**MACHINE LEARNING**

Machine Learning is the subfield of computer science that gives “**computers the ability to learn without being explicitly programmed.”**

Examples: Netflix, Prime Videos, Bank Load, Companies

**Major Methods learning techniques**

* **Regression / Estimation**

Predicting continuous values

* **Classification**

Predicting the items class/category of a case

* **Clustering**

Finding the structure of data: summarization

* **Association**

Associating frequent co-occurring items/events

* **Anomaly detection**

Discovering abnormal and unusual cases

* **Sequence mining**

Predicting next events: click-stream (Markov Model, HMM)

* **Dimension Reduction**

Reducing the size of data (PCA)

* **Recommendation system**

Recommending items

**Python Libraries for Machine Learning**

* Pandas
* Scikit-learn
* NumPy
* SciPy
* Matplotlib

**More about scikit-learn**

* Free software Machine learning library
* Classification, Regression and Clustering algorithms
* Works with NumPy and SciPy
* Great documentation
* Easy to implement

**Supervised vs Unsupervised learning**

* **Supervised Learning**
* Classification - Classifies labelled data.
* Regression – Predicts trends using previous labelled data.
* Has more evaluation methods than unsupervised learning.
* Controlled environment.
* **Unsupervised Learning**
* Clustering – Finds patterns and groupings from unlabelled data.
* Has fewer evaluation methods than supervised learning.
* Less controlled environment.

**Regression**

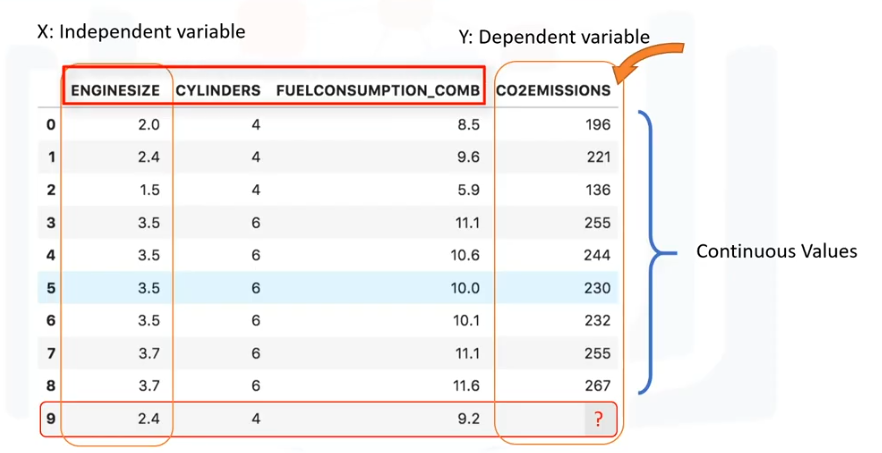
Regression is the process of predicting a continuous value. In regression, there are two types of variables, a dependent variable, and one or more independent variables.

The **dependent variable** can be seen as the state target, or final goal we study and try to predict and the **independent variables** also known as explanatory variables can be seen as the causes of those states.

* **Simple Regression - one independent variable is used to estimate a dependent variable.**
* Simple Linear Regression
* Simple Non-linear Regression
* **Multiple Regression – more than one independent variable is used to estimate a dependent variable.**
* Multiple linear Regression
* Multiple Non-linear Regression

**Regression algorithms**

* Ordinal regression
* Poisson Regression
* Fast Forest quantile regression
* Linear, Polynomial, Lasso, Stepwise, Ridge regression
* Bayesian linear regression
* Neural Network regression
* Decision forest regression
* Boosted decision tree regression
* KNN (k- nearest neighbors)

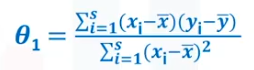
**Linear Regression**

* **Simple Linear Regression**
* Predict **co2emission** vs **EngineSize** of all cars
* Independent Variable (x): EngineSize
* Dependent Variable (y): co2emission
* **Multiple Linear Regression**
* Predict **co2emission** vs **EngineSize** and **Cylinders** of all cars
* Independent Variable (x): EngineSize, Cylinders, etc
* Dependent Variable (y): co2emission

**Prediction of linear regression**

****

**Estimate the parameters**

****



For this example

θ1 = 39 θ0 = 125.75 EngineSize = 2.4

Co2Emission = θ1 + θ0 EngineSize

Co2Emission = 125 + 39 \* 2.4

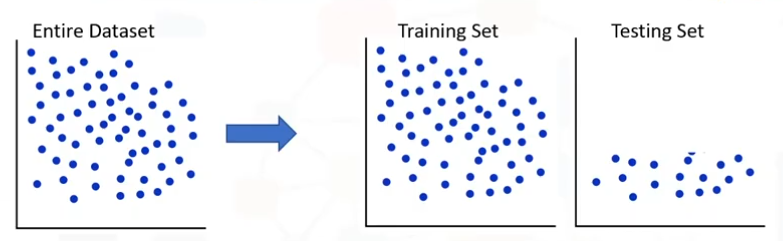
Co2Emission = 218.6

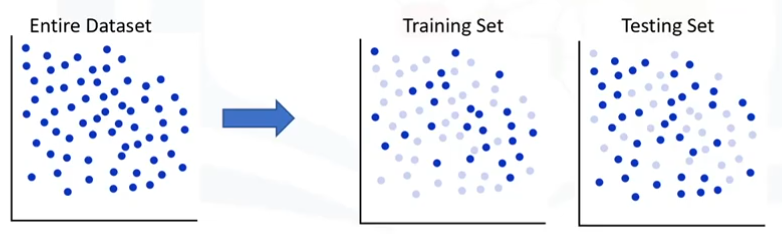
**Pros of linear regression**

* Very Fast
* No parameter tuning
* Easy to understand, and highly interpretable

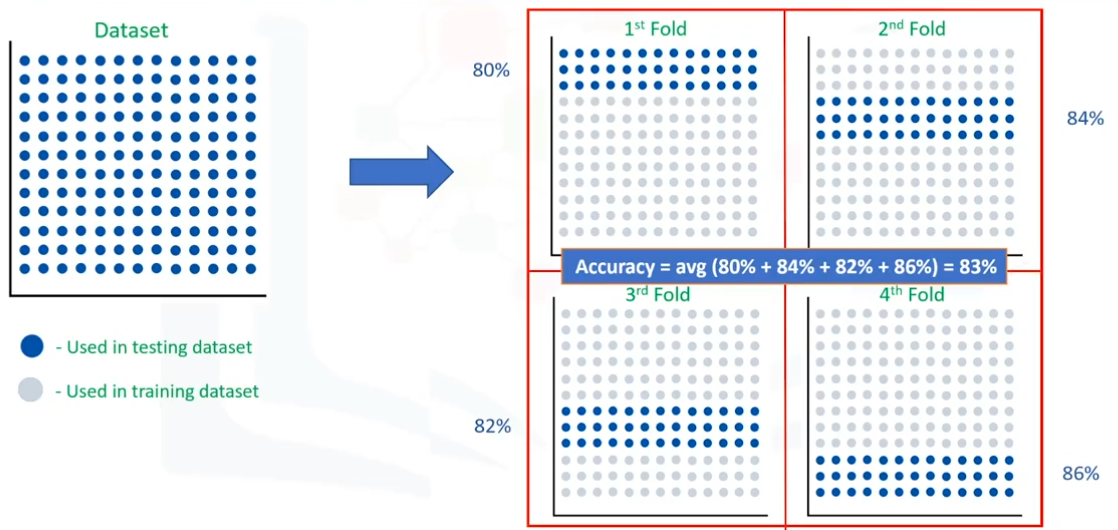
**Model Evaluation Approaches**

* **Train and Test on the Same Dataset**
* Test-set is a portion of the train-set
* High “training accuracy”
* Low “out-of-sample accuracy”

****

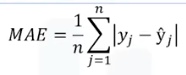
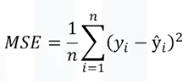
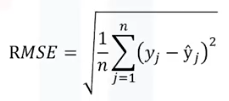
* **Train/Teat Split**
* Mutually exclusive
* More accurate evaluation on out-of-sample accuracy
* ****Highly dependent on which datasets the data is trained and tested

**K- fold cross – validation**



**Regression Accuracy**

Regression accuracy can be calculated by getting the error of accuracy.

* Mean Absolute error
* Mean Square error
* Root Mean Square error
* Relative Absolute error
* Relative Squared errors