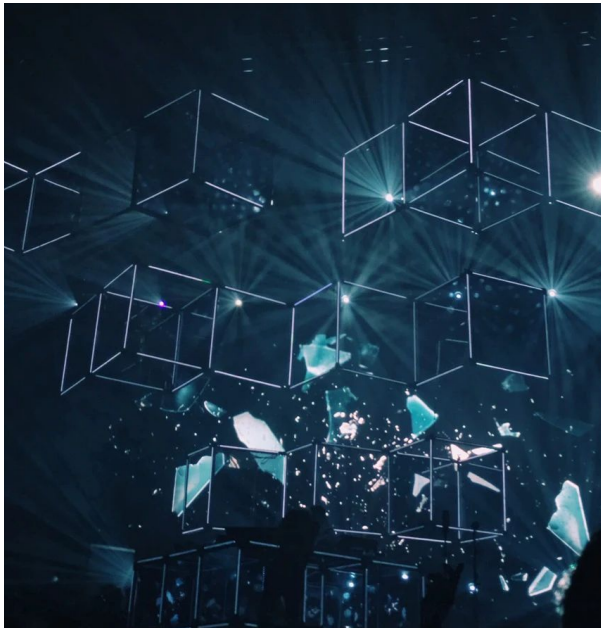
Supervised Machine learning - Linear Regression

DATA ANALYTICS | IRONHACK

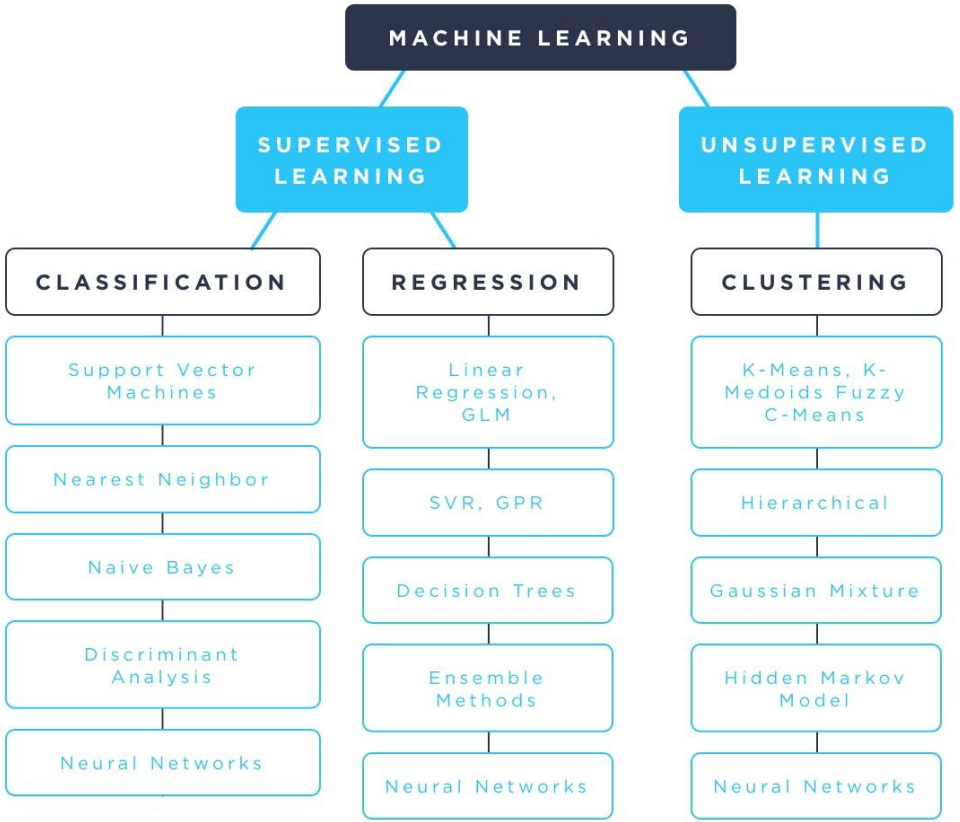


Credit: Unsplash

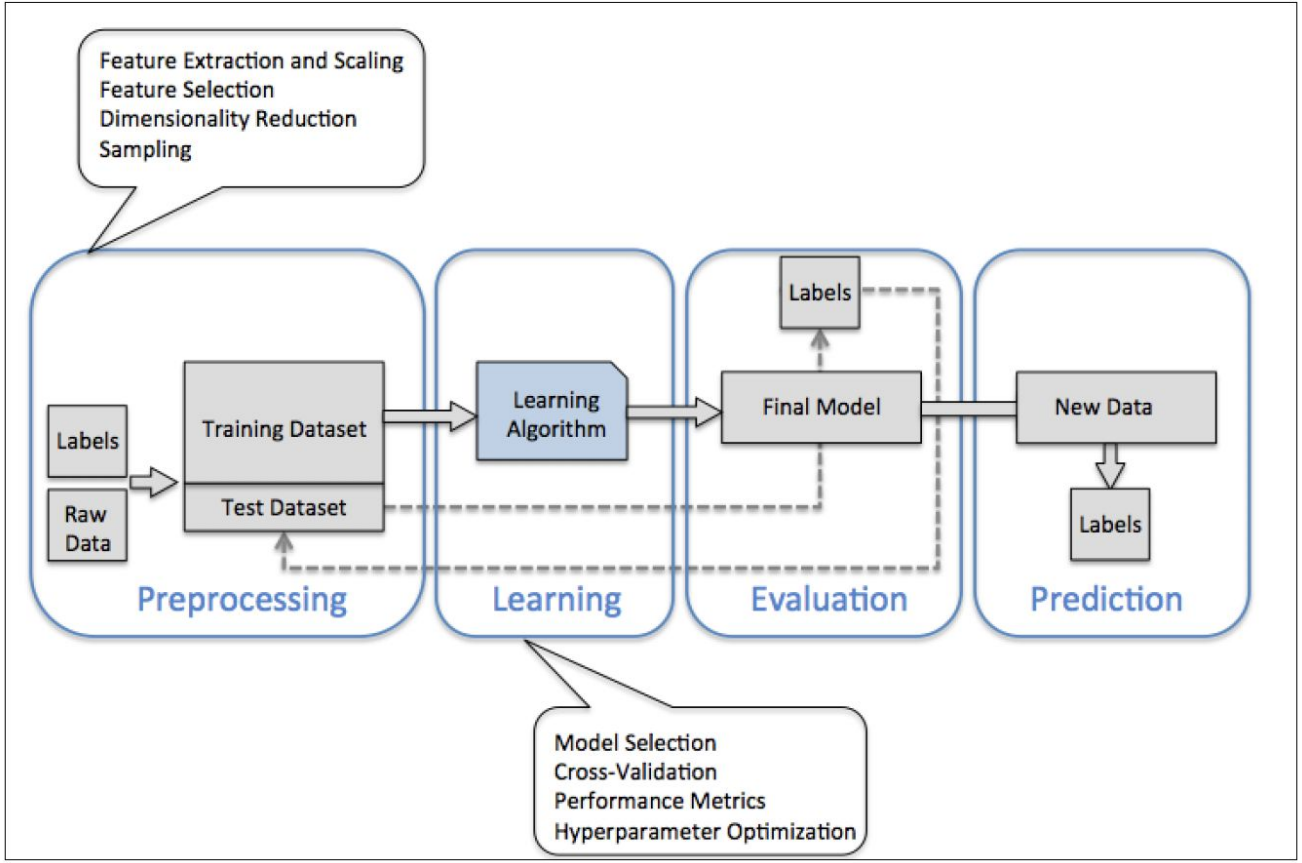
Aims of the day:

1. What is machine learning
2. Review case study aims
3. Linear regression model
    - How to apply it
    - How to evaluate it
4. Pre processing data
5. Some of the challenges

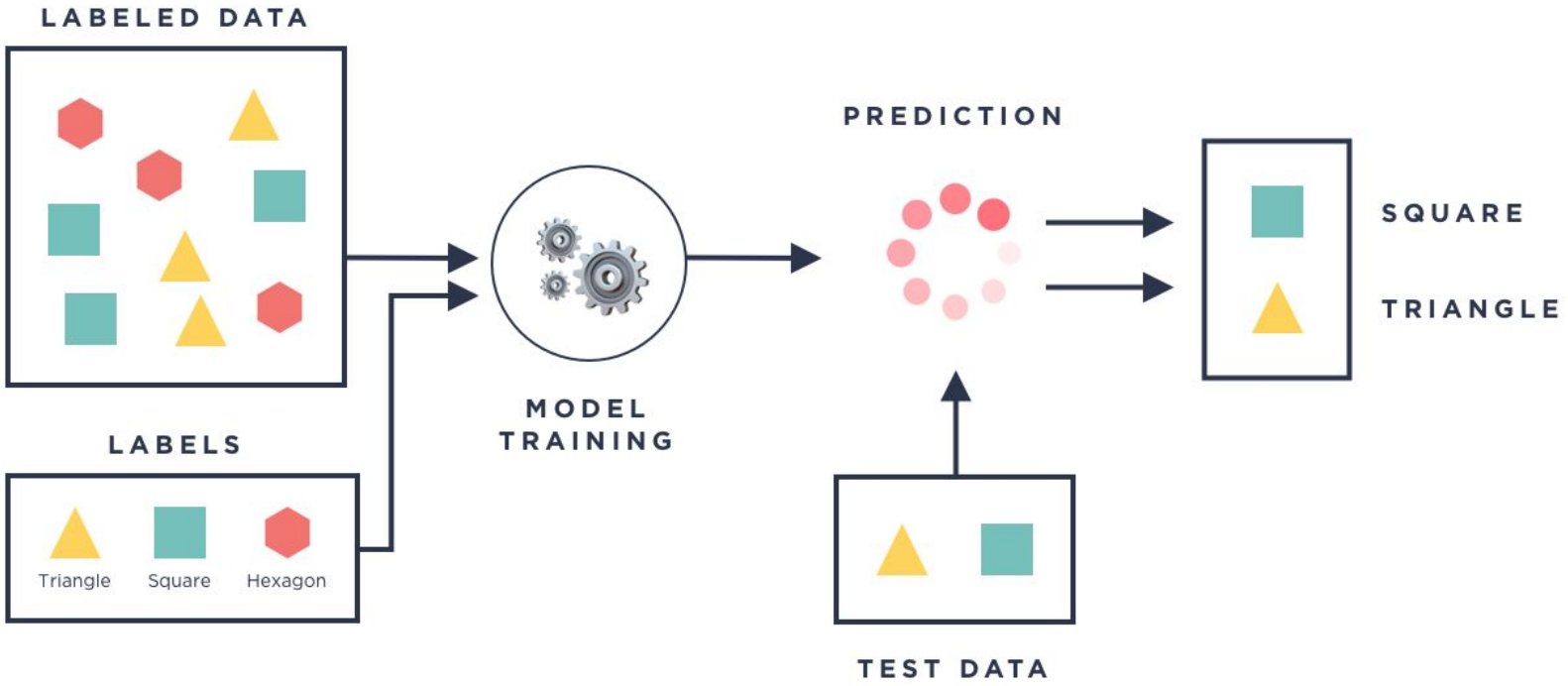
Overview:


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Nearest Neighbor
Naive Bayes
Discriminant Analysis
Neural Networks
Linear Regression, GLM
SVR, GPR
Decision Trees
Ensemble Methods
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks

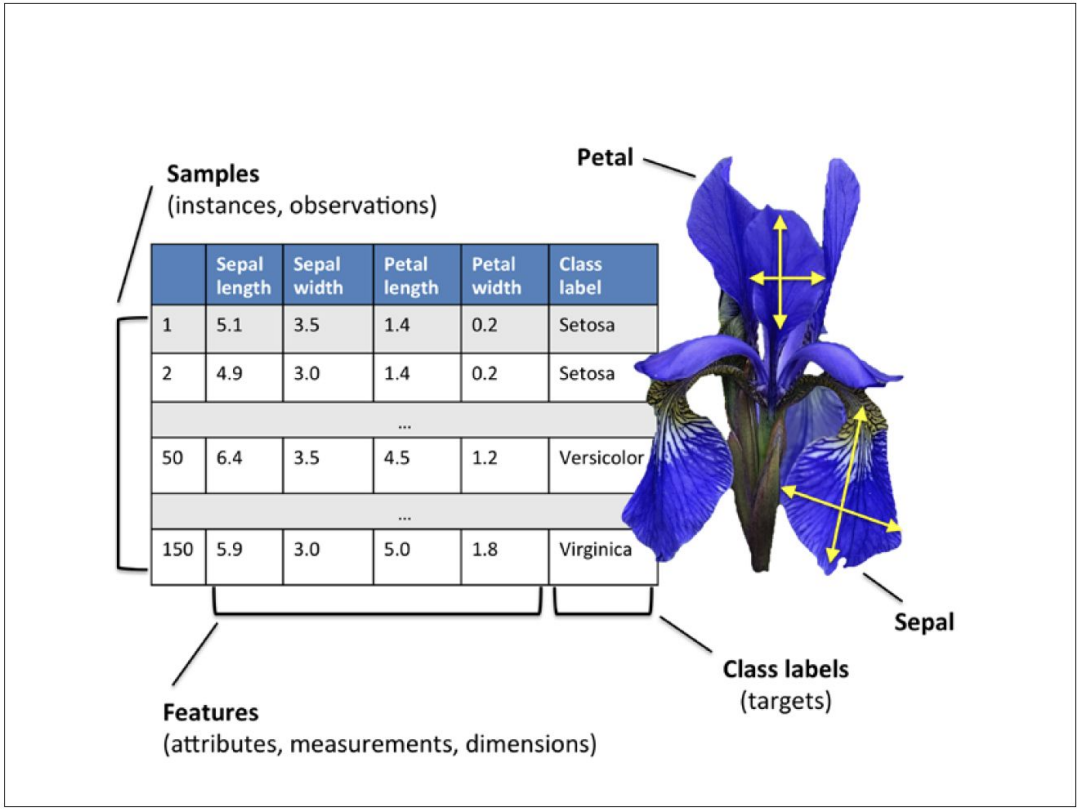
Process:


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Learning Algorithm
Labels
Final Model
New Data
Labels
Preprocessing
Learning
Evaluation
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Supervised Learning.

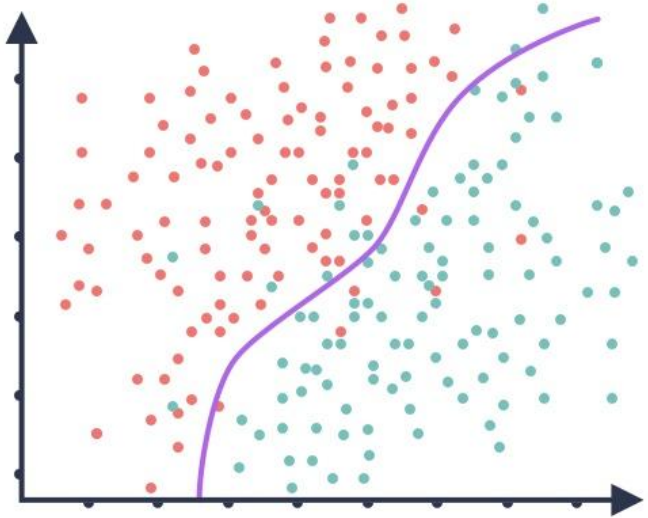

LABELED DATA
LABELS
Triangle
Square
Hexagon
MODEL TRAINING
PREDICTION
TEST DATA
SQUARE
TRIANGLE

ML Terminology:

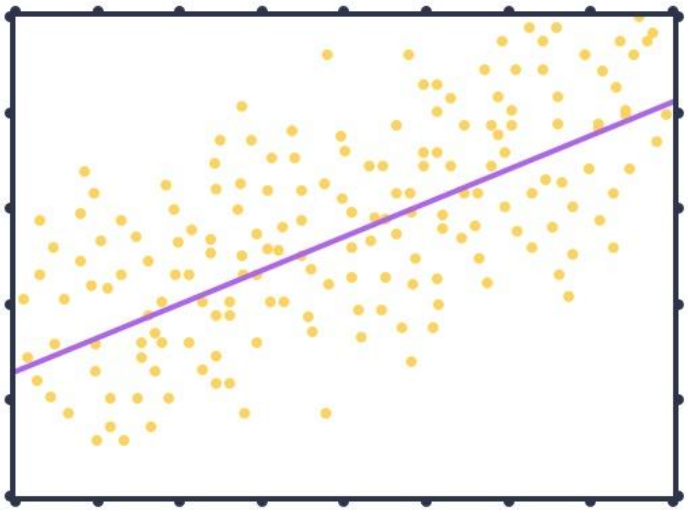


	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

What is the difference between…



CLASSIFICATION



REGRESSION

Terminology

Regression Problem:

Predicting an amount

Target_D is the dependent variable.

The features are the independent variables.

Credit: Unsplash

Features

Labels

HV1	IC1	IC2	IC3	IC4	IC5	AVGGIFT	TARGET_D
2346	420	446	468	503	14552	15.5	21
497	350	364	357	384	11696	3.08	3
1229	469	502	507	544	17313	7.5	20
325	148	181	171	209	6334	6.7	5
768	174	201	220	249	7802	8.78571429	10
557	211	188	221	205	5550	13	16
2145	474	492	522	554	18340	11.5714286	15
2184	351	376	394	419	16480	12.5	20
1442	369	394	445	488	26462	7.84615385	10
1708	437	586	551	684	29098	9.76923077	20
1054	584	644	652	726	26074	13.5384615	20
1062	486	550	555	584	17908	15.3333333	20
849	457	508	470	519	16386	12.8	25
213	222	273	283	329	12227	5.125	5
574	289	318	315	363	11250	3.55555556	4
2506	449	455	501	517	16302	8.875	50
622	347	378	401	416	15808	15	25
764	272	361	346	424	16257	7.91304348	15
681	335	398	356	419	14011	30.75	51

Terminology

Classification Problem:

Predicting target variable that are labels

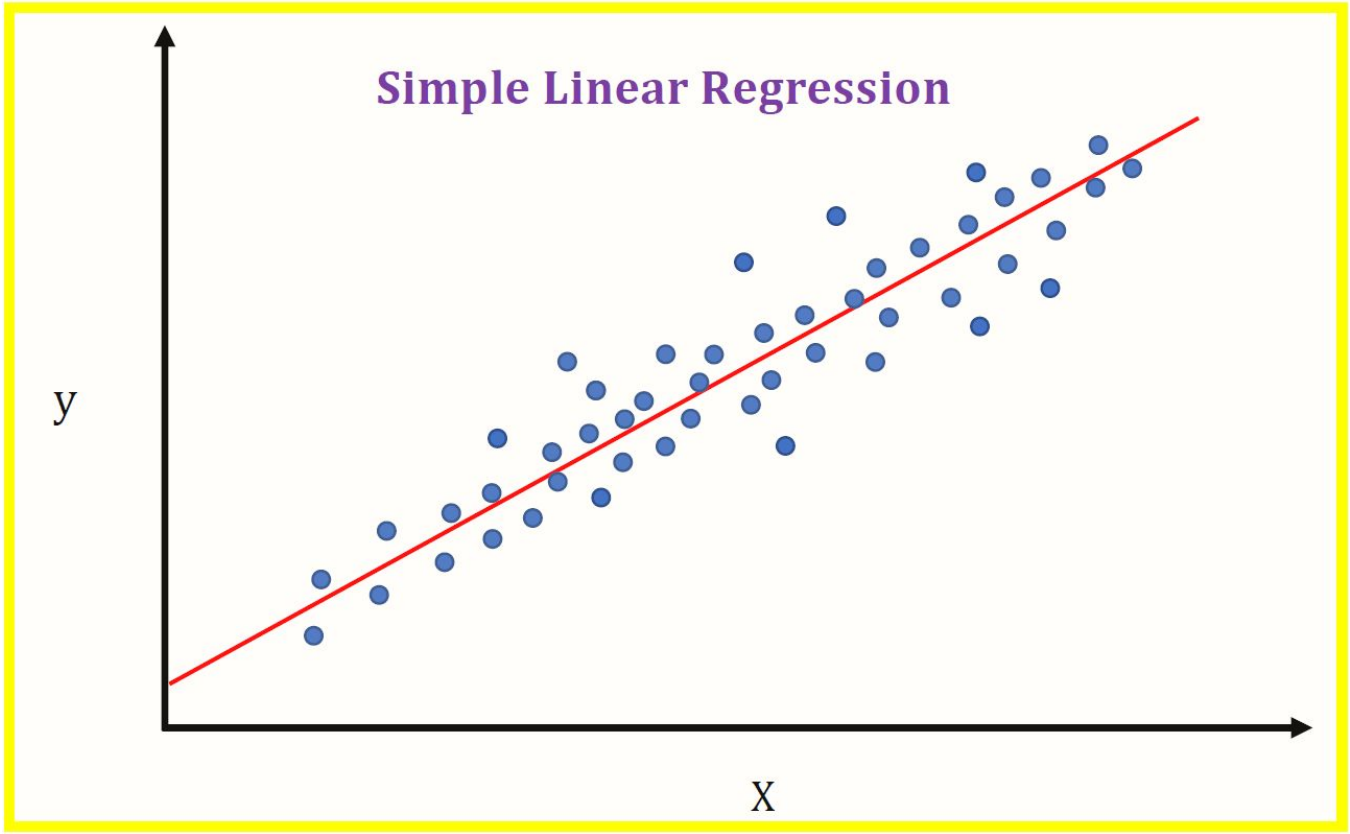
... in this case a binary (A or B, True or False, Yes or No, 1 or 0)

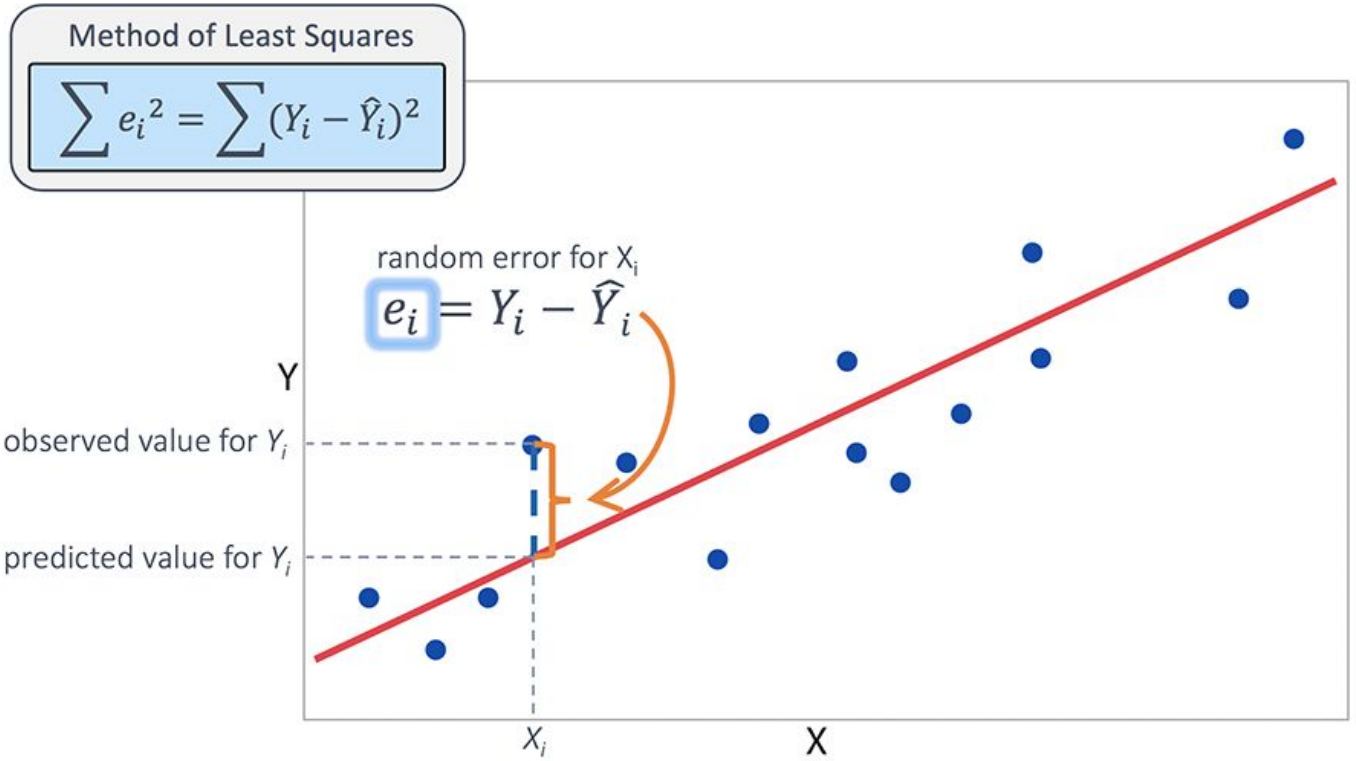
Credit: Unsplash

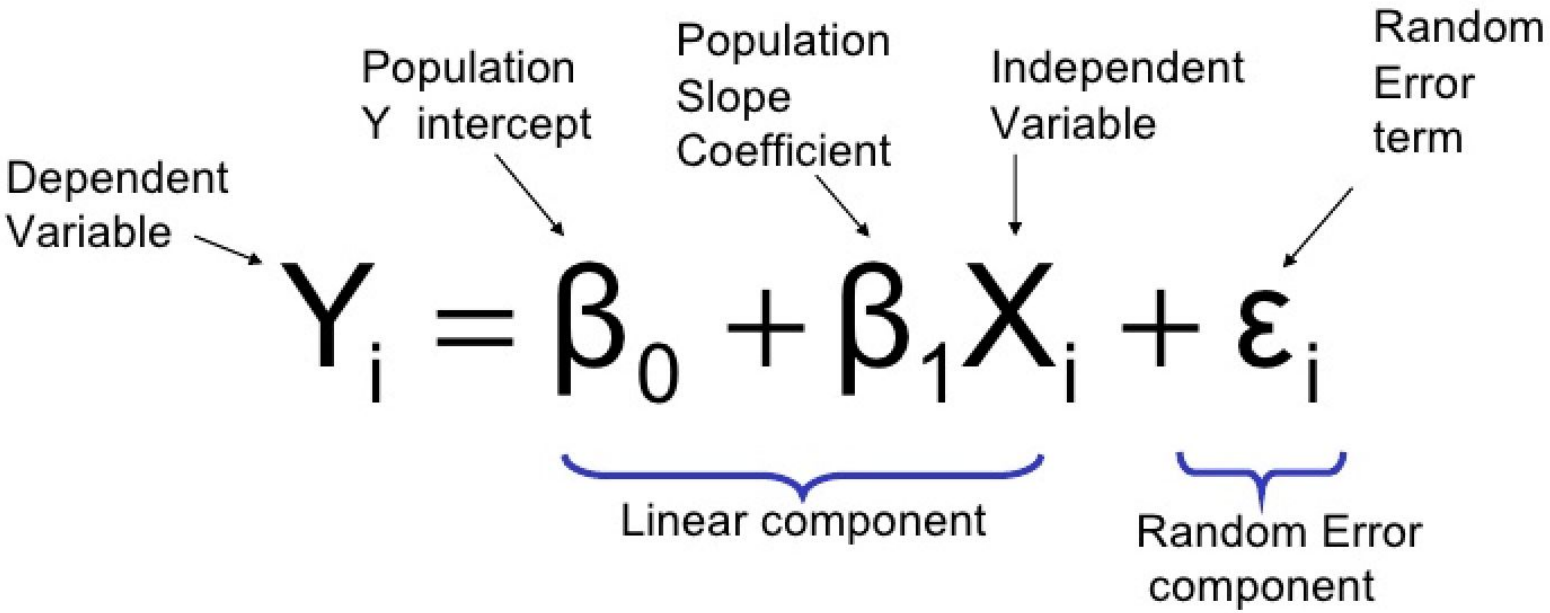
Features

Labels

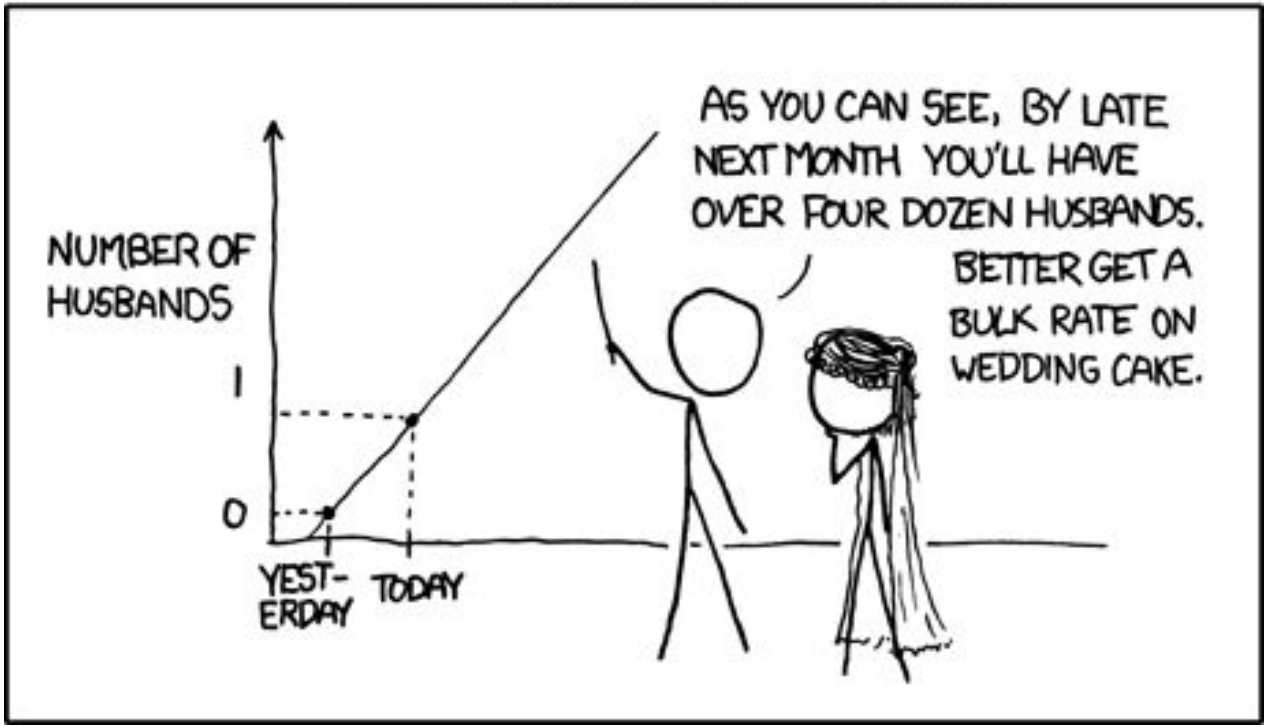
	loan_id	account_id	date	amount	duration	payments	status
0	5314	1787	930705	96396	12	8033.0	B
1	5316	1801	930711	165960	36	4610.0	A
2	6863	9188	930728	127080	60	2118.0	A
3	5325	1843	930803	105804	36	2939.0	A
4	7240	11013	930906	274740	60	4579.0	A
5	6687	8261	930913	87840	24	3660.0	A
6	7284	11265	930915	52788	12	4399.0	A
7	6111	5428	930924	174744	24	7281.0	B
8	7235	10973	931013	154416	48	3217.0	A
9	5997	4894	931104	117024	24	4876.0	A
10	7121	10364	931110	21924	36	609.0	A
11	6077	5270	931122	79608	24	3317.0	A
12	6228	6034	931201	464520	60	7742.0	B
13	6356	6701	931208	95400	36	2650.0	A
14	5523	2705	931208	93888	36	2608.0	A
15	6456	7123	931209	47016	12	3918.0	A

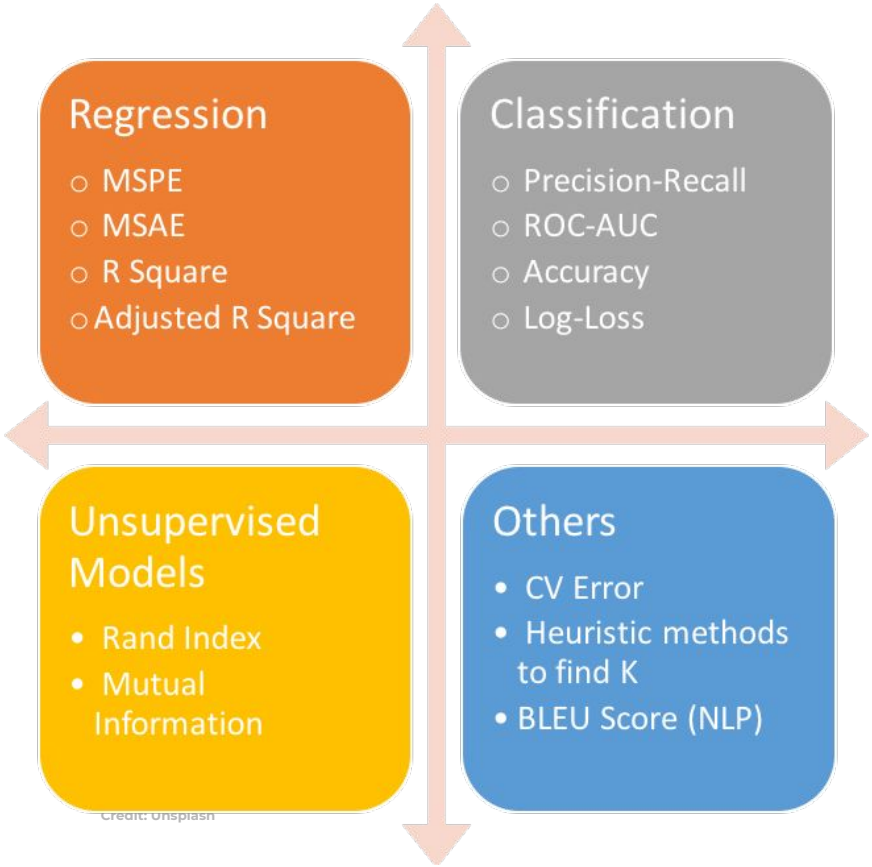

Simple Linear Regression
y
X


Method of Least Squares
$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$
random error for X_i
$e_i = Y_i - \hat{Y}_i$
Y
observed value for Y_i
predicted value for Y_i
X_i
X

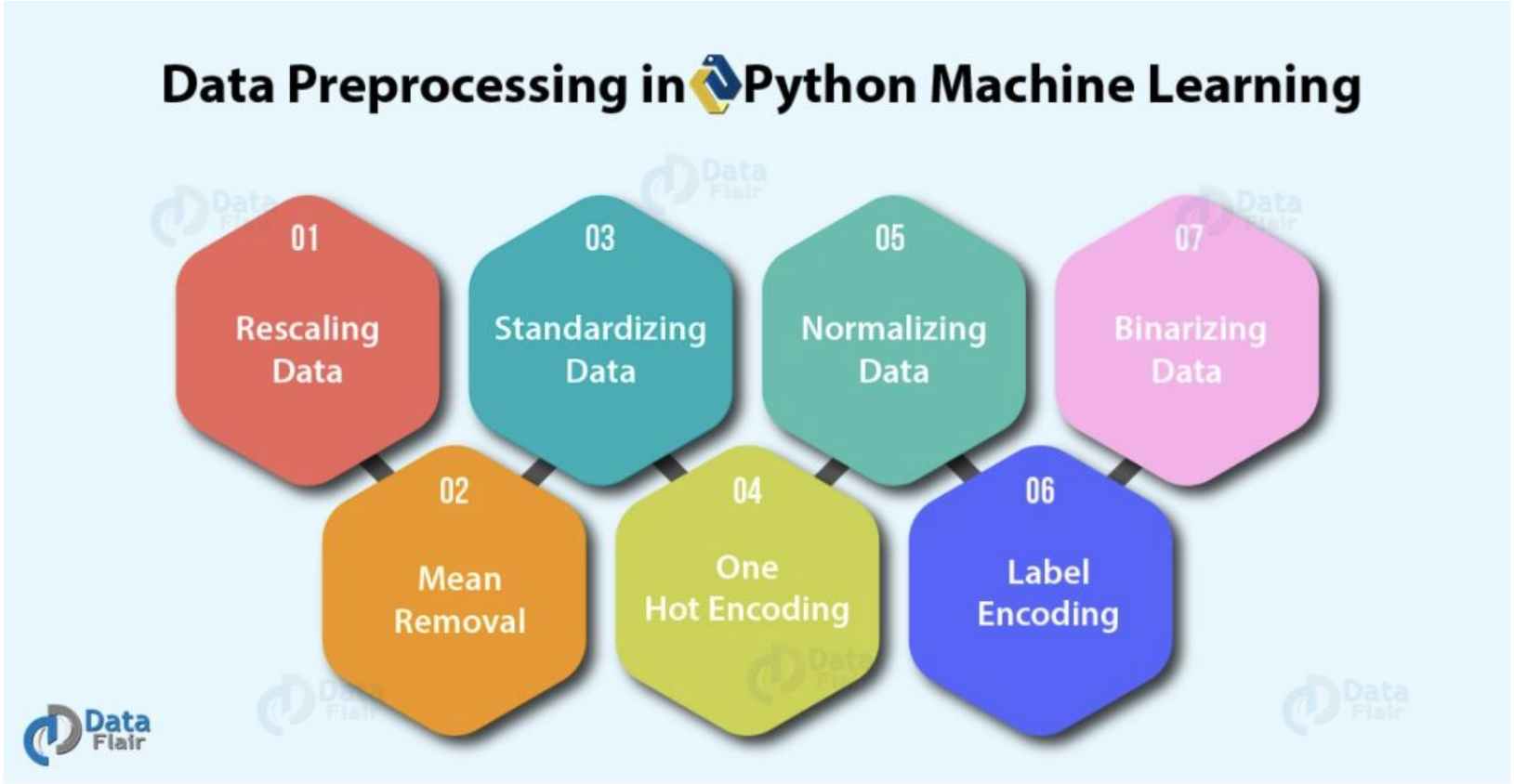

Dependent Variable
Population Y intercept
Population Slope Coefficient
Independent Variable
Random Error term
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Linear component
Random Error component

MY HOBBY: EXTRAPOLATING





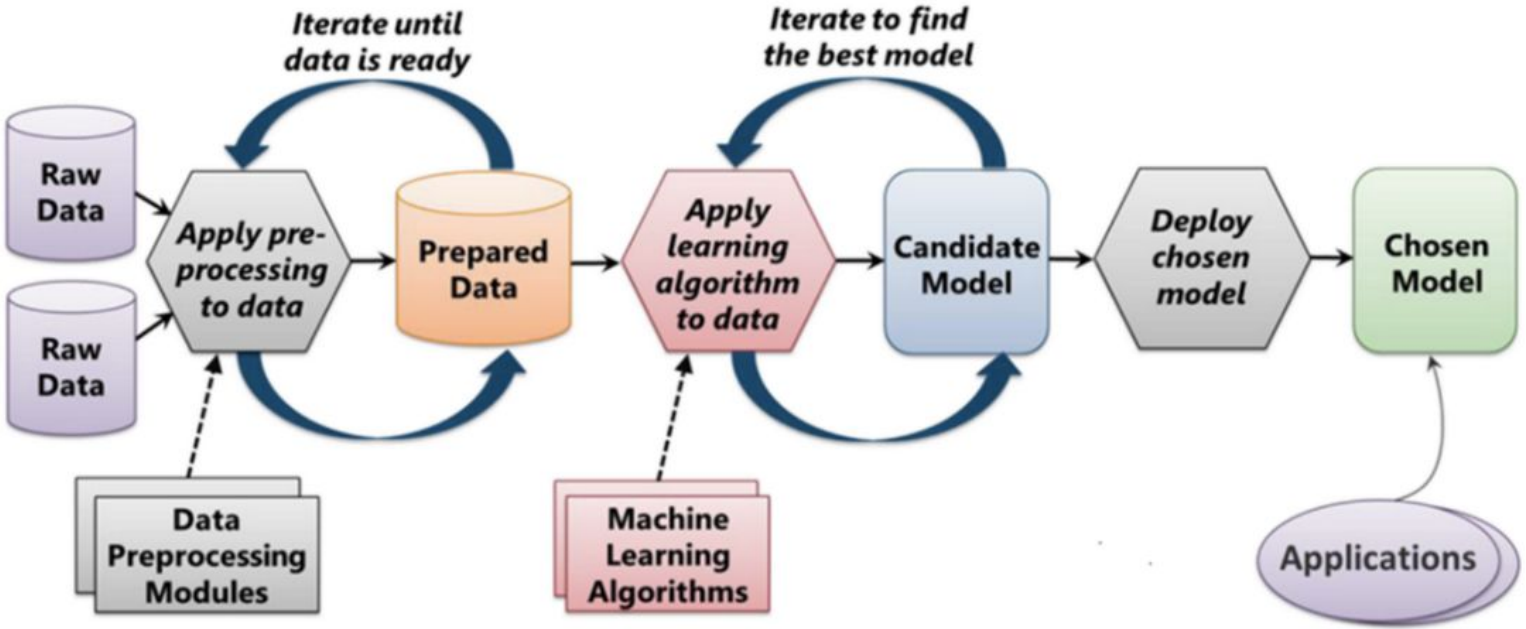
Credit: Unsplash


Data Preprocessing in Python Machine Learning
01
Rescaling Data
02
Mean Removal
03
Standardizing Data
04
One Hot Encoding
05
Normalizing Data
06
Label Encoding
07
Binarizing Data
Data Flair

Challenges of Machine Learning

- Not enough training data
- Training data is non representative
- Data is of poor quality
- Irrelevant features available (garbage in, garbage out)
- Overfitting to the training data
- All models are wrong; some are useful
- No free lunch theorem - we could be using the wrong model, we may need to find the best one through evaluation

The Machine Learning Process



From "Introduction to Microsoft Azure" by David Chappell

REGRESSION AND P-VALUES

- When we train a linear model, we can obtain a p-value associated to each coefficient of the linear model.
- What they are?
- The **Ho hypothesis in regression is that each coefficient in the linear regression is null**. In other words, the corresponding variable has no impact.
- Therefore, if we get a p-value < 0.05 we reject the null hypothesis, and therefore the associated variable has an impact on the model.

THANKS !