.
- Reason for feature scaling? Answer is simple, in a dataset you have multiple columns of numbers and data representing different contexts. For algorithms to accurately perform, we need to make sure all data is scaled, hence we apply **Normalization** and **Standardization** methods to scale data.

Next Steps -> Data Preprocessing

Data Preprocessing