# Machine Learning In Python

## Subject :Regression in Supervised Learning

Lecturer : Reza Akbari Movahed

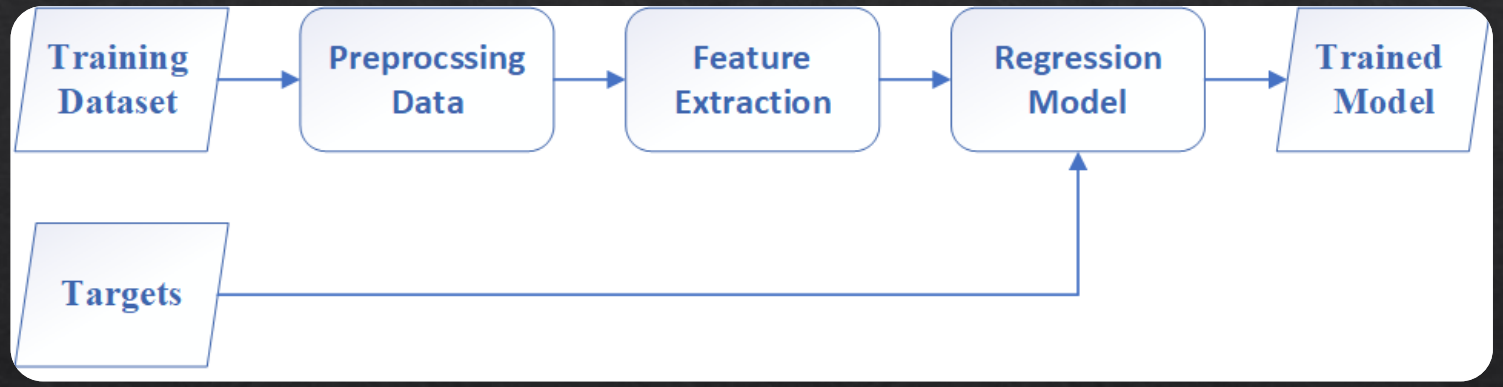
Hamedan University of Technology

Spring 2020

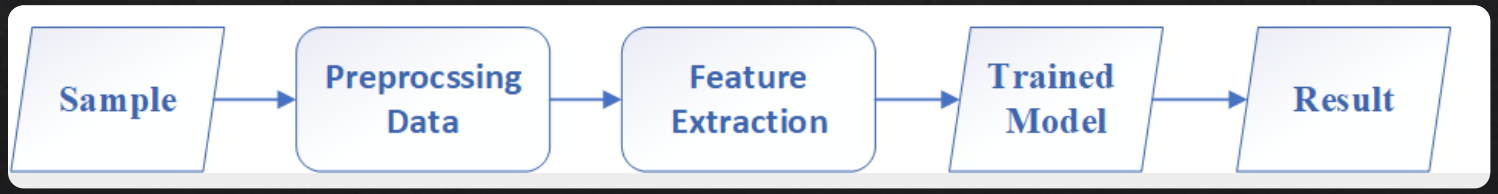
# Regression Framework

Regression In Supervised Learning Framework

**Training Phase**




**Testing Phase**



# Regression in Supervised Learning

## Evaluation Metrics for Regression Models

$$e_t = y_t - \hat{y}_t$$

$y_t$ : *real target value*

$\hat{y}_t$: *estimated target value*

$e_t$ : *error*

| | |
|---|---|
| Mean squared error | $\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$ |
| Root mean squared error | $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$ |
| Mean absolute error | $\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$ |
| Mean absolute percentage error | $\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$ |

# Regression in Supervised Learning

## Common Regression Datasets

### 7.2.3. Diabetes dataset

Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

**Data Set Characteristics:**

| | |
|---|---|
| **Number of Instances:** | 442 |
| **Number of Attributes:** | First 10 columns are numeric predictive values |
| **Target:** | Column 11 is a quantitative measure of disease progression one year after baseline |
| **Attribute Information:** | • Age<br>• Sex<br>• Body mass index<br>• Average blood pressure<br>• S1<br>• S2<br>• S3<br>• S4<br>• S5<br>• S6 |

# Regression in Supervised Learning

## Common Regression Datasets

### 7.2.1. Boston house prices dataset

**Data Set Characteristics:**

| | |
|---|---|
| **Number of Instances:** | 506 |
| **Number of Attributes:** | 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target. |
| **Attribute Information (in order):** | • CRIM per capita crime rate by town<br>• ZN proportion of residential land zoned for lots over 25,000 sq.ft.<br>• INDUS proportion of non-retail business acres per town<br>• CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)<br>• NOX nitric oxides concentration (parts per 10 million)<br>• RM average number of rooms per dwelling<br>• AGE proportion of owner-occupied units built prior to 1940<br>• DIS weighted distances to five Boston employment centres<br>• RAD index of accessibility to radial highways<br>• TAX full-value property-tax rate per $10,000<br>• PTRATIO pupil-teacher ratio by town<br>• B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town<br>• LSTAT % lower status of the population<br>• MEDV Median value of owner-occupied homes in $1000's |