Machine Learning Workshop Part I

Dr Sara Soltaninejad

WiSER (Women in Science, Engineering & Research)

AltaML

Who Am I?

- Machine Learning Engineer, AltaML, 2020-Now
- PhD, Computing Science, UofA

My Support Team:

- Navaneeth: Lead Machine Learning developer AltaML, 2019-Now
- Graham: Lead Machine Learning developer AltaML, 2018-Now
- Mark: Vice President, People

ML Workshop Outline

Part I

- General Concepts of Machine Learning
- ML standard process Steps from data preparation to evaluation
- Hands-On

13 May 2021

6 May 2021

Part II

- Machine Learning Algorithms
- Hands-On

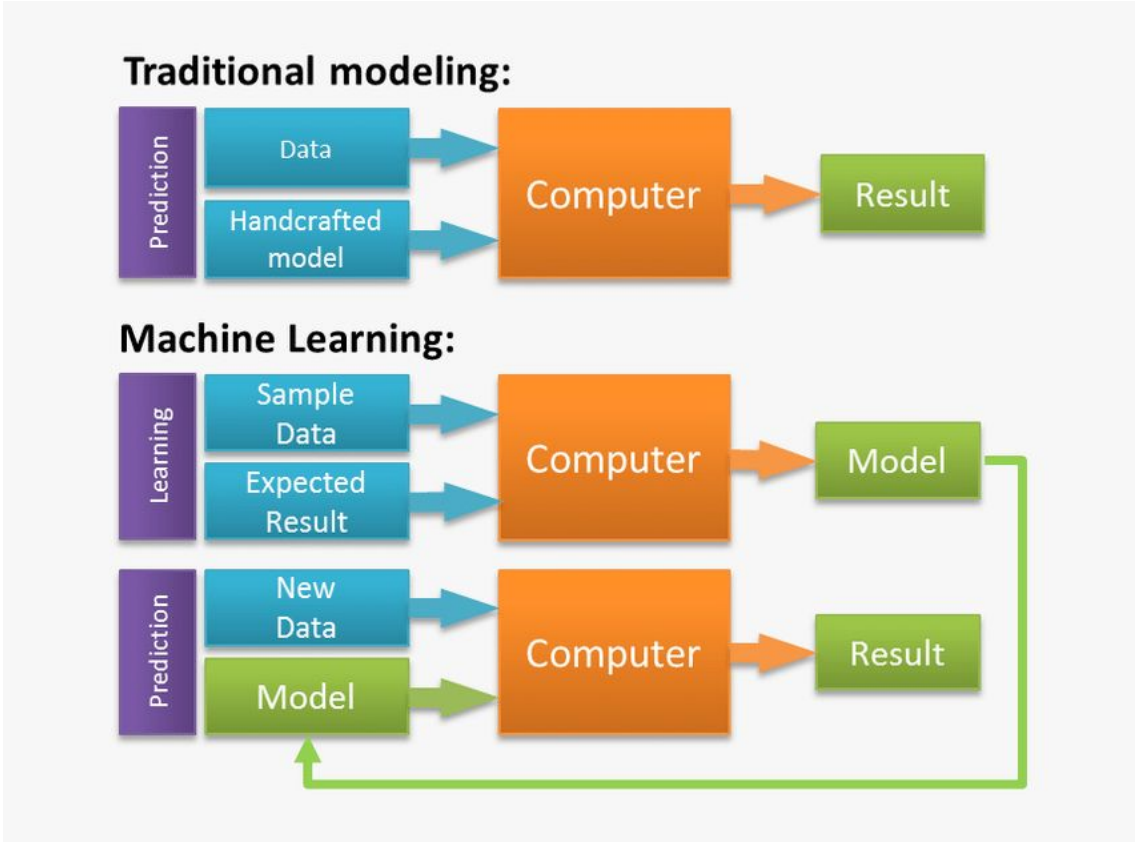
Part I Agenda

- **WiSER Introduction (3-5 min)**
- **General Introduction of the team and the AltaML company (3-5 min)**
- **Machine Learning Theoretical Concepts (1h 15min)**
  - Machine Learning General Concepts
  - Machine learning Data Preprocessing
  - Machine Learning Modelling and Evaluation
- **Break (5min)**
- **Hands-On (20-25 min)**
- **Q&A (5 min)**
- **WiSER Closing (2-3 min)**

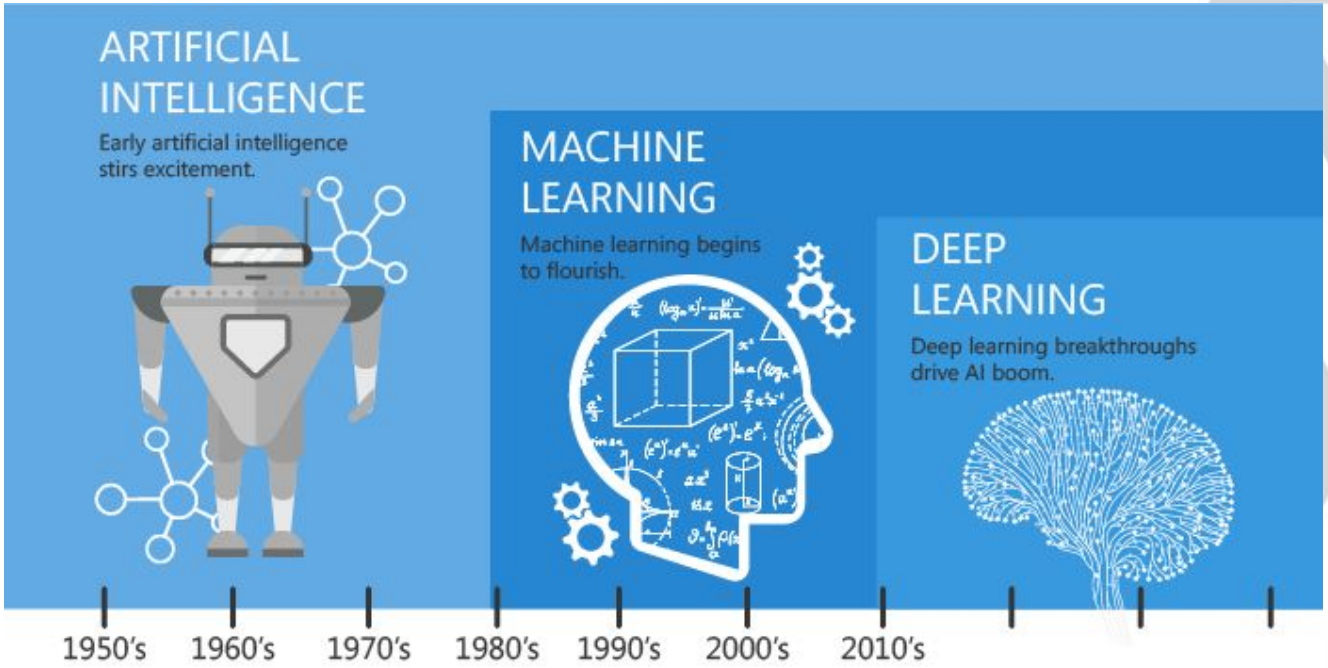
Machine Learning

Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions

ML vs Traditional Programming



AI History

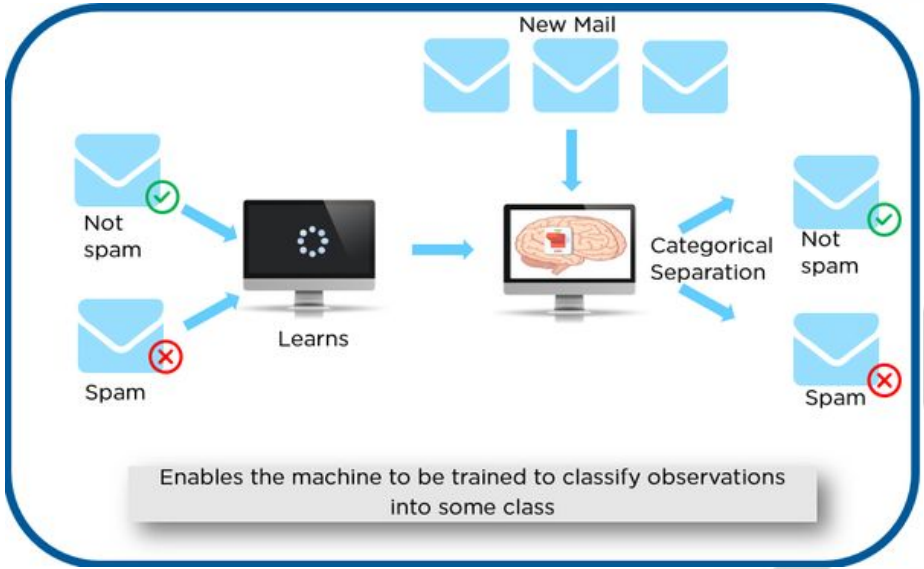


Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

ML Algorithms Types

Supervised ML Algorithm

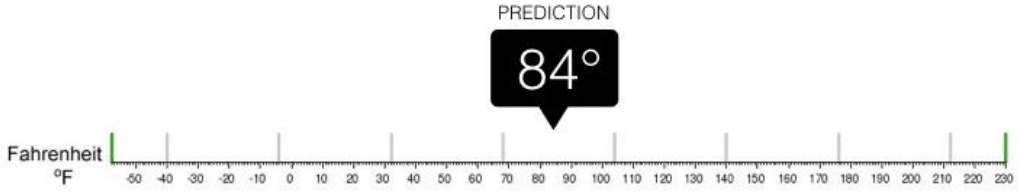
- Supervised learning is the most popular paradigm for machine learning.
- It is very similar to teaching a child with the use of flash cards.
- Supervised learning is often described as task-oriented. It is highly focused on a singular task, feeding more and more examples to the algorithm until it can accurately perform on that task.



Supervised ML Algorithm

Regression

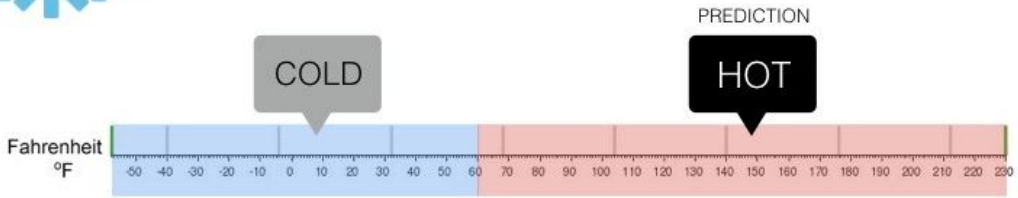
What is the temperature going to be tomorrow?



the output variable is a real value, such as “dollars” or “weight”.

Classification

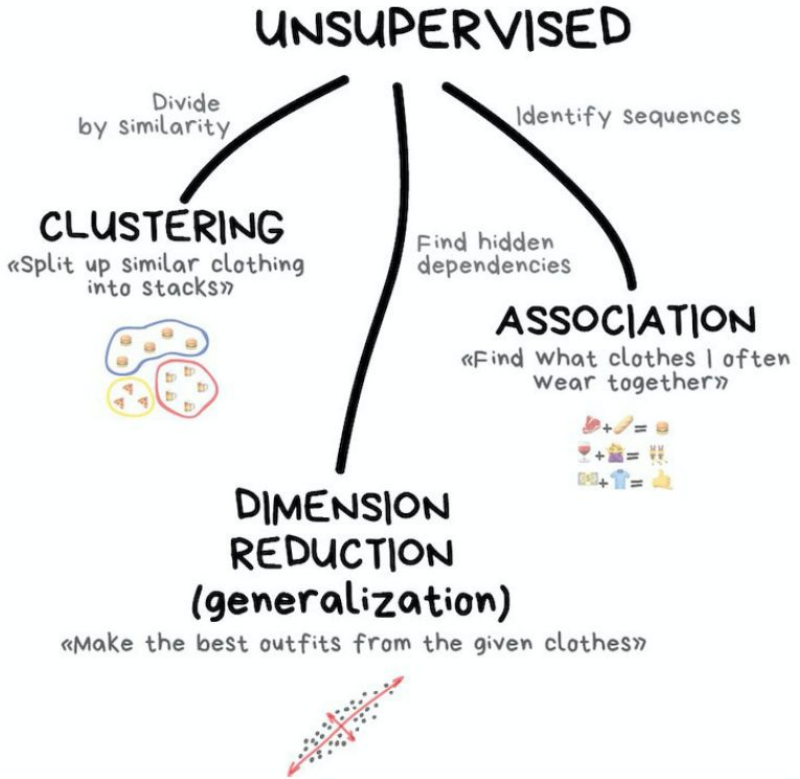
Will it be Cold or Hot tomorrow?



output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

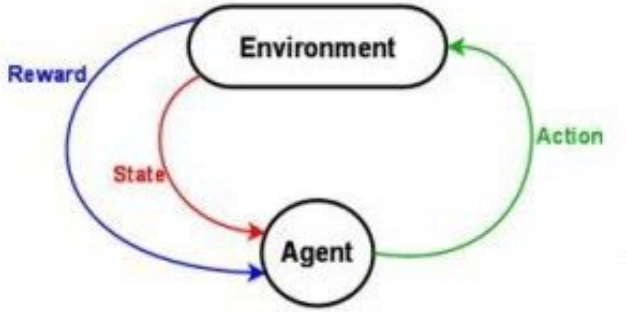
UnSupervised ML Algorithm

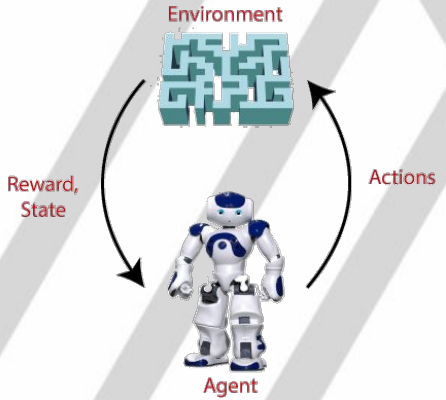
- Unsupervised Learning is a machine learning technique in which the users do not need to supervise the model. It mainly deals with the unlabelled data.
- Instead, it allows the model to work on its own to discover patterns and information that was previously undetected.
- Allow users to perform more complex processing tasks compared to supervised learning. Although, unsupervised learning can be more unpredictable compared with other natural learning methods.



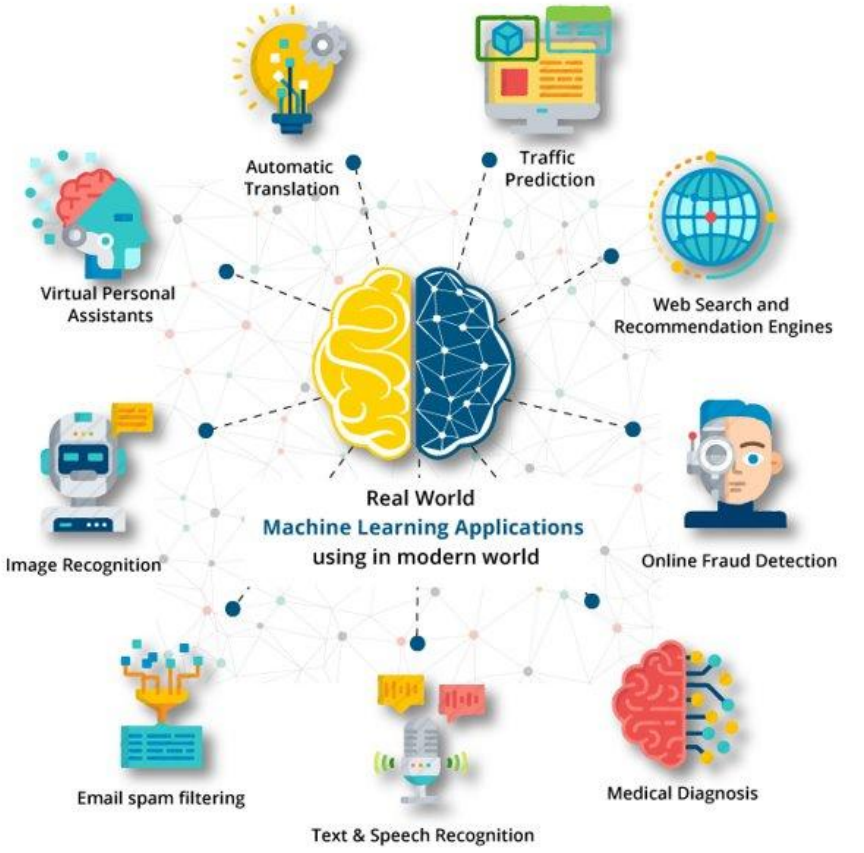
Reinforcement Learning

- RL enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.
- Though both supervised and RL use mapping between input and output, unlike supervised learning where feedback provided to the agent is correct set of actions for performing a task, reinforcement learning uses rewards and punishment as signals for positive and negative behavior.
- As compared to unsupervised learning, RL is different in terms of goals. While the goal in unsupervised learning is to find similarities and differences between data points, in reinforcement learning the goal is to find a suitable action model that would maximize the total cumulative reward of the agent.



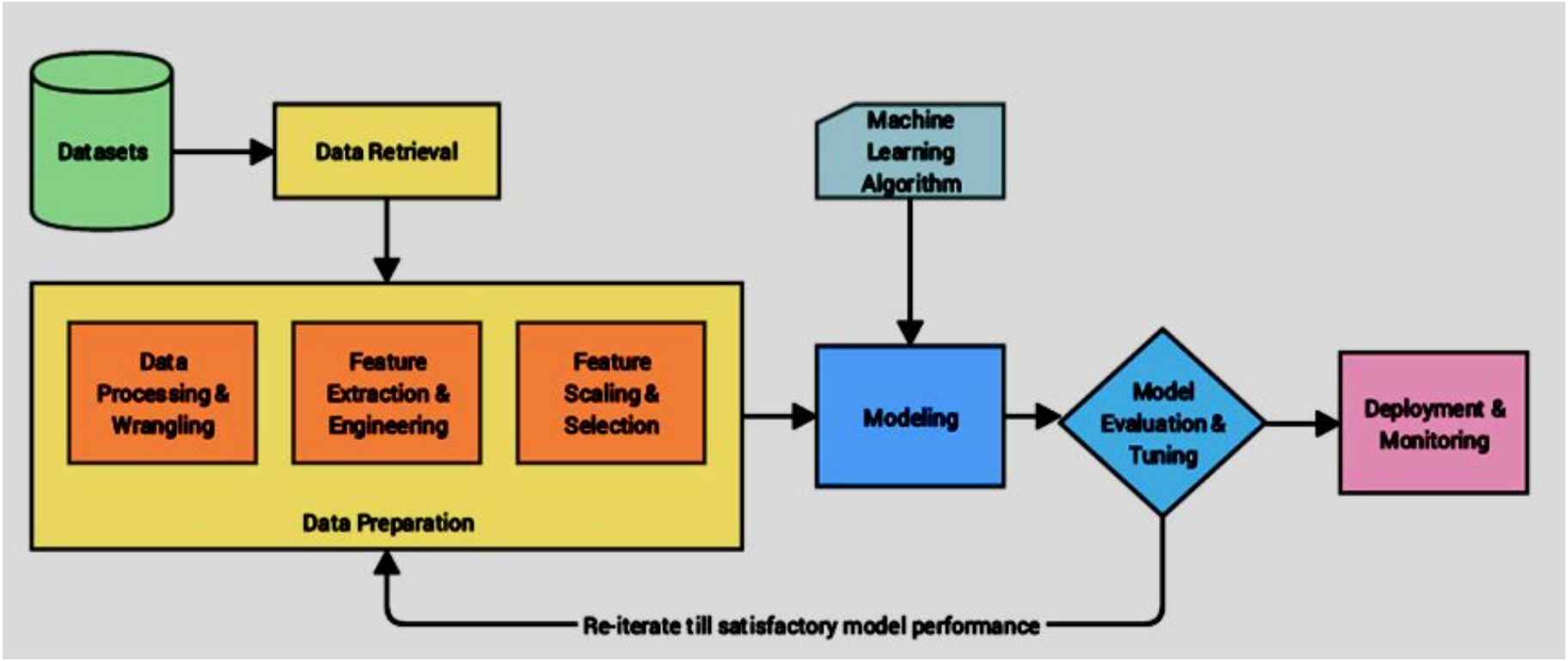


ML Applications

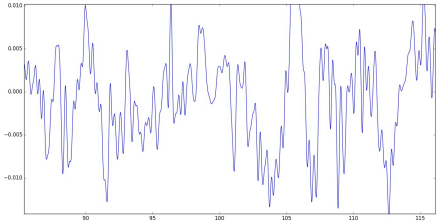


ML WorkFlow

AltaML



ML Input Data



AltaML

IMAGE

VIDEO

SIGNAL

TEXT

DATA FRAME

DataFrame object

Column names

Label index
(country code)

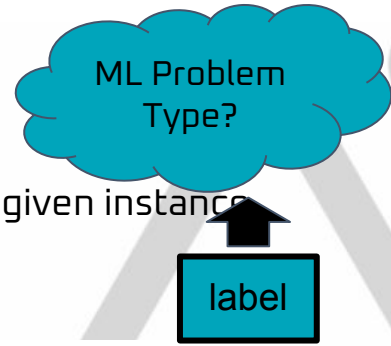
	Country	Popu	Percent
IT	Italy	61	0.83
ES	Spain	46	0.63
GR	Greece	11	0.15
FR	France	65	0.88
PO	Portugal	10	0.14

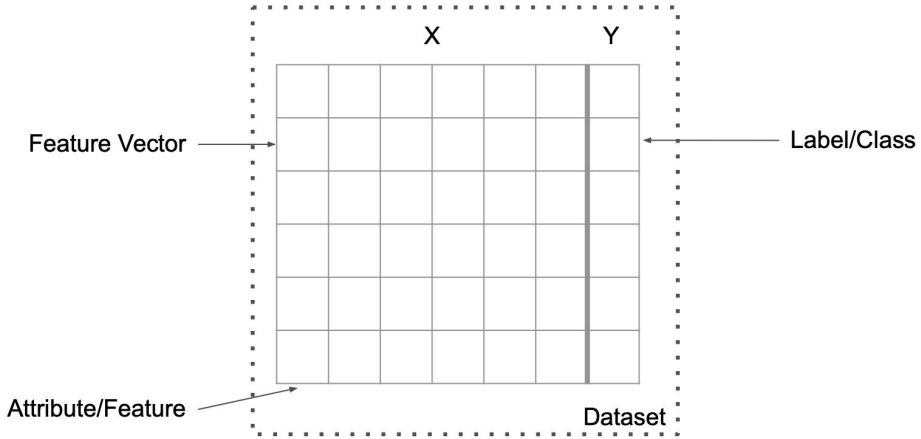
Data
(different type in each column)

ML Input: Data Frame Terminology

AltaML

- Feature/Attribute: A single variable (binary, nominal, numerical)
- Instance/Feature vector: One entity described by features
- Label/Class/Target Variable: An extra information that categorizes/classifies a given instance
- Dataset: Collection instances

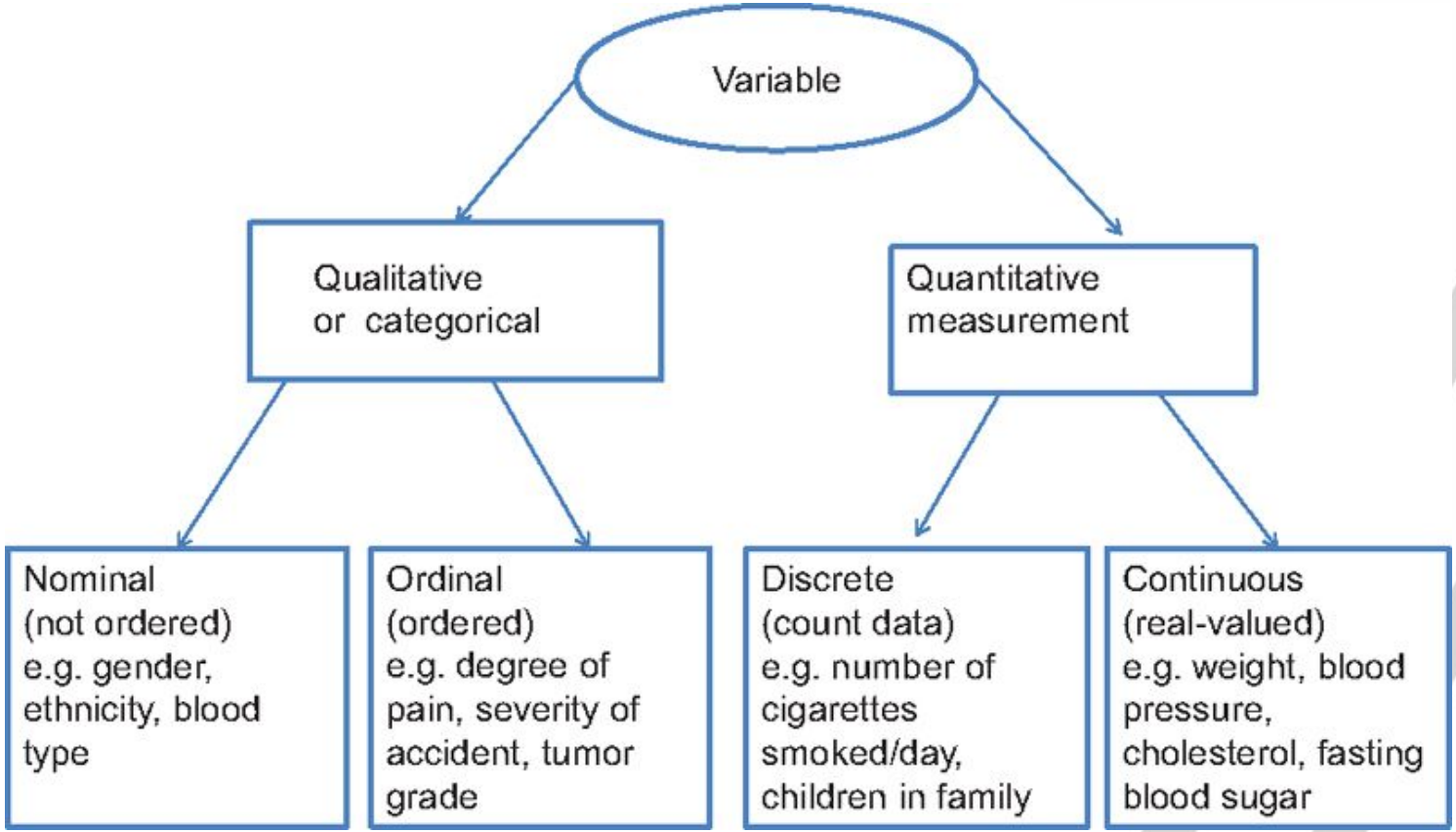




	state	color	food	age	height	score
Jane	NY	blue	Steak	30	165	4.6
Niko	TX	green	Lamb	2	70	8.3
Aaron	FL	red	Mango	12	120	9.0
Penelope	AL	white	Apple	4	80	3.3
Dean	AK	gray	Cheese	32	180	1.8
Christina	TX	black	Melon	33	172	9.5
Cornelia	TX	red	Beans	69	150	2.2

Data Frame Variable Data Types

AltaML



EDA helps us understand various facets of our data. In this step, we analyze different attributes of data, uncover interesting insights, and even visualize data on different dimensions to get a better understanding.

Data Preprocessing

Preprocessing - Missing Values

AltaML

User forgot to fill in a field.

Data was lost while transferring manually from a legacy database.

There was a programming error.

Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

solutions

Dropping
- *Row-wise*
- *Column-wise*

Imputation
- *Constant value*
- *Statistical value*

Gender	Age
Male	29
Male	NA
NA	43
Female	25
Male	34
NA	50
Female	NA

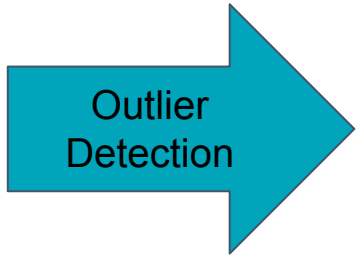
Gender	Age
Male	29
Male	34
Female	43
Female	25
Male	34
Male	50
Female	25

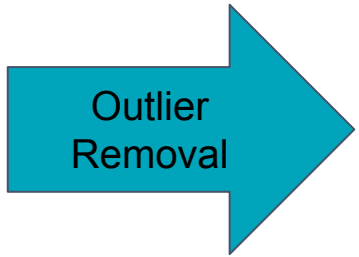
Random Sample Imputation

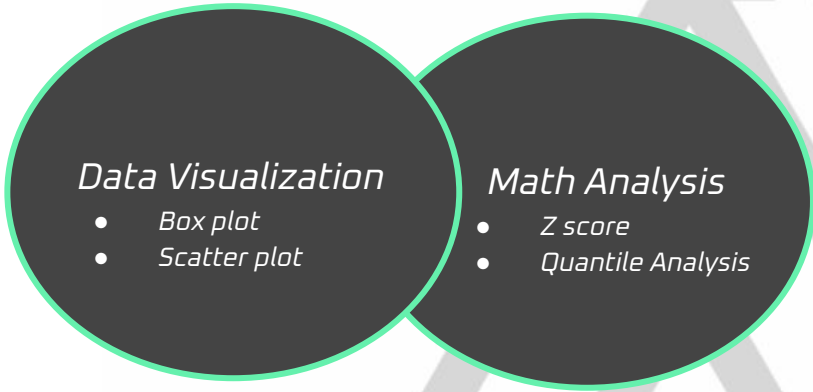
Preprocessing - outliers

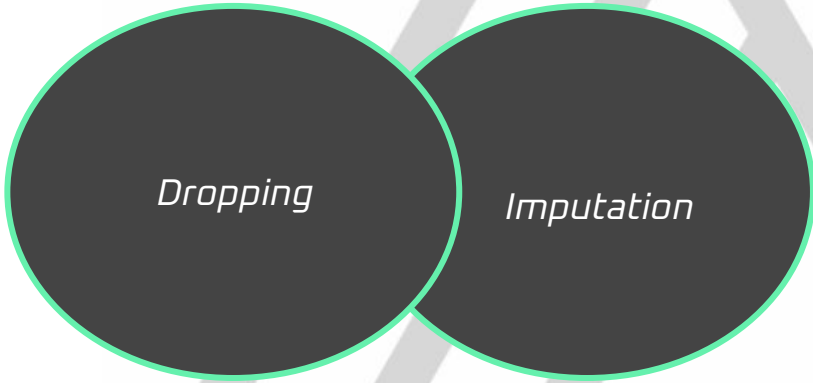
AltaML

In statistics, an outlier is an observation point that is distant from other observations. Possible reasons for outliers are recording errors, unusual sampling and laboratory procedures or conditions.



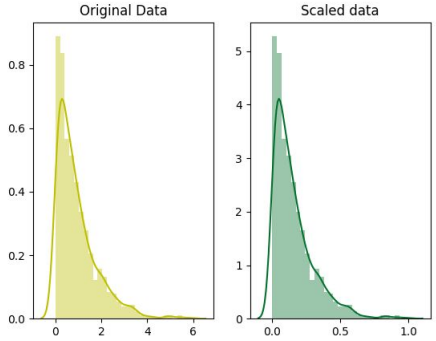


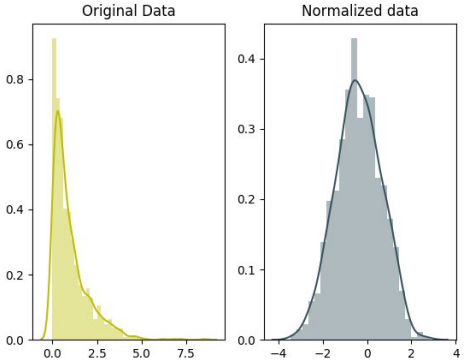




Preprocessing - Scaling

Data may contain attributes with a mixture of scales for various quantities such as dollar, kilogram, and sales volume.





$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Scaling for numerical variable

$$x_{scaled} = \frac{x - mean}{sd}$$

Normalization

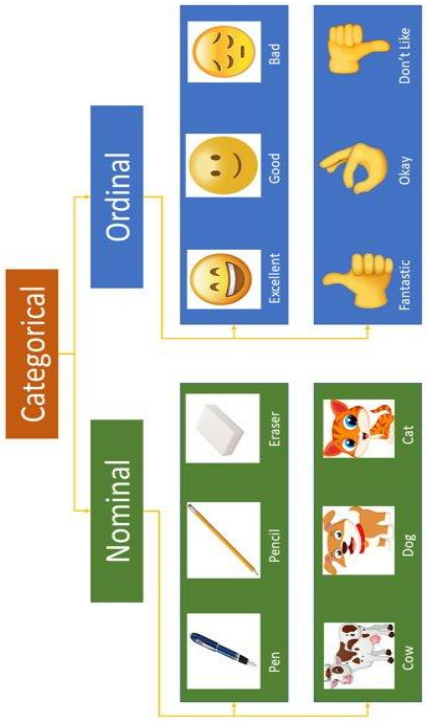
Sensitive to outliers

Standardization

Values are not bounded.

Preprocessing - Scaling

AltaML



Scaling for categorical variable

Label Encoding

- *Dataset label*
- *Ordinal Variable*

Breakfast
Every day
Never
Rarely
Most days
Never

→

Breakfast
3
0
1
2
0

Careful about variable order!!

One Hot Encoding Nominal Variables

color
red
green
blue
red

→

color_red	color_blue	color_green
1	0	0
0	0	1
0	1	0
1	0	0

Feature Creation

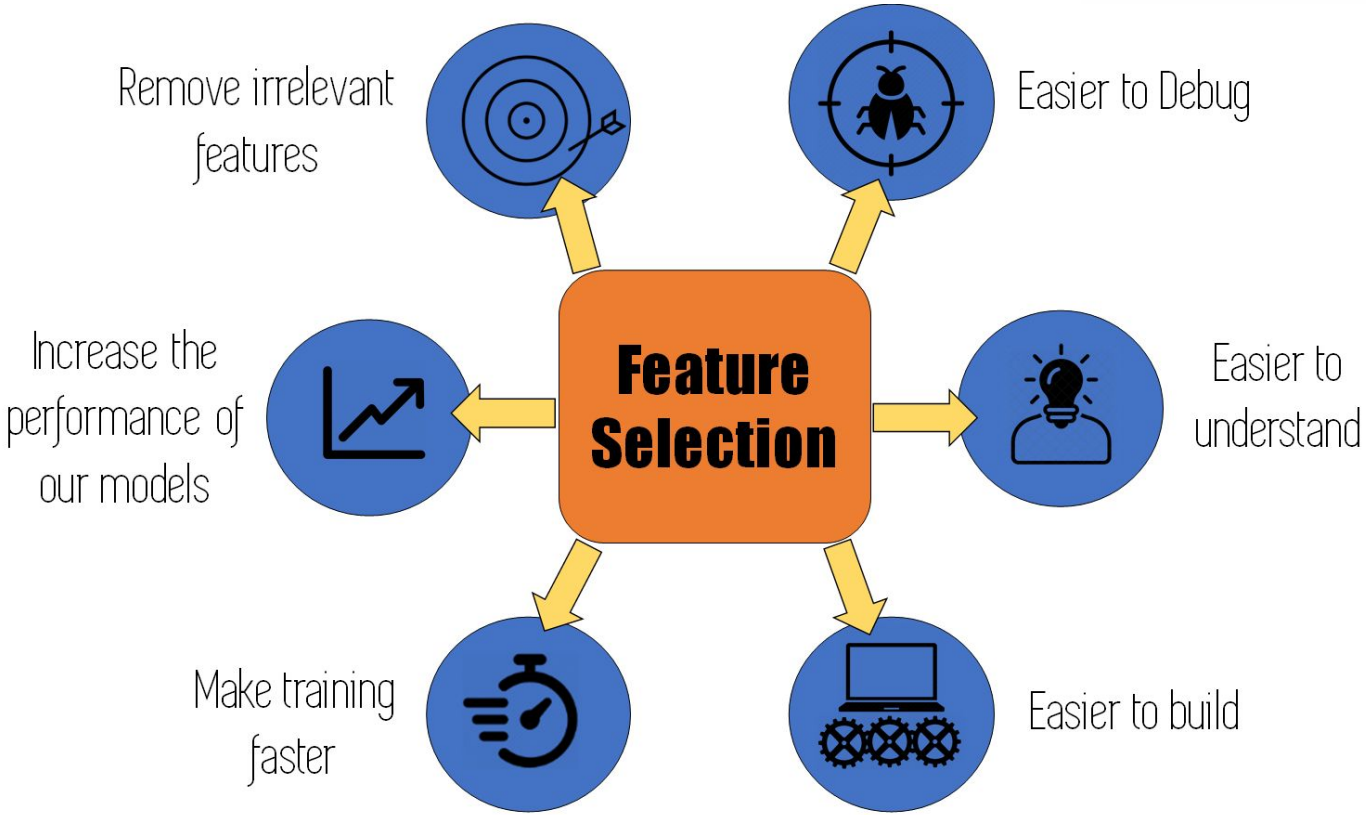
AltaML

Feature engineering is the process of using domain knowledge to transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

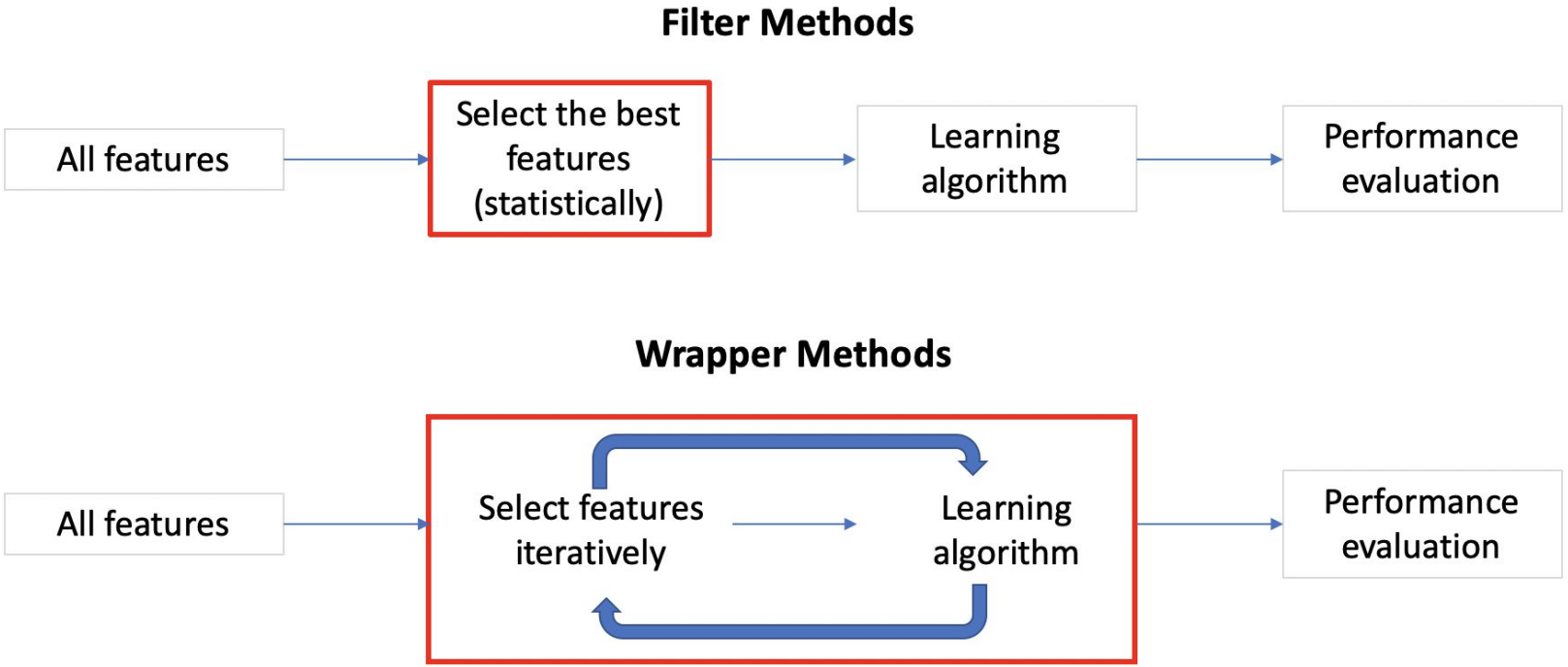
New feature

client_id	joined	income	credit_score	join_month	log_income
46109	2002-04-16	172677	527	4	12.059178
49545	2007-11-14	104564	770	11	11.557555
41480	2013-03-11	122607	585	3	11.716739
46180	2001-11-06	43851	562	11	10.688553
25707	2006-10-06	211422	621	10	12.261611

Feature Selection

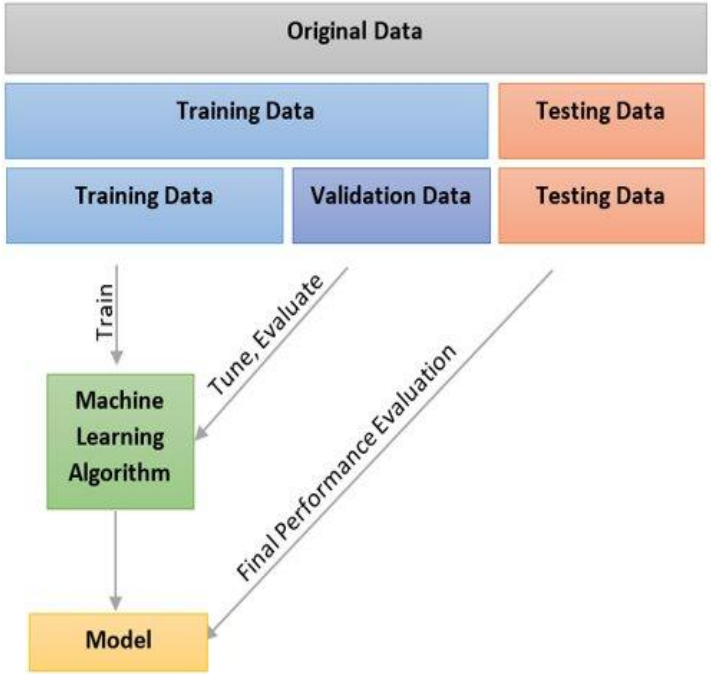


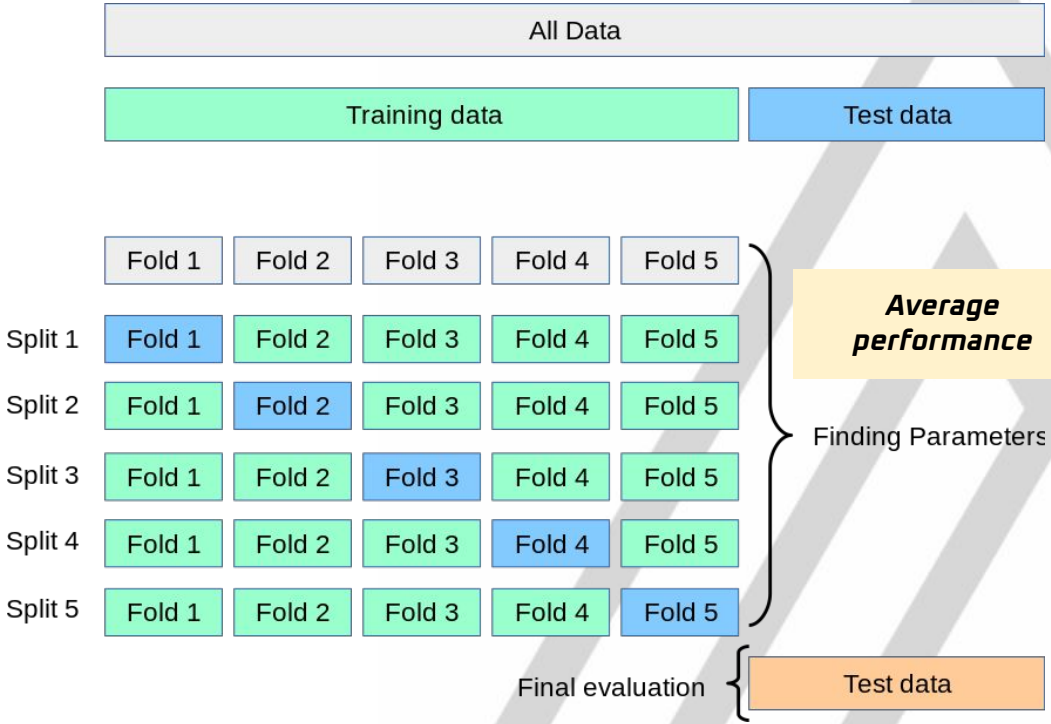
Feature Selection



© Jack Tan

Data Division





Evaluation

Regression

- MSPE
- MSAE
- R Square
- Adjusted R Square

Classification

- Precision-Recall
- ROC-AUC
- Accuracy
- Log-Loss

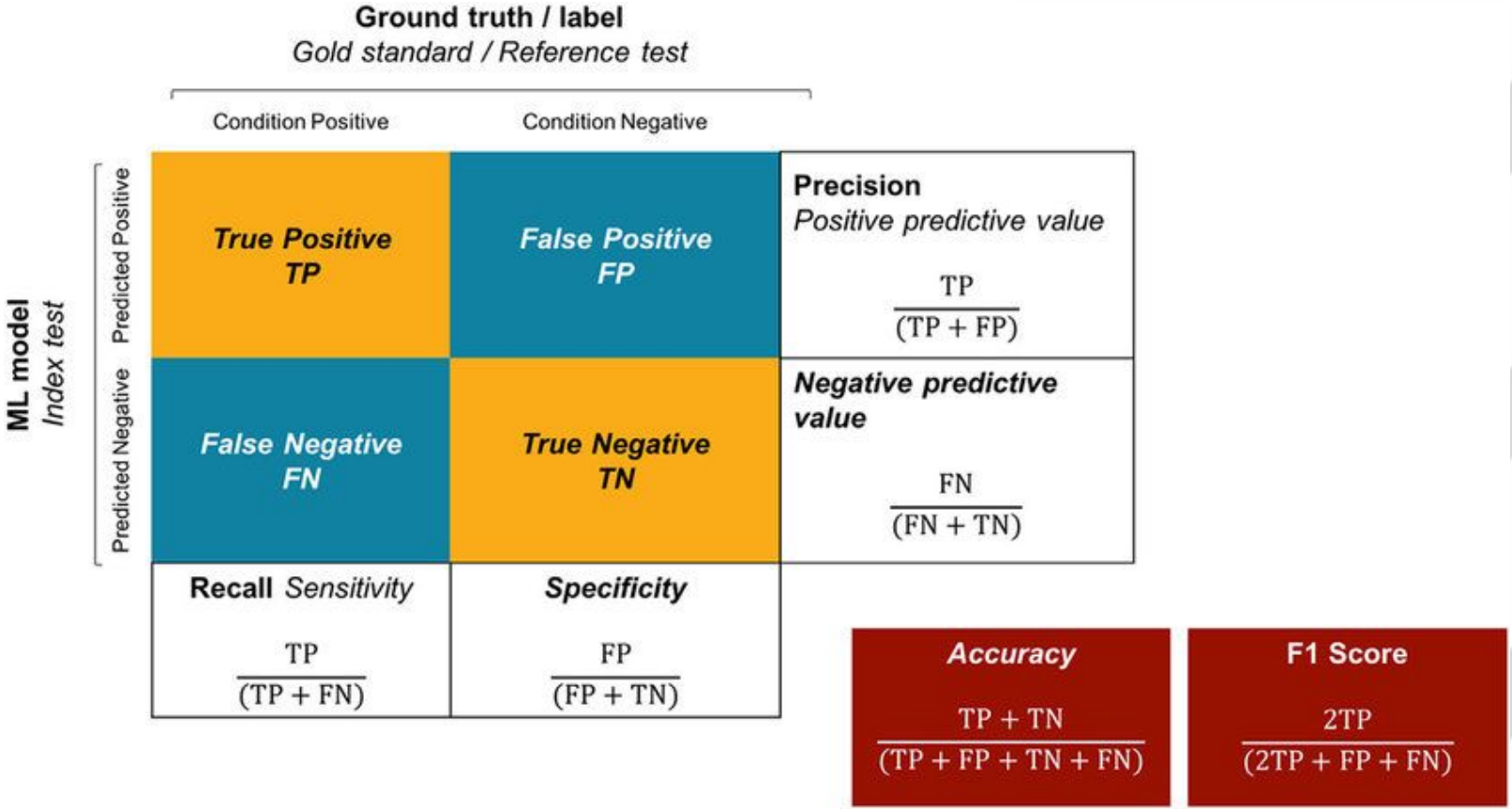
Unsupervised Models

- Rand Index
- Mutual Information

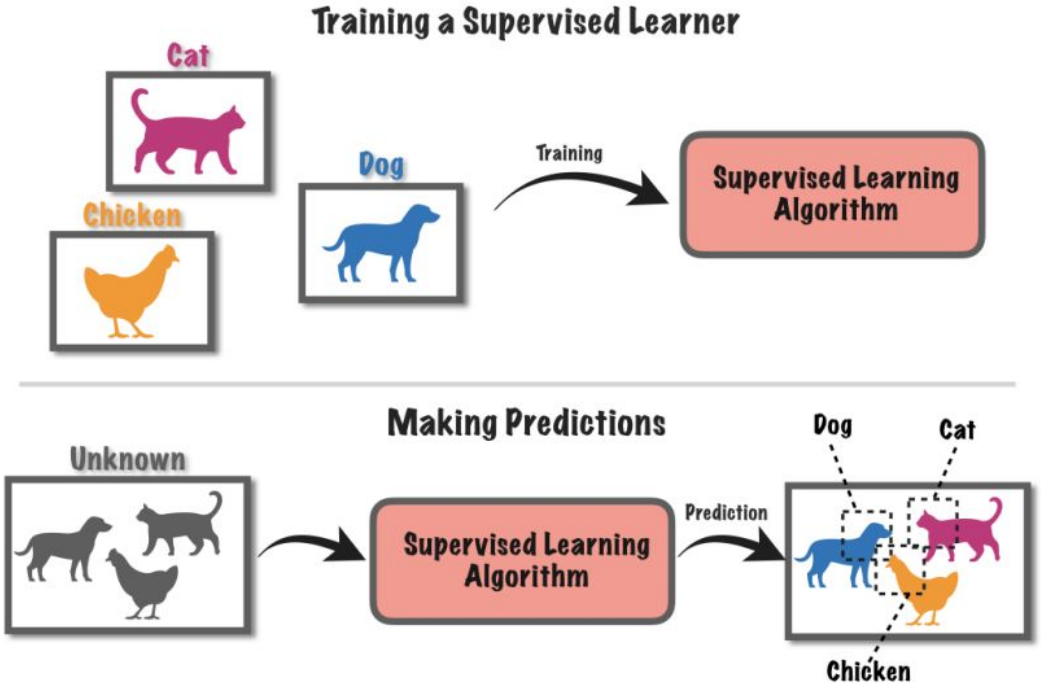
Others

- CV Error
- Heuristic methods to find K
- BLEU Score (NLP)

Evaluation


Ground truth / label
Gold standard / Reference test
Condition Positive
Condition Negative
ML model
Index test
Predicted Positive
Predicted Negative
True Positive
TP
False Positive
FP
False Negative
FN
True Negative
TN
Precision
Positive predictive value
TP / (TP + FP)
Negative predictive value
FN / (FN + TN)
Recall Sensitivity
TP / (TP + FN)
Specificity
FP / (FP + TN)
Accuracy
(TP + TN) / (TP + FP + TN + FN)
F1 Score
2TP / (2TP + FP + FN)

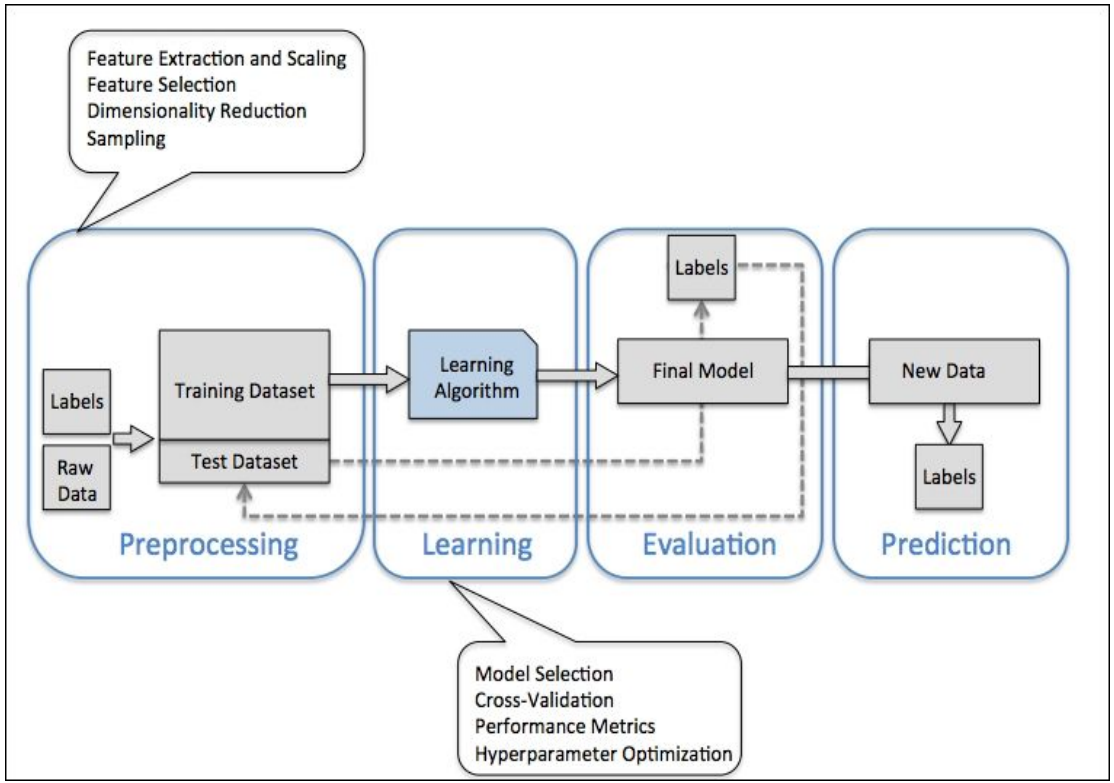
Evaluation for more than 2 classes

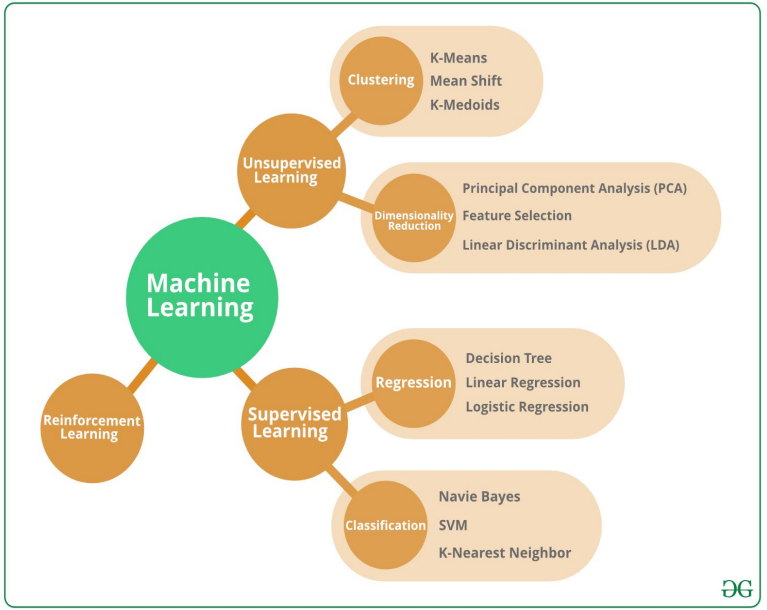


		True Class				
		A	B	C	D	Total
Predicted Class	A					
	B					
	C					
	D					
	Total					

Multiclass Accurary = Sum of yellow cells / Green cell

Modelling





Which one is the best ml algorithm to choose?
What are the best parameter values for that best model?

AltaML

hands-on
TRAINING

Q&A

sara@altaml.com

https://www.linkedin.com/in/sarasoltanineja d/?originalSubdomain=ca

THANK YOU