ClusterKit

HPCNC

Machine Learning with Apache Spark & Zeppelin

นันท์นภัส ม่วงมิ่งสุข (เพชร)
Nunnapus Moungmingsuk
nunnapus@clusterkit.co.th

อุทิศแด่

ผศ.ดร.ภุชงค์ อุทโยภาศ

ผู้ก่อตั้งแลป HPCNC

ครูผู้ประสิทธิ์ประสาทวิชา

Outline

- Machine Learning
  - Machine Learning Workflow
  - Types of Machine Learning
  - Choosing the Right Learning Algorithm
- Apache Spark
  - Spark MLlib

Outline

- Machine Learning with Apache Spark
  - Data Preprocessing
  - Clustering
  - Classification
  - Regression
  - Collaborative Filtering
  - Frequent Pattern Mining
  - ML Pipelines

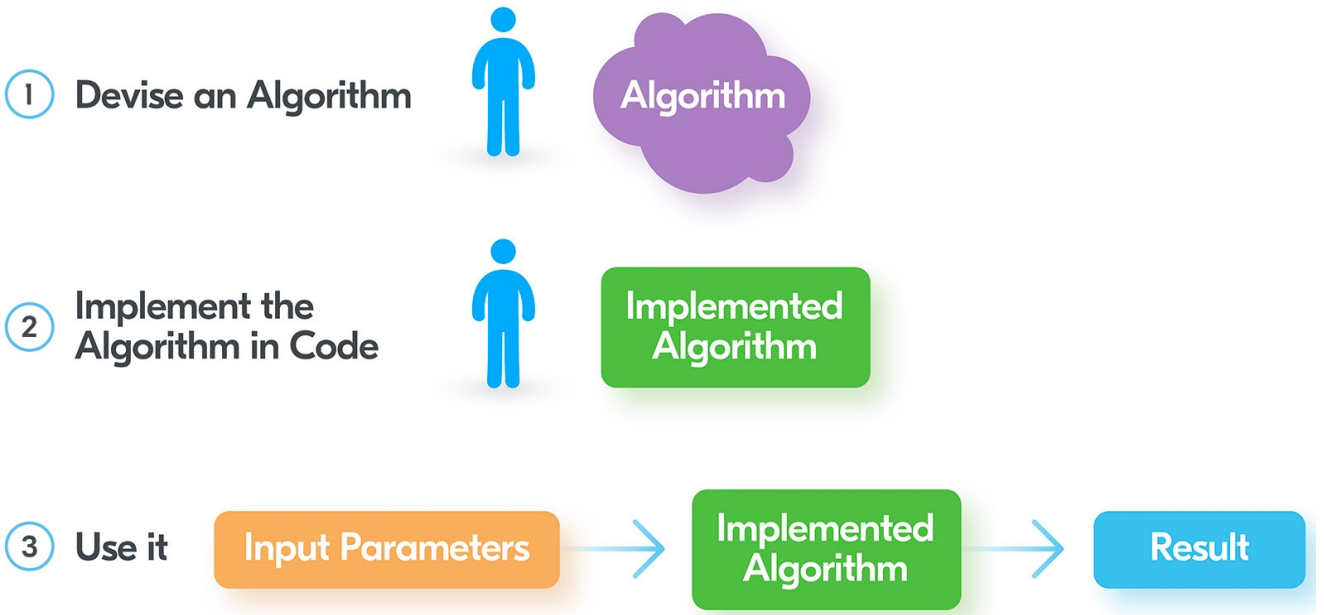
Machine Learning (ML)

What is Machine Learning (ML)?

- ML is a field of computer science that evolved from studying pattern recognition and computational learning theory in AI.
- It is the learning and building of algorithms that can learn from and make predictions on data sets.

Traditional Programming Approach

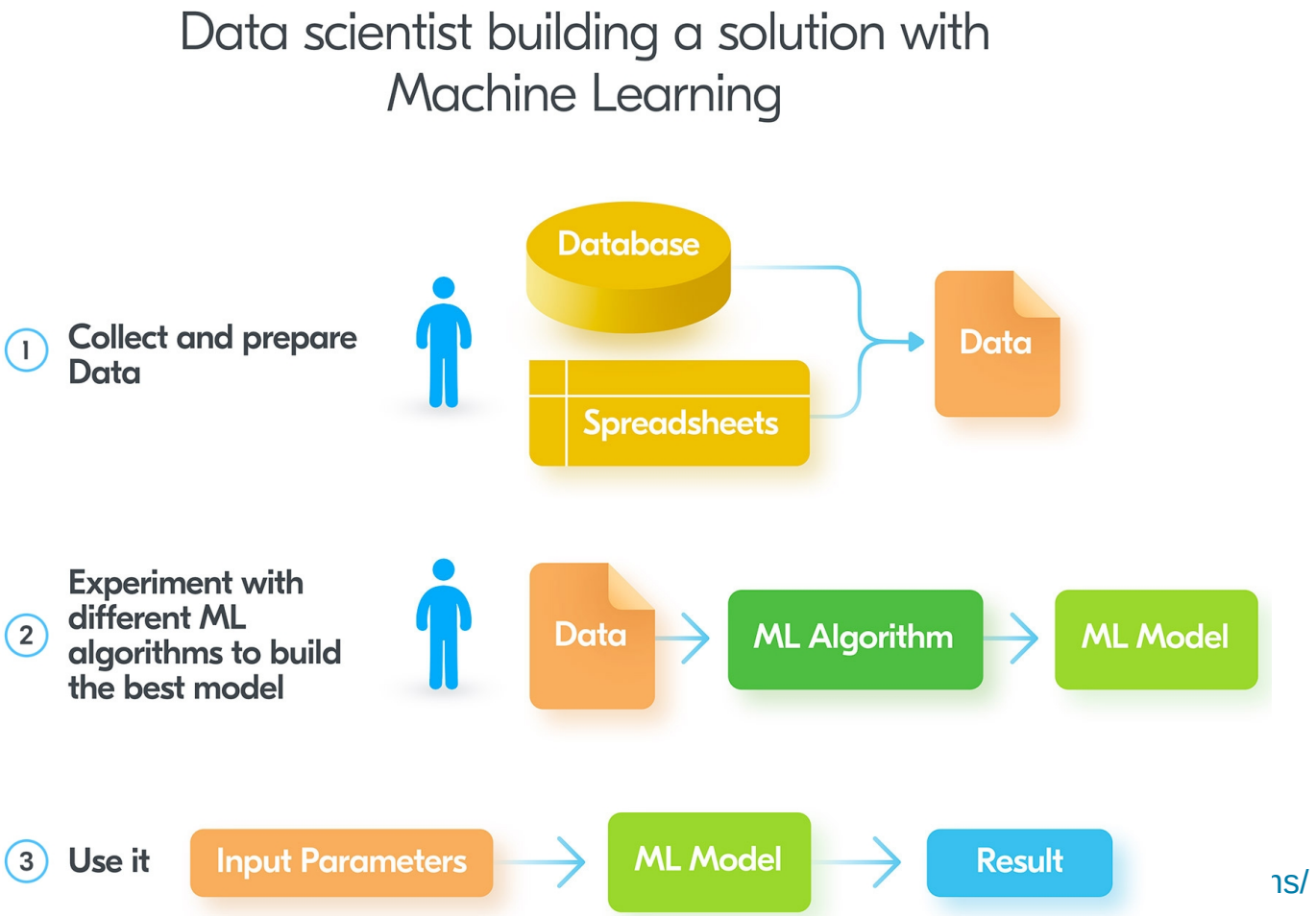
Engineer building a solution with traditional programming



https://railsware.com/blog/practical-machine-learning-for-solving-real-world-problems/

ClusterKit

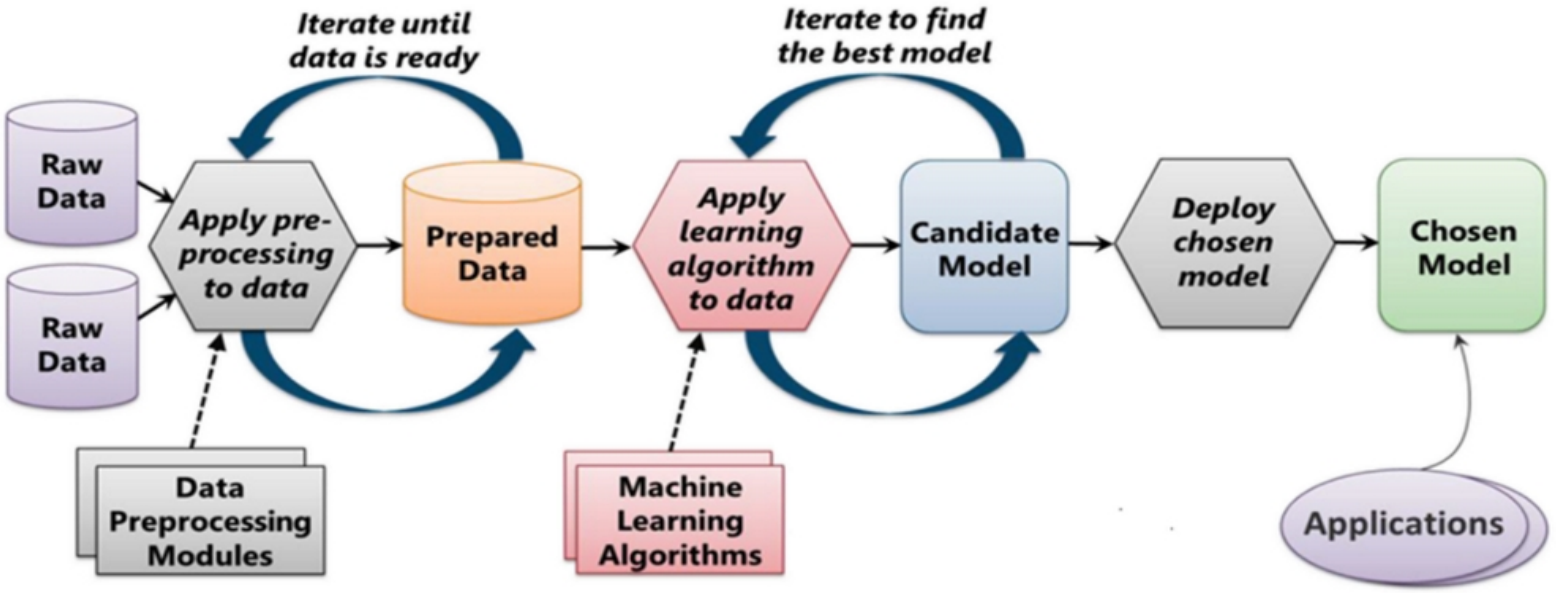
Machine Learning Approach



ClusterKit

ıs/

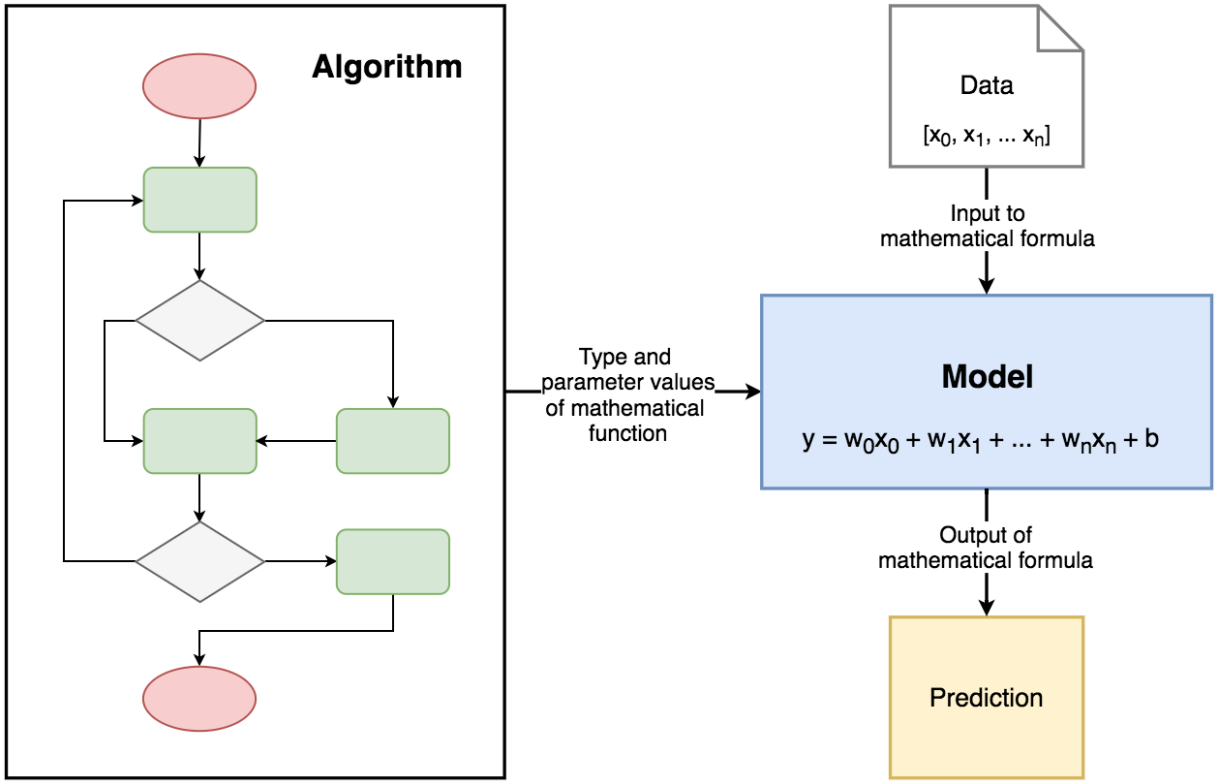
Machine Learning Process



From "Introduction to Microsoft Azure" by David Chappell

ClusterKit

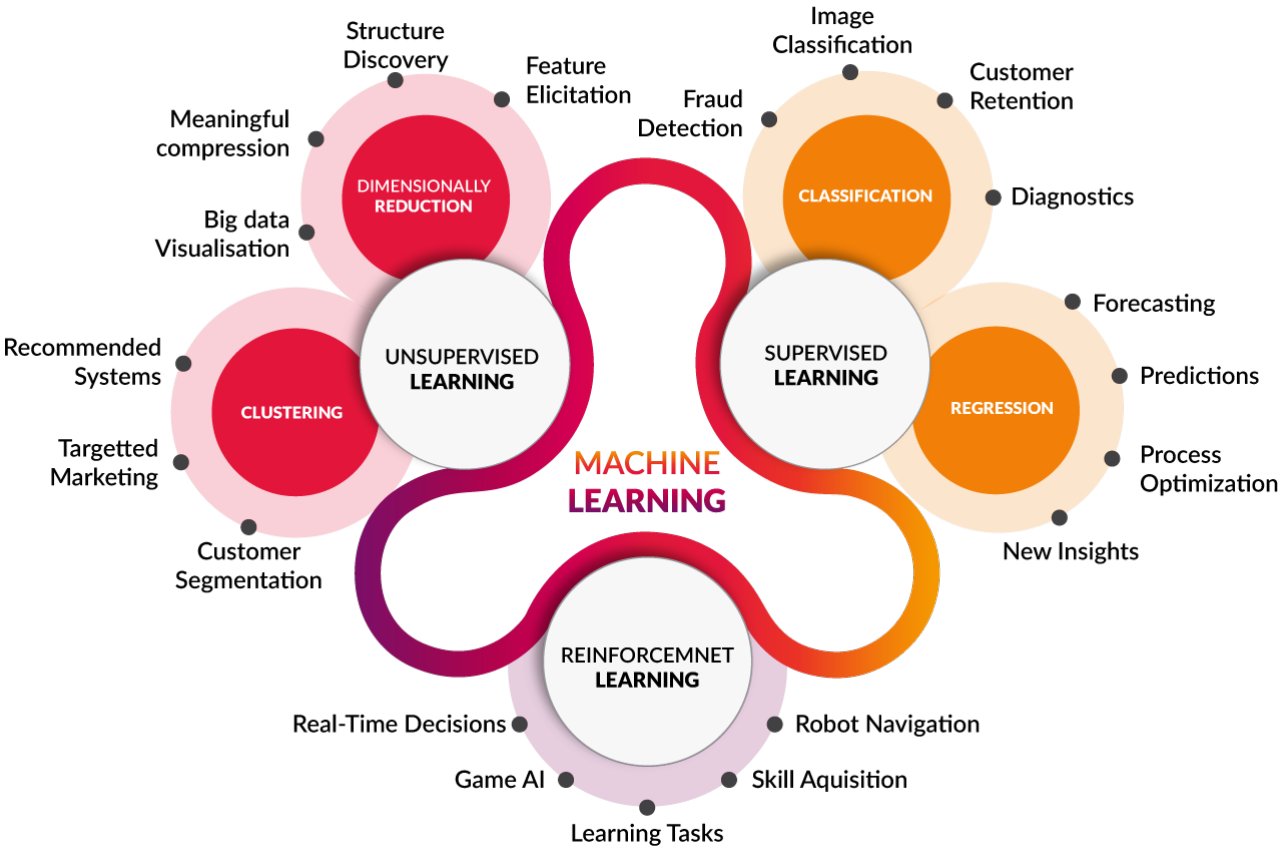
Machine Learning Model



ClusterKit

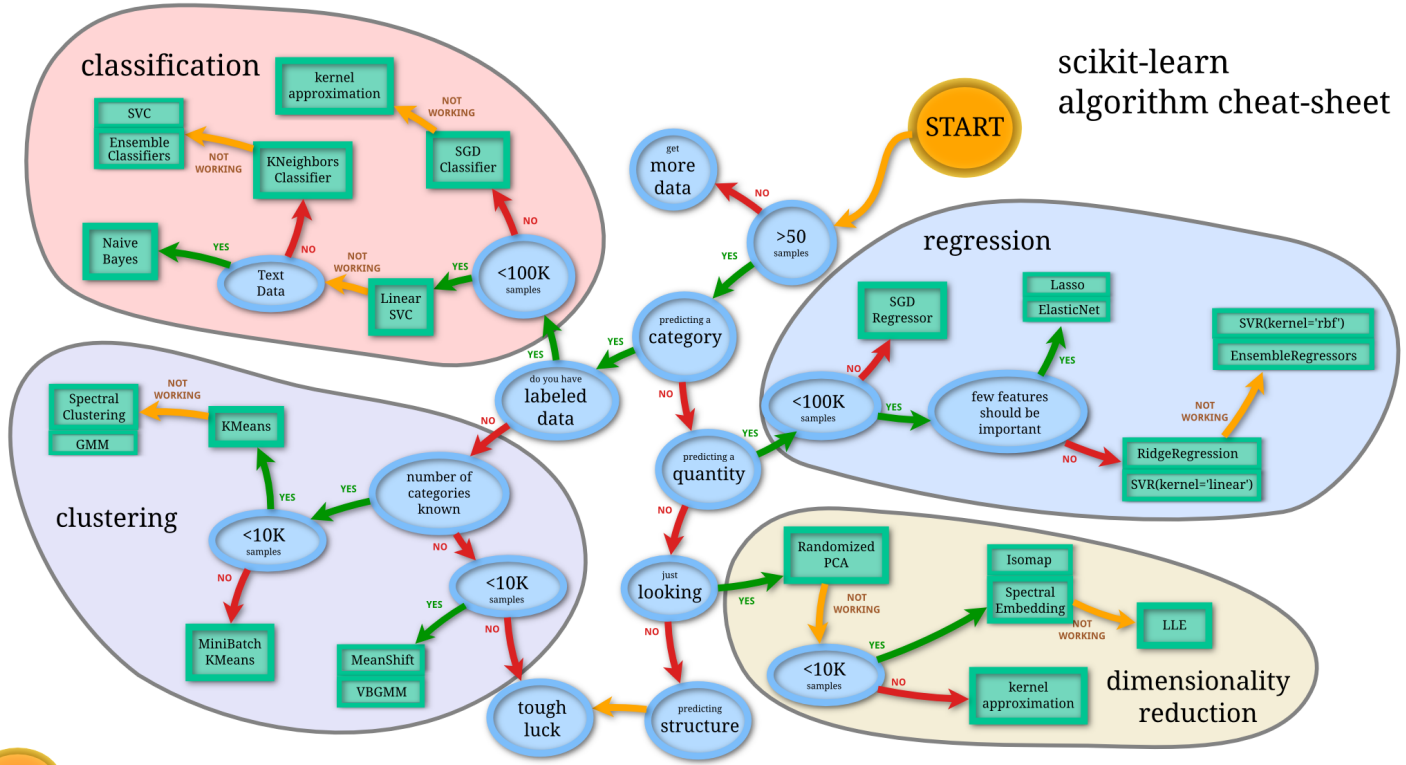
https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d

Types of Machine Learning



https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d

Choosing the Right Learning Algorithm



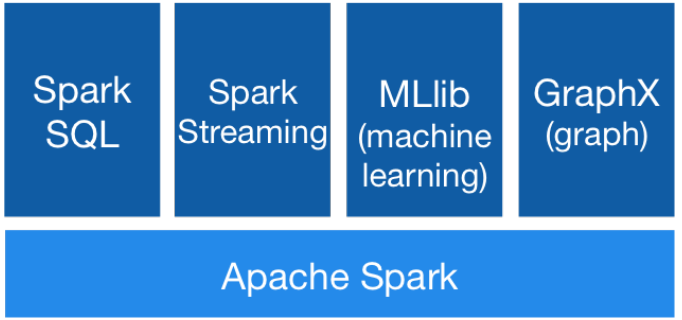
https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Apache Spark

ClusterKit

What is Apache Spark?

- Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.
- It provides high-level APIs in Java, Scala, Python and R.
- It also supports a rich set of higher-level tools including
  - **Spark SQL** for SQL and structured data processing
  - **Spark Streaming**
  - **MLlib** for machine learning
  - **GraphX** for graph processing



http://spark.apache.org/docs/latest/index.html

Simple. Fast. Scalable. Unified.

Key features

Batch/streaming data

Unify the processing of your data in batches and real-time streaming, using your preferred language: Python, SQL, Scala, Java or R.

SQL analytics

Execute fast, distributed ANSI SQL queries for dashboarding and ad-hoc reporting. Runs faster than most data warehouses.

Data science at scale

Perform Exploratory Data Analysis (EDA) on petabyte-scale data without having to resort to downsampling

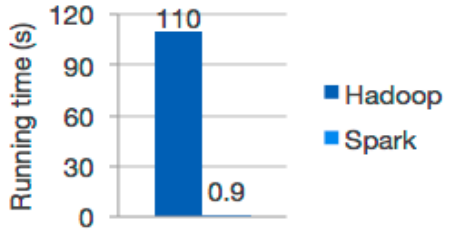
Machine learning

Train machine learning algorithms on a laptop and use the same code to scale to fault-tolerant clusters of thousands of machines.

ClusterKit

MLlib

- **MLlib** is Apache Spark's scalable machine learning library.
- Ease of use
  - Usable in Java, Scala, Python, and R.
  - You can use any Hadoop data source (e.g. HDFS, HBase, or local files), making it easy to plug into Hadoop workflows.
- Performance
  - High-quality algorithms, 100x faster than MapReduce.
- Runs everywhere
  - Spark runs on Hadoop, Apache Mesos, Kubernetes, standalone, or in the cloud, against diverse data sources.



Logistic regression in Hadoop and Spark

Spark MLlib

- Algorithms
  - ML algorithms include:
    - Classification: logistic regression, naive Bayes,...
    - Regression: generalized linear regression, survival regression,...
    - Decision trees, random forests, and gradient-boosted trees
    - Recommendation: alternating least squares (ALS)
    - Clustering: K-means, Gaussian mixtures (GMMs),...
    - Frequent itemsets, association rules, and sequential pattern mining

MLlib

- Algorithms
  - ML workflow utilities include:
    - Feature transformations: standardization, normalization, hashing,...
    - ML Pipeline construction
    - Model evaluation and hyper-parameter tuning
    - ML persistence: saving and loading models and Pipelines

Data Preprocessing

What is Data Preprocessing?

- A technique that involves transforming raw data into an understandable format.
- Real-world data is often incomplete, inconsistent, and is likely to contain many errors.
- This technique is performed before the execution of Iterative Analysis.
- The set of steps is known as Data Preprocessing.
    - Data Cleaning
    - Data Integration
    - Data Transformation
    - Data Reduction

Task of Data Preprocessing

- **Data Cleaning**: handling missing data, noisy data, detection, and removal of outliers
- **Data Integration**: used when data is gathered from various data sources and data are combined to form consistent data.
- **Data Transformation**: used to convert the raw data into a specified format according to the need of the model.
- **Data Reduction**: redundancy within the data is removed and efficiently organize the data

Hands-on:

1. Data Preprocessing
2. Data Visualization

Machine Learning Algorithm:

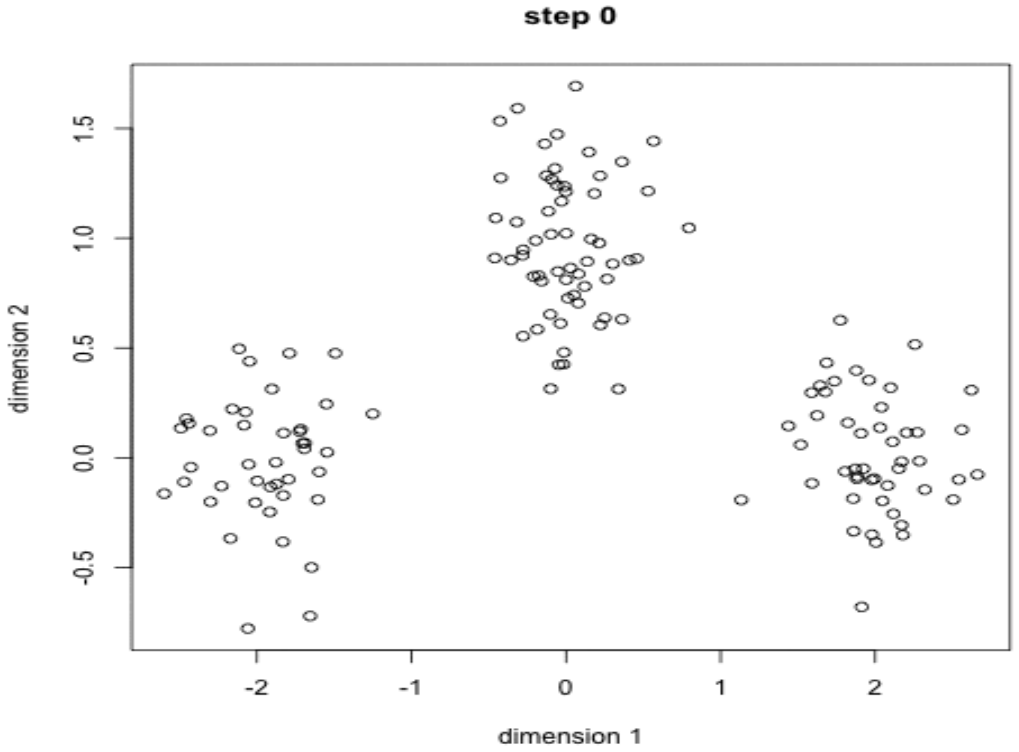
Clustering

Clustering

- Clustering is a Machine Learning technique that involves the grouping of data points.
- A clustering algorithm: classify each data point into a specific group.
- In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features.

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68#:~:text=Given%20a%20set%20of%20data,dissimilar%20properties%20and%2For%20features.

K-Means Clustering



ClusterKit

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68#:~:text=Given%20a%20set%20of%20data,dissimilar%20properties%20and%2For%20features.

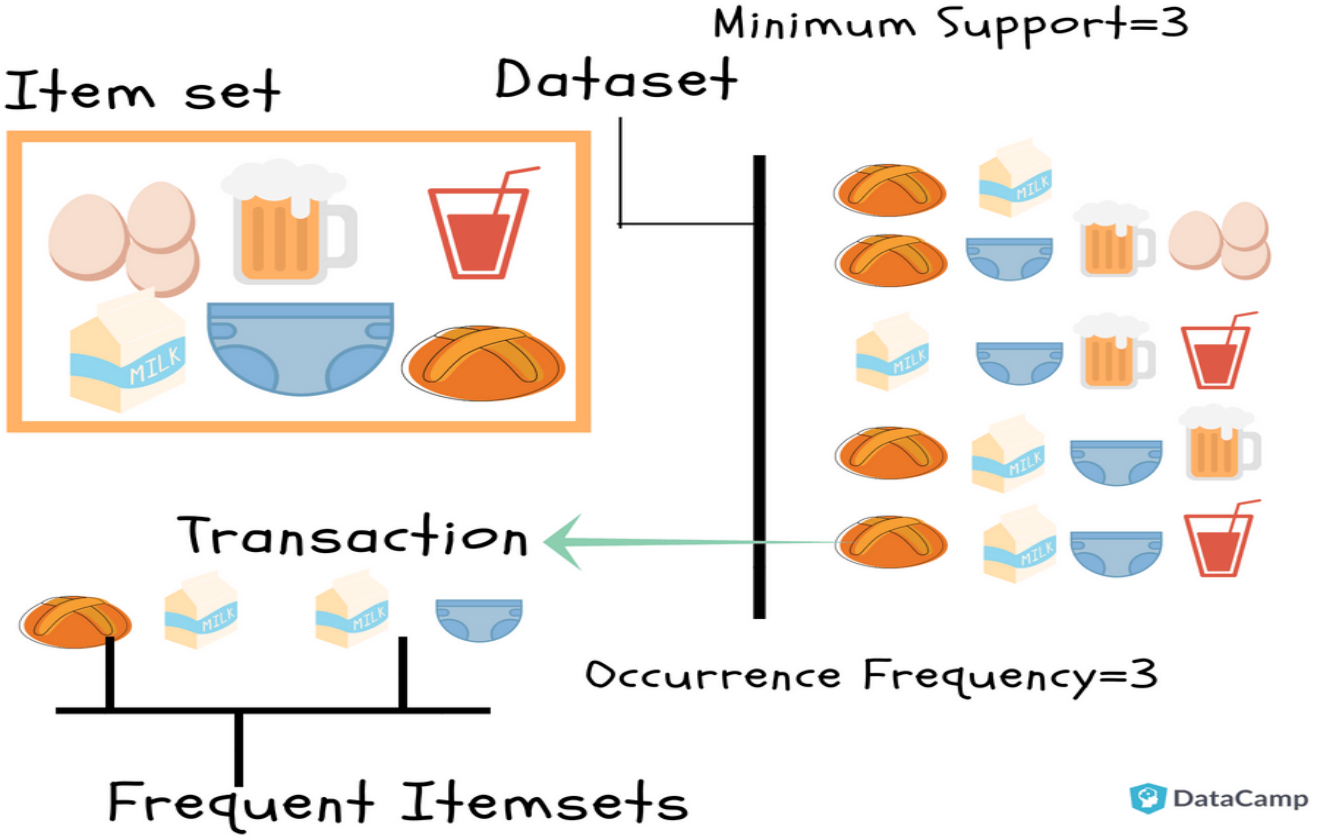
Hands-on:

3. Clustering

Machine Learning Algorithm:

Frequent Pattern Mining

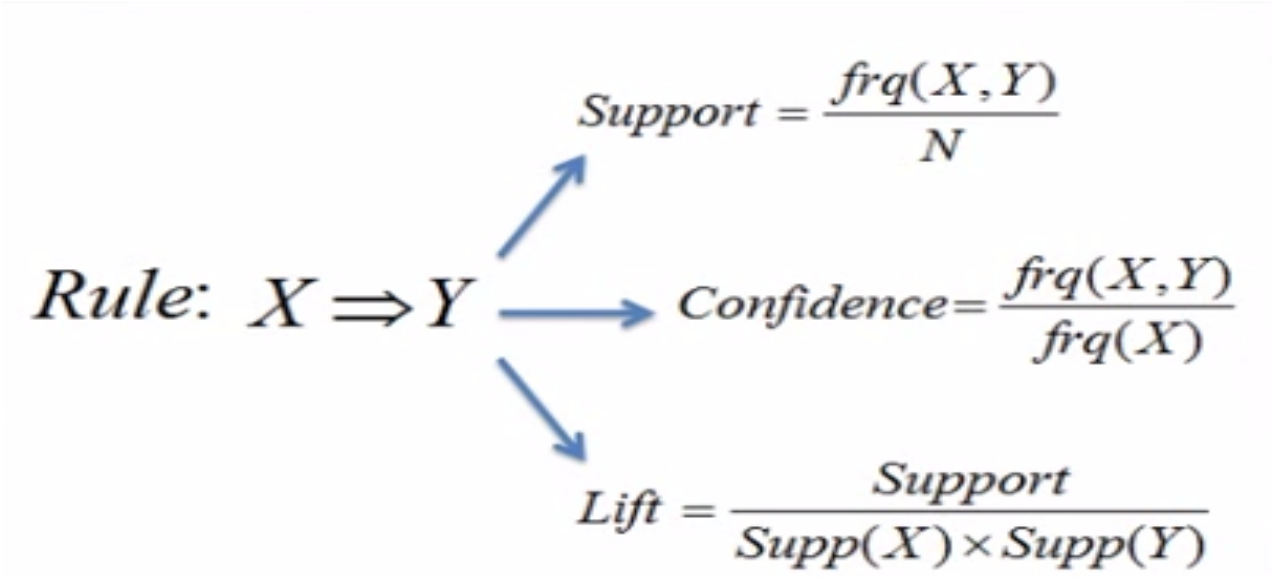
Association Rule Mining



ClusterKit

https://www.datacamp.com/community/tutorials/market-basket-analysis-r

Association Rule Mining



ClusterKit

Frequent Pattern Mining: FP-Growth

TID	Items Bought	(Ordered) Frequent Items
100	f, a, c, d, g, i, m, p	f, c, a, m, p
200	a, b, c, f, l, m, o	f, c, a, b, m
300	b, f, h, j, o	f, b
400	b, c, k, s, p	c, b, p
500	a, f, c, e, l, p, m, n	f, c, a, m, p

Table 1: A transaction database as running example.

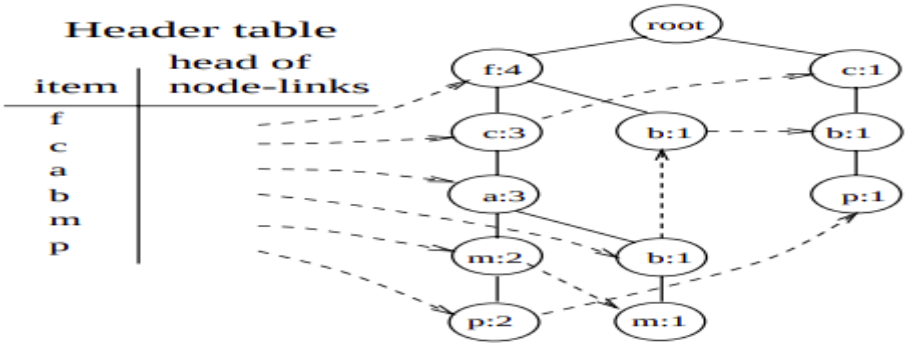


Figure 1: The FP-tree in Example 1.

ClusterKit

Han et al., Mining frequent patterns without candidate generation

Hands-on:

4. Frequent Pattern Mining

Machine Learning Algorithm:

Collaborative Filtering

Recommendation system

- Recommendation systems usually make use of either or both **collaborative filtering** and content-based filtering
- **Collaborative filtering approaches** build a model from a user's past behavior (items previously purchased, ratings given to items) as well as similar decisions made by other users.
- This model is then used to predict items (or ratings for items) that the user may have an interest in.

Prem Melville and Vikas Sindhwani, Recommender Systems, Encyclopedia of Machine Learning, 2010.

Collaborative filtering

- An example of collaborative filtering based on a ratings system




https://en.wikipedia.org/wiki/Recommender_system#Collaborative_filtering

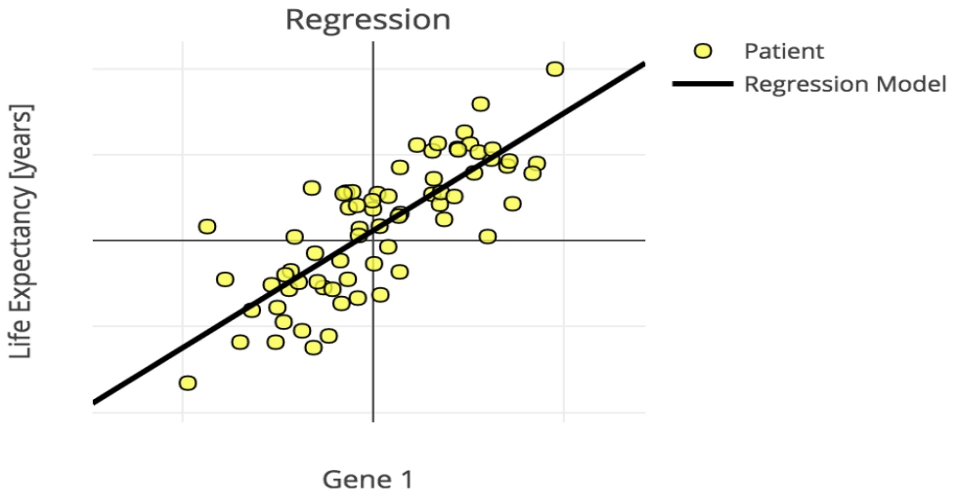
Hands-on:

5. Collaborative Filtering

Machine Learning Algorithm:
Classification and Regression

Regression

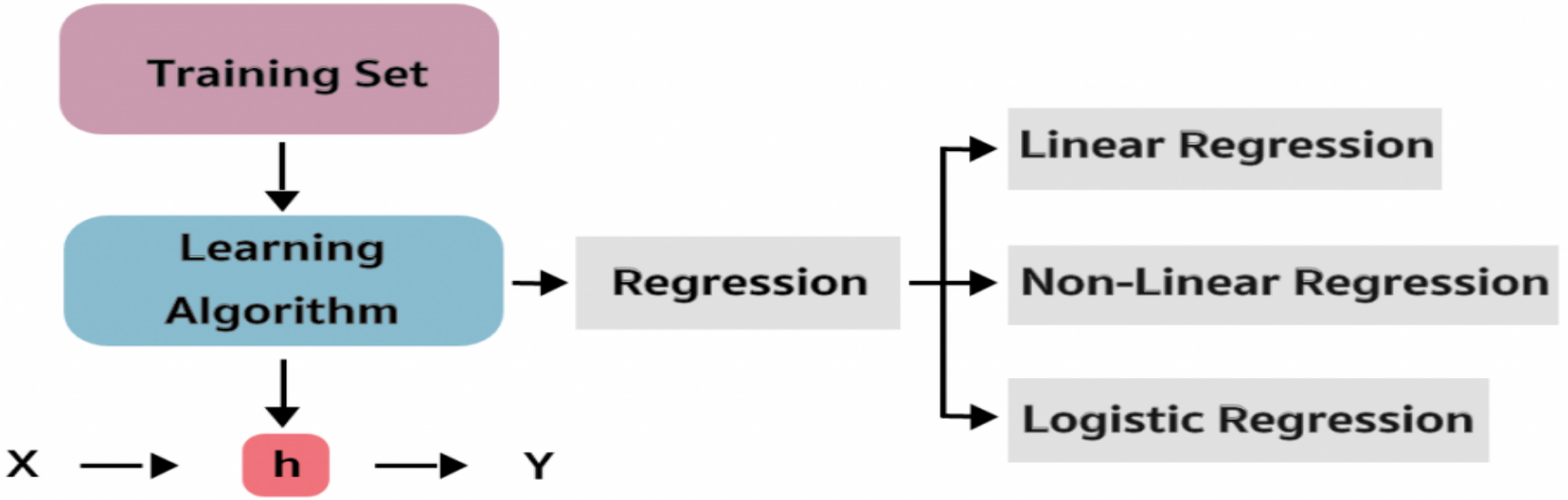
- Regression - process of predicting a continuous, numerical value for input data sample.
- Example usages: assessing the house price, forecasting grocery store food demand, temperature forecasting.



ClusterKit

https://towardsdatascience.com/coding-deep-learning-for-beginners-types-of-machine-learning-b9e651e1ed9d

Regression



ClusterKit

https://ilog.ai/linear-regression-for-ml/

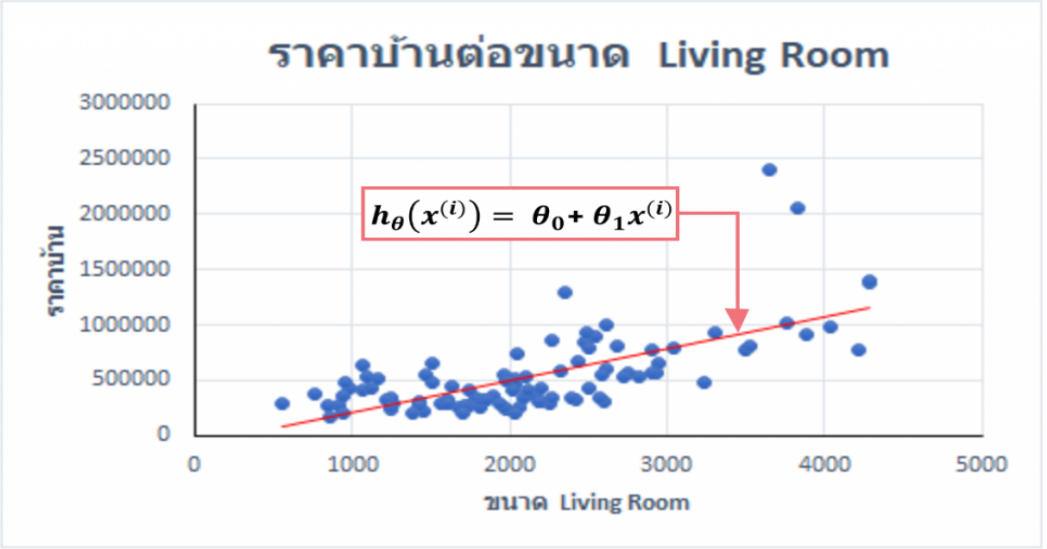
Training Set

ขนาด living (X)	ราคา (Y)
1430	310000
2950	650000
1710	233000
2320	580500
1090	535000
2620	605000
4220	775000
2250	292500
.....	.....

m = จำนวนชุดข้อมูลที่ใช้(100ชุด)

$X^{(i)}$ = ขนาด Living room(i)

$Y^{(i)}$ = ราคาบ้านที่(i)

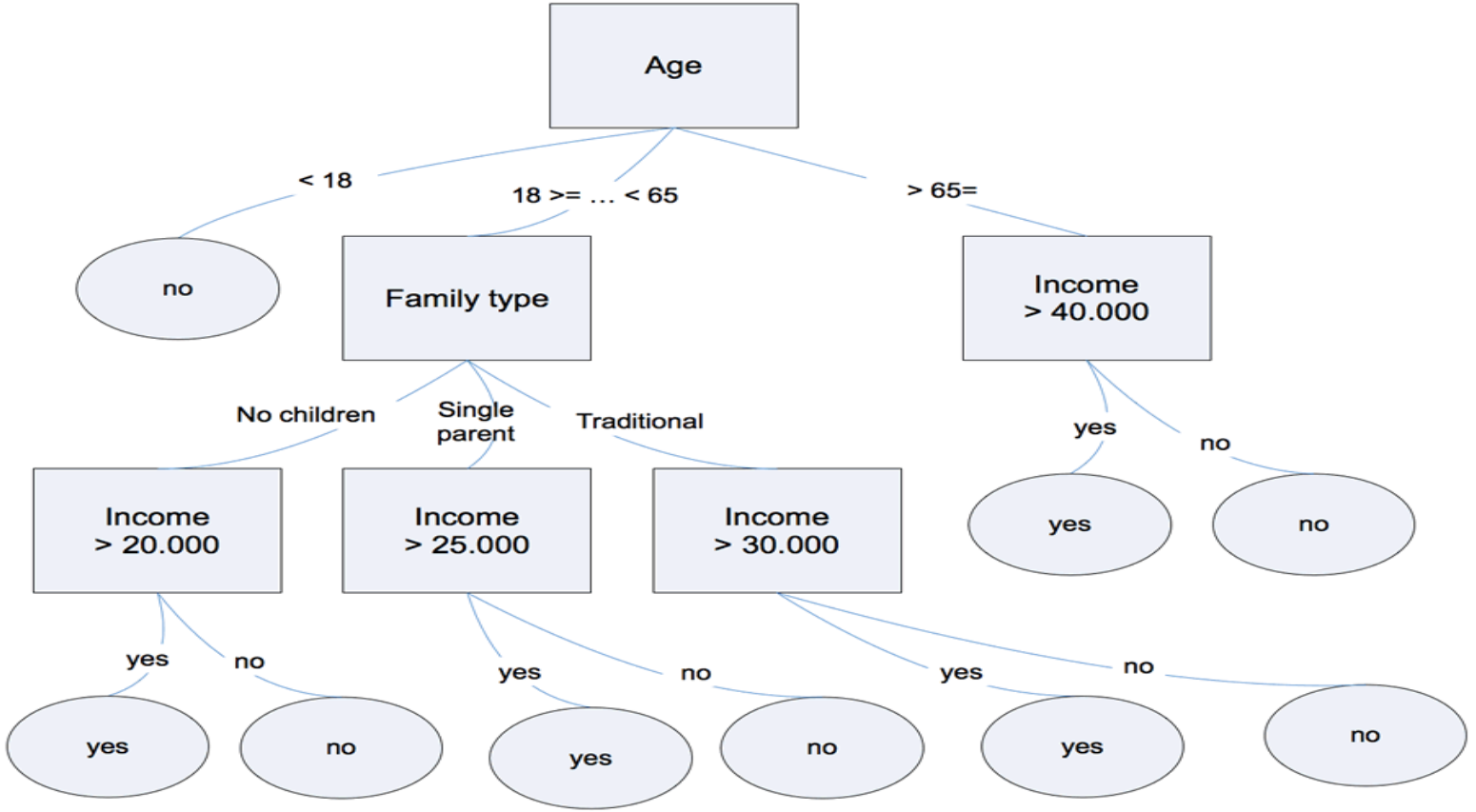


สมการเส้นตรง : $h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$

ClusterKit

https://ilog.ai/linear-regression-for-ml/

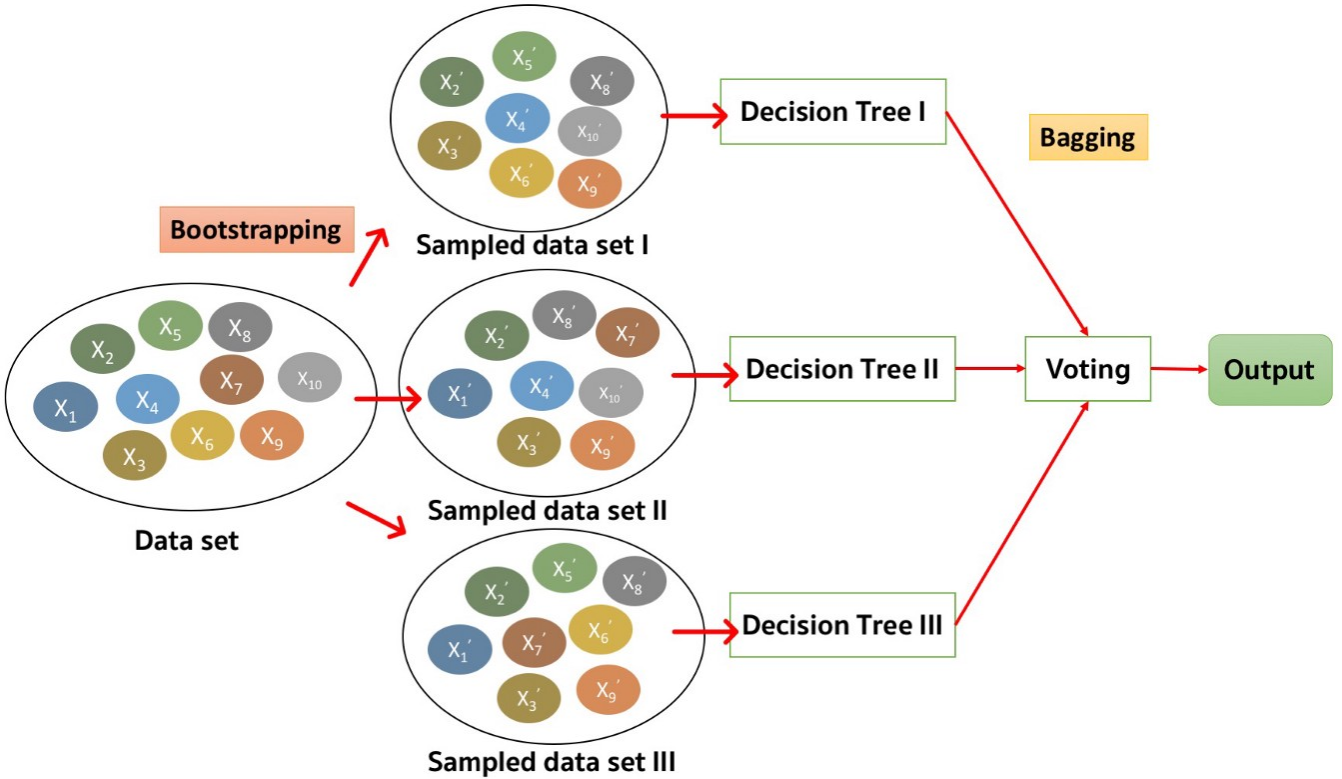
Decision Tree



ClusterKit

http://www.brcommunity.com/b624-2.htm

Random Forest



ClusterKit

https://medium.com/@witchapongdaroontham/เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c

Hands-on:

6. Regression

7. ML Pipelines

ML workflow utilities:

ML Pipelines

ML Pipelines

- A Pipeline is specified as a sequence of stages, and each stage is either a Transformer or an Estimator.
- Pipeline components
  - Transformers : take a DataFrame and output a new DataFrame
    - Feature transformers
    - Learned models : output a new DataFrame with predicted labels appended as a column
  - Estimators : any algorithm that fits or trains on data, implements a method fit(), which accepts a DataFrame and produces a model
- ML persistence: Saving and Loading Pipelines

How to create Pipelines

- Declare its stages
- Configure their parameters
- Chain them in a pipeline object

ClusterKit

Sawasdee :-)