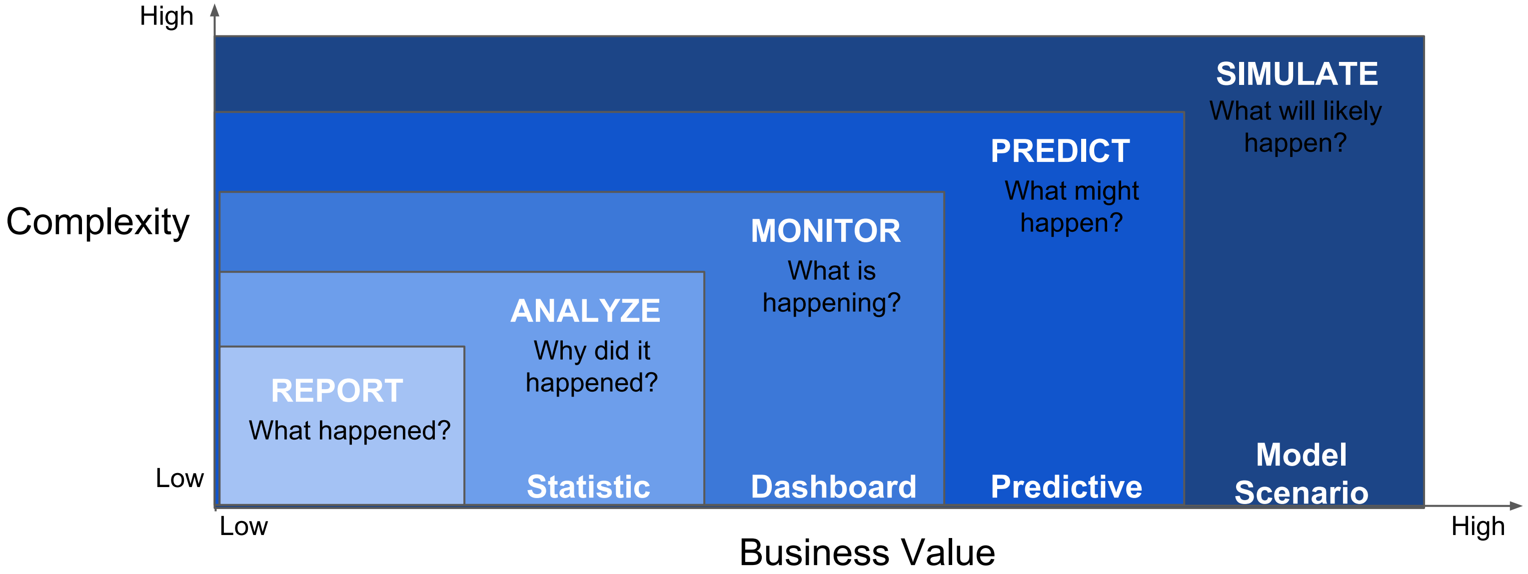
Melek Machine Learning

Data Science Indonesia

Contents

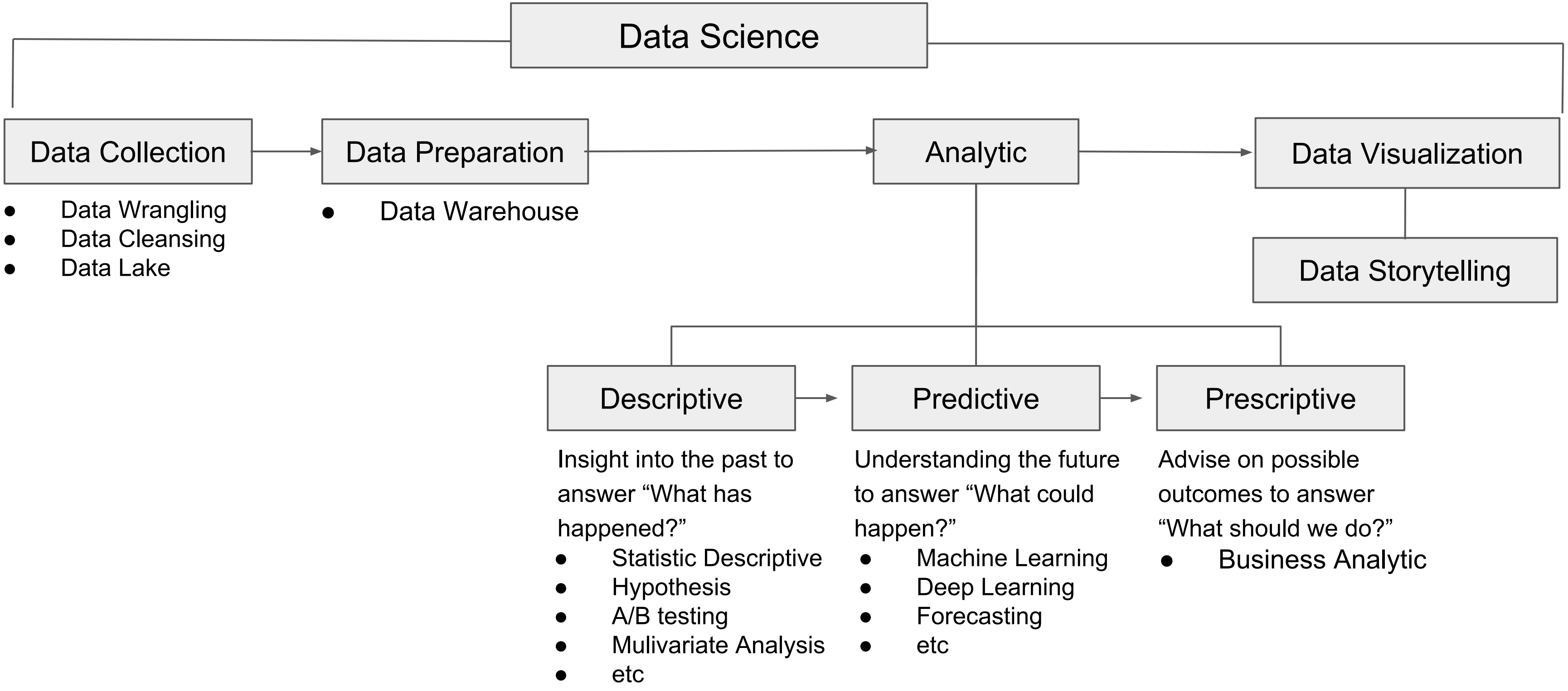
Where is the location of Machine learning?

From impact to business side:



Where is the location of Machine learning?

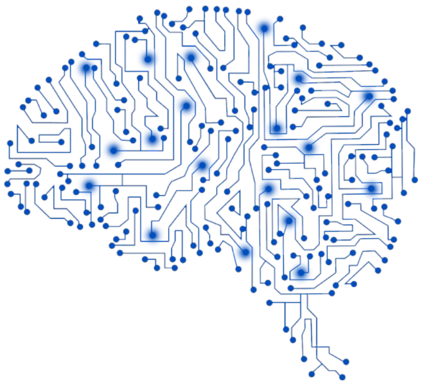
From Data Science Side:



So, What is Machine Learning?

Wikipedia:

Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.



So, What is Machine Learning?

Wikipedia:

Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.

Me:

Machine Learning is a thing labeler!

So, What is Machine Learning?

Thing labeler, talking your description of something and telling you what label it should get

Example:

We have data:

- *Kaki empat*
- *Memiliki Ekor*
- *Berbulu*
- *Suka colek kaki orang yang makan di warteg*

→ Machine Learning Model →

- 98% Probability is "Kucing"
- 65% Probability is "Anjing"
- 6% Probability is "Kuda nil"

What should label that appropriate with this object?

So, What is Machine Learning?

why you should be excited about thing labeler

What is this animal?

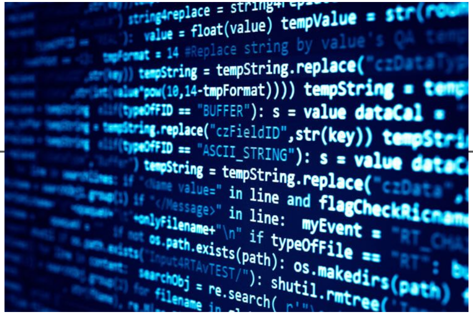
So, What is Machine Learning?

Why you should be excited about thing labeler

Kucing!

So, What is Machine Learning?

Why you should be excited about thing labeler

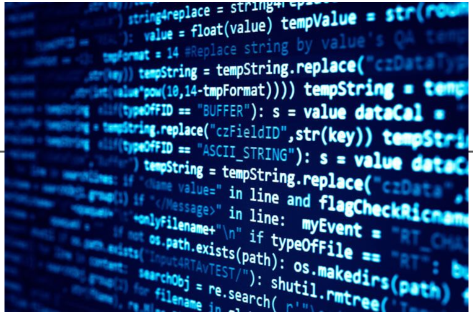


Kucing!

Machine learning is a new programming paradigm, a new way of communicating your wishes to a computer.

So, What is Machine Learning?

Why you should be excited about thing labeler



Kucing!

Explain with examples (data), not instructions

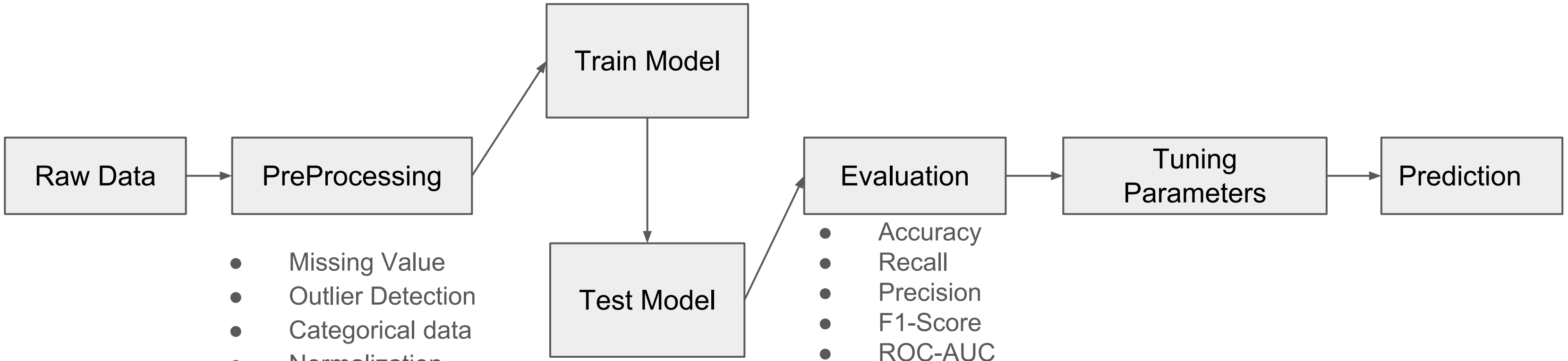
Why Machine Learning?

Machine Learning is used to solve business problem

My boss asked me that he wants to:

- wants to know about our customer behavior ---> *ML can help you!*
- wants to know which customer dare to pay expensive ---> *ML can help you!*
- wants to know customer who will stop using our product ---> *ML can help you!*
- wants to sell the same product with different price on each customer at the same time ---> *ML can help you!*
- wants to know the customer who will pay the credit until the end ---> *ML can help you!*
- wants to recommend products according to the needs of every customer ---> *ML can help you!*
- wants to know if there is fraud to the customer in using our product ---> *ML can help you!*
- wants to know the market potential for new innovation products ---> *ML can help you!*
- And many more

How to build Machine Learning Models?



- Missing Value
- Outlier Detection
- Categorical data
- Normalization

- Accuracy
- Recall
- Precision
- F1-Score
- ROC-AUC
- Log-loss
- Gini index
- F-Score
- MSE
- MAPE
- etc

What are kinds of Machine Learning model?

Machine Learning

Supervised

Classification

- Naive Bayes
- Decision Trees
- Ensemble Methods (Bagging, AdaBoost, Random Forest, Gradient Boosting)
- K-Nearest Neighbors
- Stochastic Gradient Descent
- Support Vector Machines (SVM)
- Logistic Regression

Regression

- Linear Regression
- Polynomial Regression
- Decision Trees Regressor
- Random Forest Regressor
- Support Vector Regressor

Unsupervised

- K-Means
- Gaussian Mixture Model
- Principal Component Analysis
- Hierarchical Clustering
- DBScan

Reinforcement Learning

Supervised Learning

Supervised Learning

What is Supervised Learning?

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Algorithms:

- K-Nearest Neighbor
- Decision Tree
- Support Vector Machine

Supervised Learning

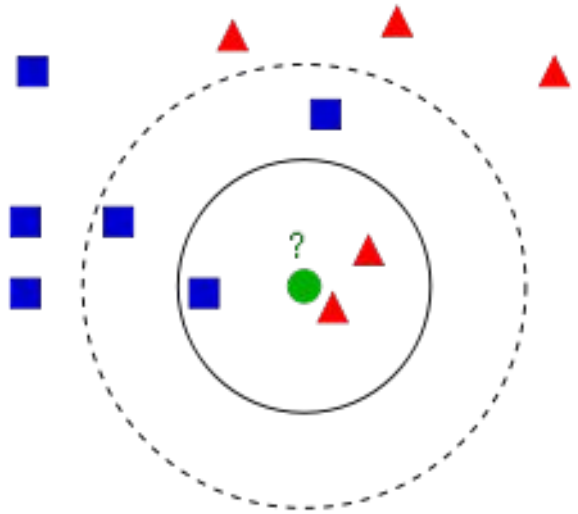
- Regression
- Classification

Supervised learning in classification handles estimator for discrete labels rather than continuous labels.

Classification > binary classification, multiclass classification, multilabel classification

K-Nearest Neighbor

- K-NN is an instance-based learning or lazy learning
- K-NN using distance to measure the likelihood of the class
- Number of K take as a comparison of the likelihood



Pros and Cons

Pros

- Simple Algorithm
- Versatile (Classification and Regression)
- Does not assume any probability distribution on the input data

Cons

- Requires high memory
- Computationally Expensive
- Lazy Learning

Decision Tree

Metrics that DT consider:

- Information Gain
- Entropy

Algorithms:

- ID3
- C4.5
- C5.0
- CART

Decision Tree

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Step 1: Calculate Entropy

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)***E**(3,2) + **P**(Overcast)***E**(4,0) + **P**(Rainy)***E**(2,3)

= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971

= 0.693

Step 2: Calculate Information Gain

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

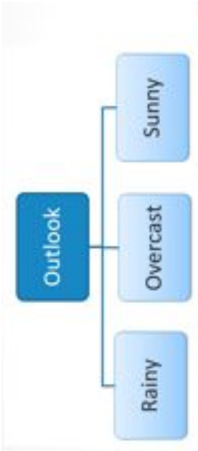
		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

Step 3: Choose Root Node

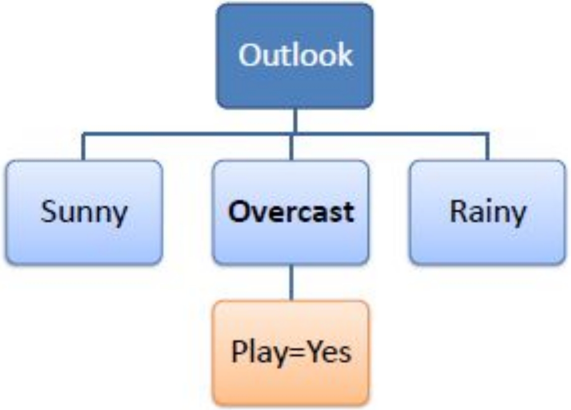
		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			



Outlook	Temp	Humidity	Windy	Play Golf
Sunny	Mild	High	FALSE	Yes
Sunny	Cool	Normal	FALSE	Yes
Sunny	Cool	Normal	TRUE	No
Sunny	Mild	Normal	FALSE	Yes
Sunny	Mild	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Hot	High	FALSE	No
Rainy	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	No
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	TRUE	Yes

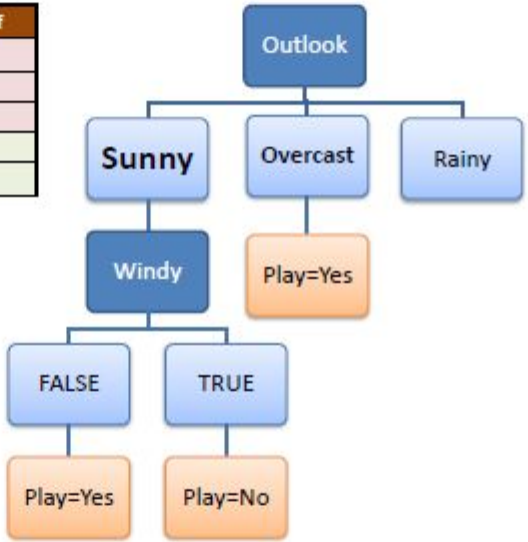
Step 3: Node Leaf

Temp	Humidity	Windy	Play Golf
Hot	High	FALSE	Yes
Cool	Normal	TRUE	Yes
Mild	High	TRUE	Yes
Hot	Normal	FALSE	Yes



Step 4: Calculate Entropy & Information Gain

Temp	Humidity	Windy	Play Golf
Mild	High	FALSE	Yes
Cool	Normal	FALSE	Yes
Mild	Normal	FALSE	Yes
Cool	Normal	TRUE	No
Mild	High	TRUE	No



Step 5: Calculate Information Gain

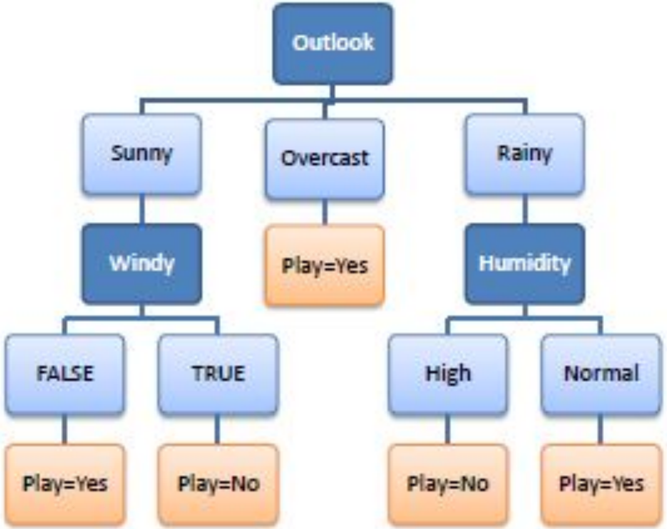
R_1: IF (Outlook=Sunny) AND (Windy=FALSE) THEN Play=Yes

R_2: IF (Outlook=Sunny) AND (Windy=TRUE) THEN Play=No

R_3: IF (Outlook=Overcast) THEN Play=Yes

R_4: IF (Outlook=Rainy) AND (Humidity=High) THEN Play=No

R_5: IF (Outlook=Rain) AND (Humidity=Normal) THEN Play=Yes



Pros and Cons

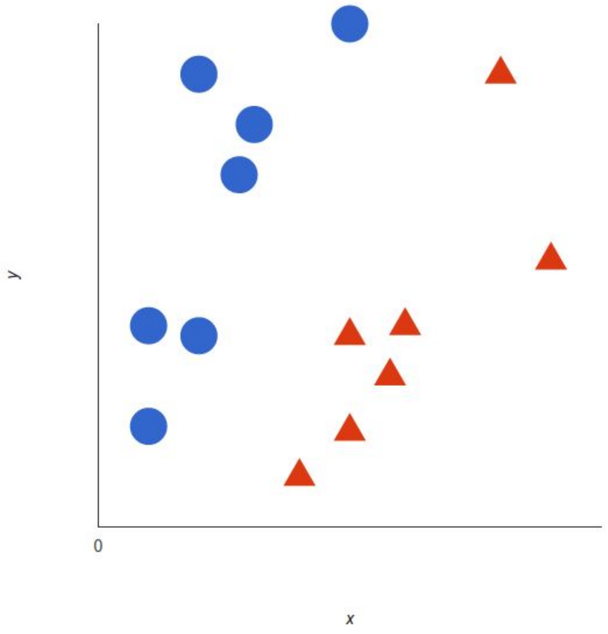
Pros:

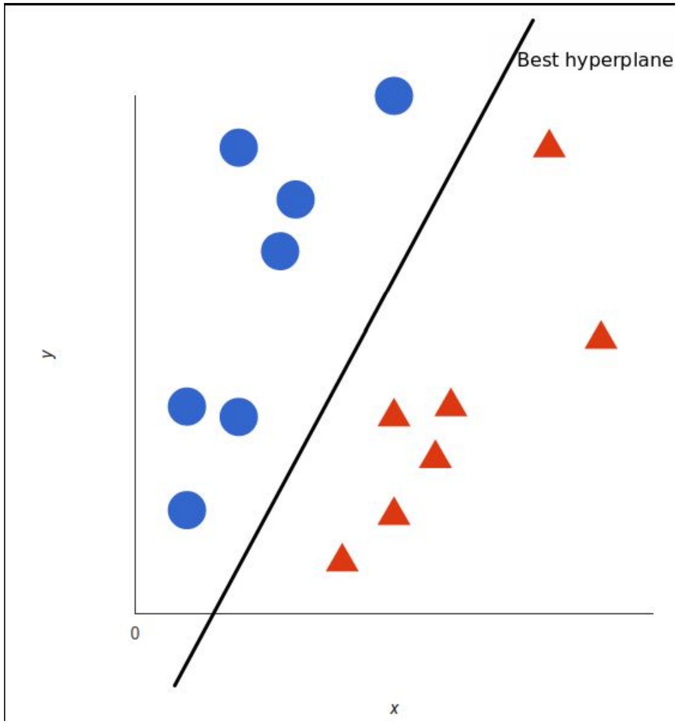
- Easy to Explain
- Follows the same approach with human
- Interpretation of a complex decision tree can be simplified by visualization

Cons:

- High probability of overfitting
- Calculations can be complex when there are many class labels

Support Vector Machine





Pros and Cons

Pros:

- Works well with clear margin
- Effective in High Dimensional Spaces
- Effective in cases where num. Of dimension is greater than num. Of samples

Cons:

- High computation
- Bad at a lot of noise

Let’s Code

Unsupervised Learning

Unsupervised Learning

What is Unsupervised Learning?

Algorithms:

- K-Means
- Hierarchy clustering
- DBSCAN

K-Means

K-Means is the ‘go-to’ clustering algorithm for many simply because it is fast, easy to understand, and available everywhere (there’s an implementation in almost any statistical or machine learning tool you care to use) <- Really Suitable with large data

“Here’s a list of 100,000 warehouses full of data. I’d like you to condense them down to one meaningful warehouse.”

Pros and Cons

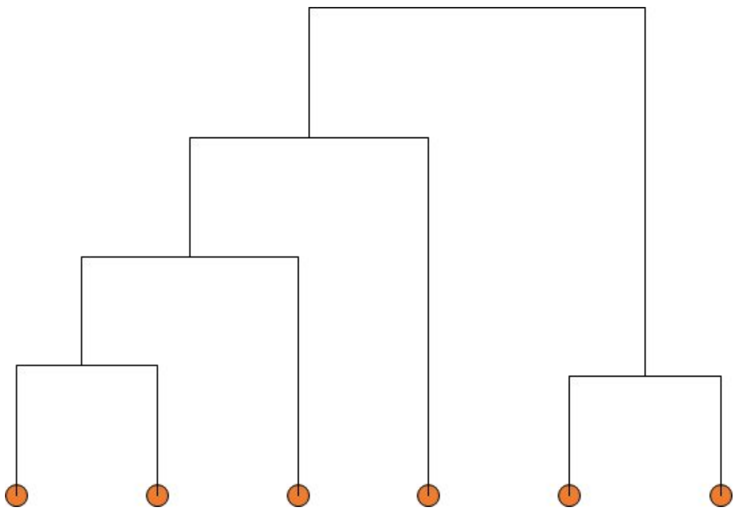
- Applicable only when mean is defined, then what about categorical data?
- Need to specify k, the number of clusters, in advance. If you have a good intuition for how many clusters the dataset your exploring has then great, otherwise you might have a problem
- Unable to handle noisy data and outliers
- Not suitable to discover clusters with non-convex shapes
- Doesn’t consider the proportion of different cluster
- Doesn’t consider the variance of different cluster

Hierarchy Clustering

Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion

In short, decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.

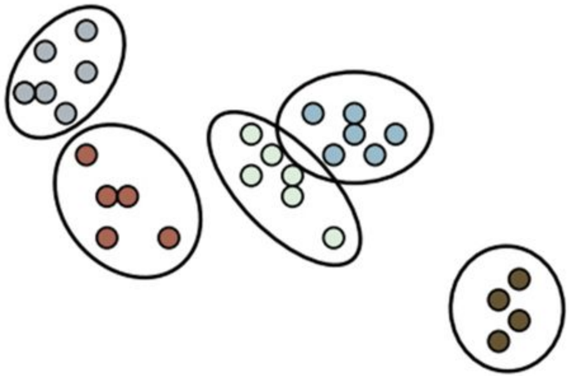


Pros and Cons

- do not scale well: time complexity of at least O(n^2), where n is the number of total objects
- can never undo what was done previously
- Similar to K-Means we are stuck choosing the number of clusters (not easy in EDA), or trying to discern some natural parameter value from a plot that may or may not have any obvious natural choices.

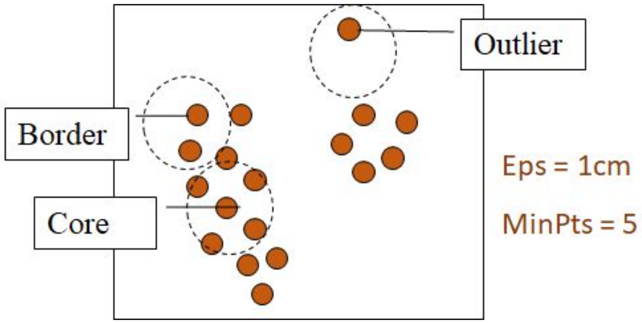
DBSCAN

DBSCAN is a density based algorithm – it assumes clusters for dense regions. It is also the first actual clustering algorithm we've looked at: it doesn't require that every point be assigned to a cluster and hence doesn't partition the data, but instead extracts the 'dense' clusters and leaves sparse background classified as 'noise'.



Pros and Cons

- Epsilon is a distance value, so you need to survey the distribution of distances in your dataset to attempt to get an idea of where it should lie. In practice, however, this isn't an especially intuitive parameter, nor is it easy to get right.
- Sensitive to the parameter



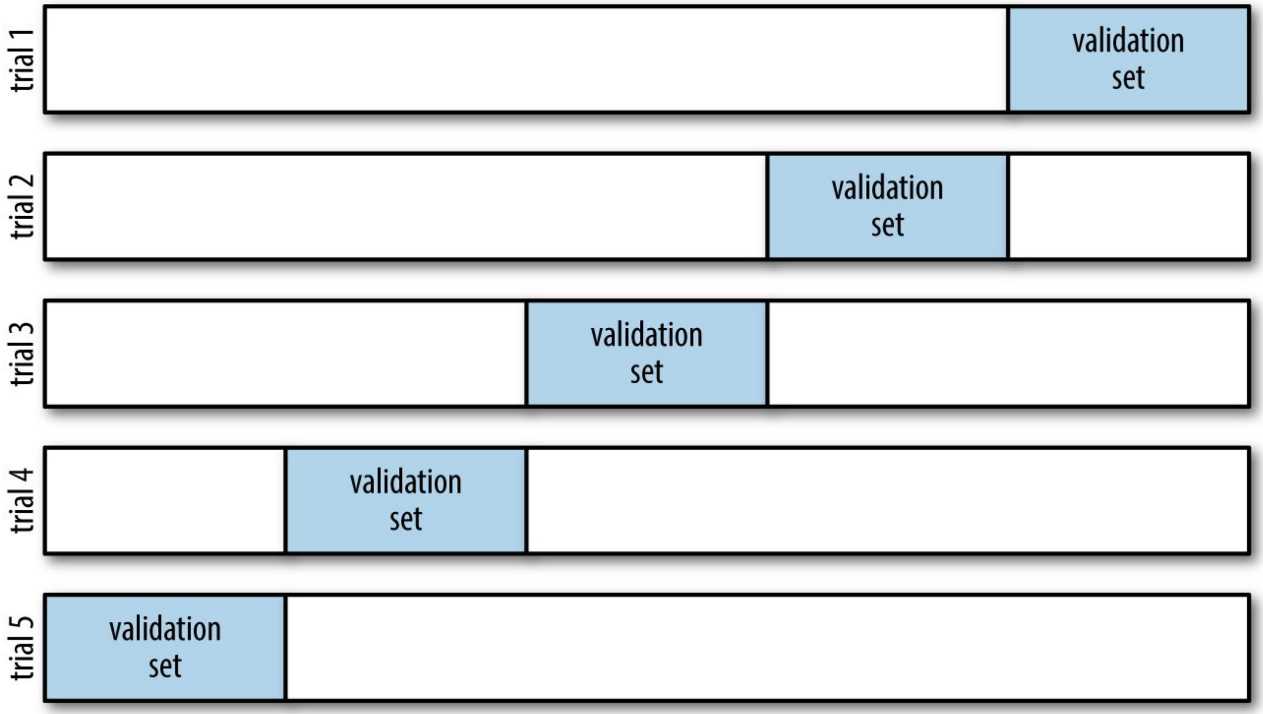
Let’s Code

Evaluation

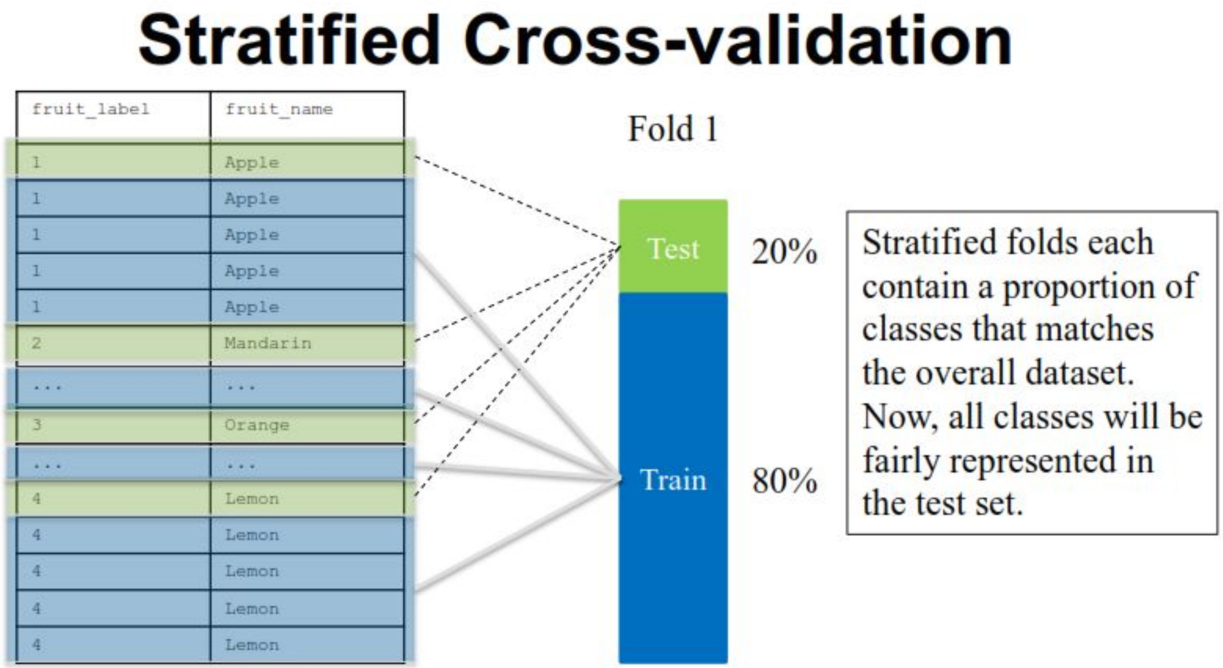
Evaluation - Classification

- Cross Validation
  - Stratified Cross Validation
  - One Leave Out Cross Validation
- Metrices
  - Accuracy
  - Precision
  - Recall
  - F1-Score

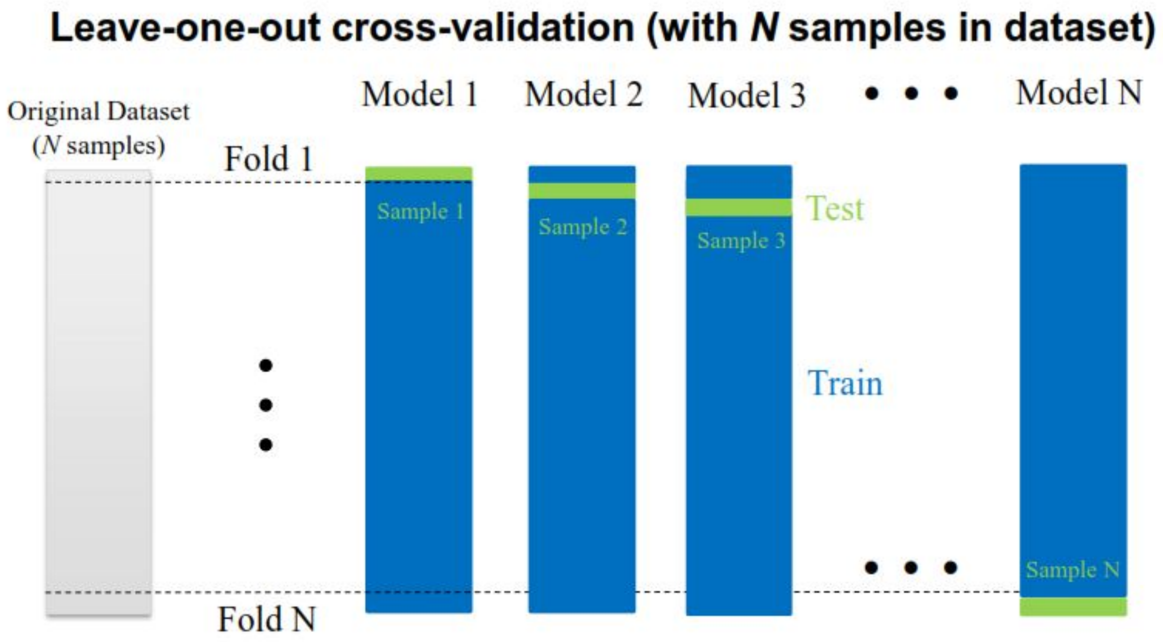
Cross Validation



Stratified Cross Validation



Leave One Out Cross Validation



Confusion Matrix

Binary prediction outcomes

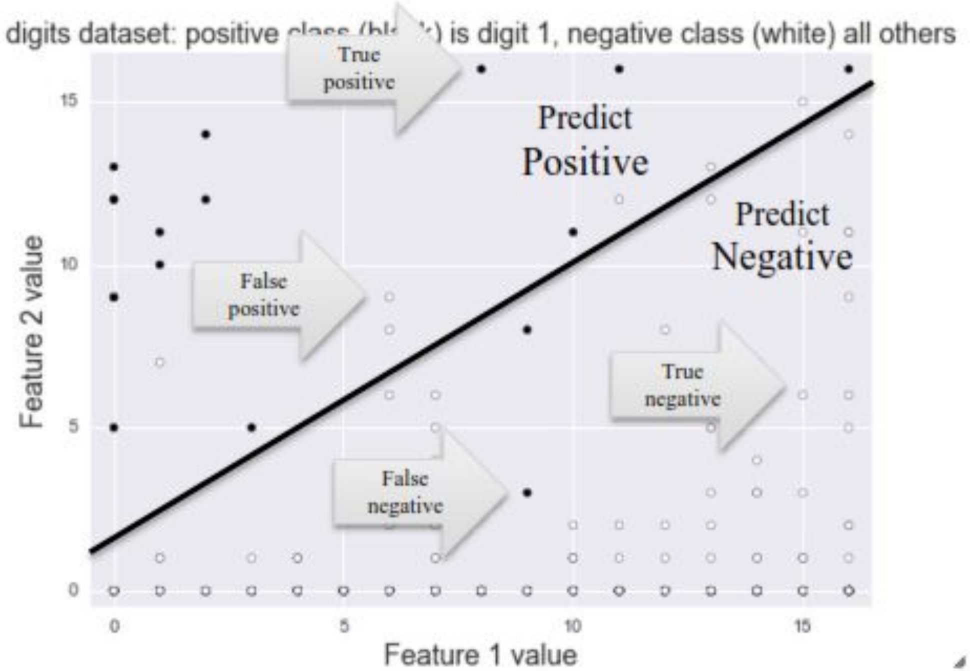
	Predicted negative	**Predicted** positive
True negative	TN	FP
True positive	FN	TP

Label 1 = positive class
(class of interest)

Label 0 = negative class
(everything else)

TP = true positive
FP = false positive (Type I error)
TN = true negative
FN = false negative (Type II error)

Confusion Matrix Visualization



Accuracy

	Predicted negative	Predicted positive	
True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
			$N = 450$

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP}$$

$$= \frac{400+26}{400+26+17+7}$$

$$= 0.95$$

Recall

True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
	Predicted negative	Predicted positive	$N = 450$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$= \frac{26}{26+17}$$

$$= 0.60$$

Recall is also known as:

- True Positive Rate (TPR)
- Sensitivity
- Probability of detection

Precision

	Predicted negative	Predicted positive	
True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
			$N = 450$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$= \frac{26}{26+7}$$

$$= 0.79$$

F1-Score

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}$$

Evaluation - Clustering

Tuning Parameter (Decision Tree)

max_depth

The first parameter to tune is max_depth. This indicates how deep the tree can be. The deeper the tree, the more splits it has and it captures more information about the data.

min_samples_split

min_samples_split represents the minimum number of samples required to split an internal node. This can vary between considering at least one sample at each node to considering all of the samples at each node. When we increase this parameter, the tree becomes more constrained as it has to consider more samples at each node.

Tuning Parameter (Decision Tree)

min_samples_leaf

min_samples_leaf is The minimum number of samples required to be at a leaf node. This parameter is similar to *min_samples_splits*, however, this describe the minimum number of samples of samples at the leafs, the base of the tree.

max_features

max_features represents the number of features to consider when looking for the best split.

Analytics

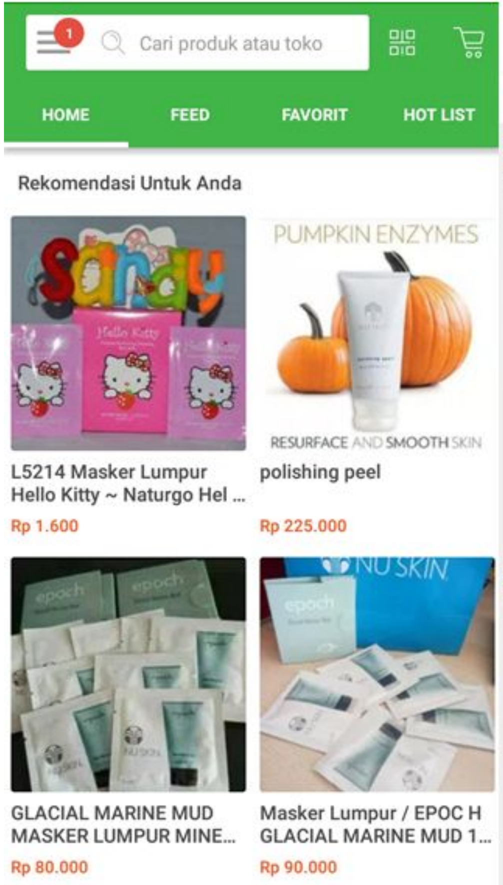
Machine Learning Use Cases

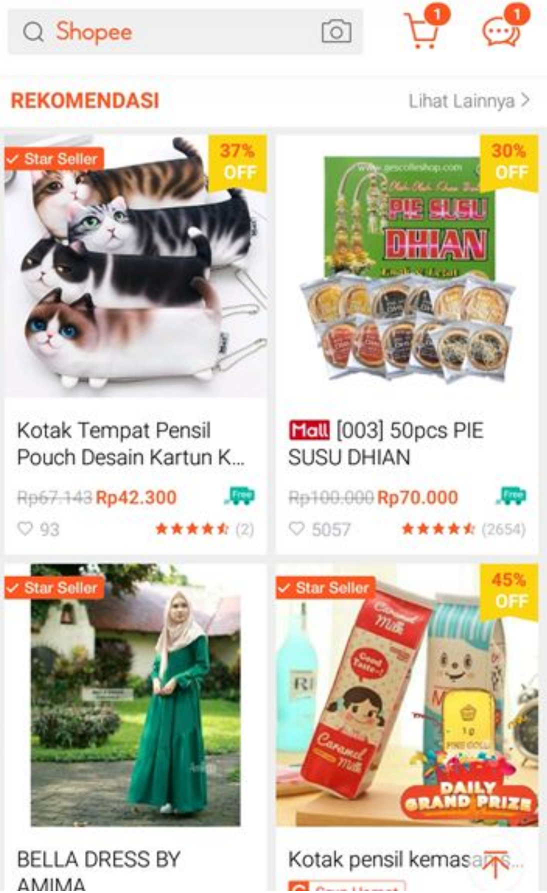
- Supervised Learning
    a. Use case 1
    b. Use case 2
    c. Use case 3
- Unsupervised Learning
    a. Use case 1
    b. Use case 2
    c. Use case 3

Use Case Machine Learning

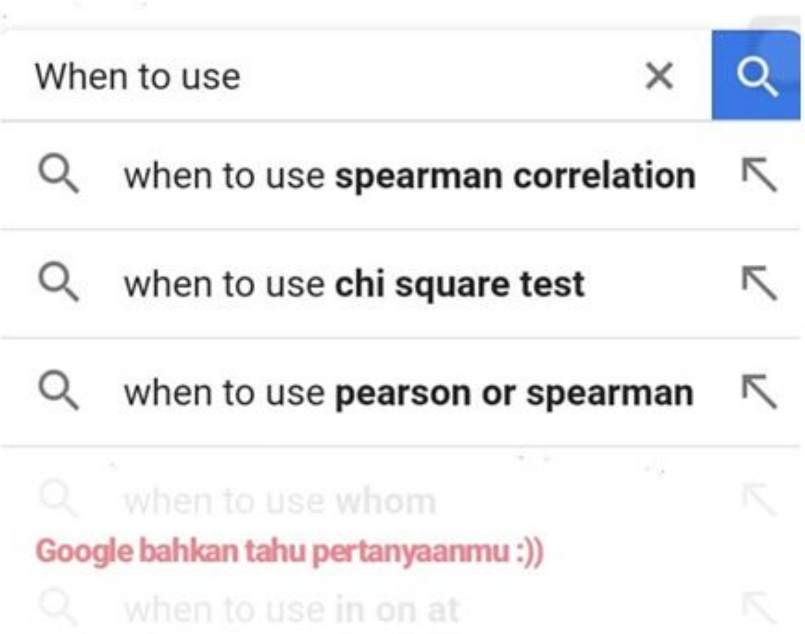
How machine learning implemented in real case

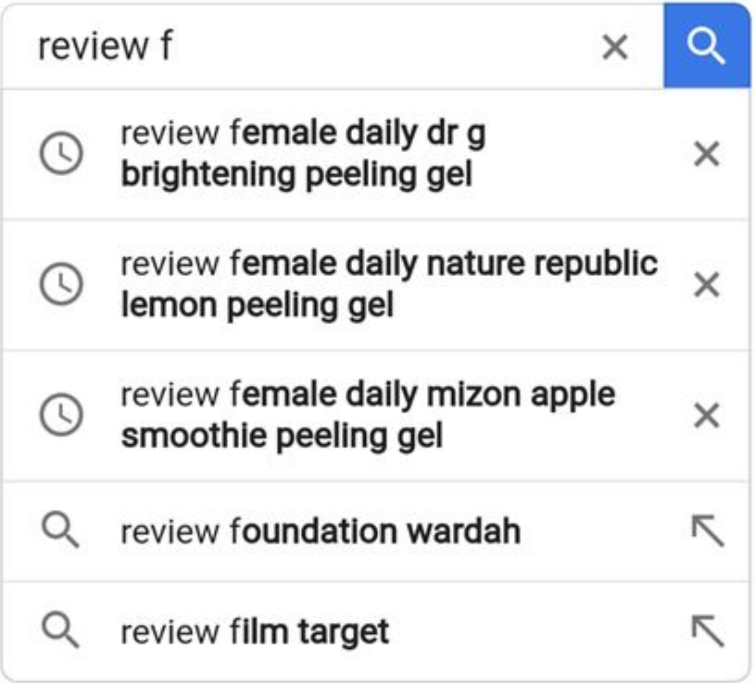
Rekomendasi



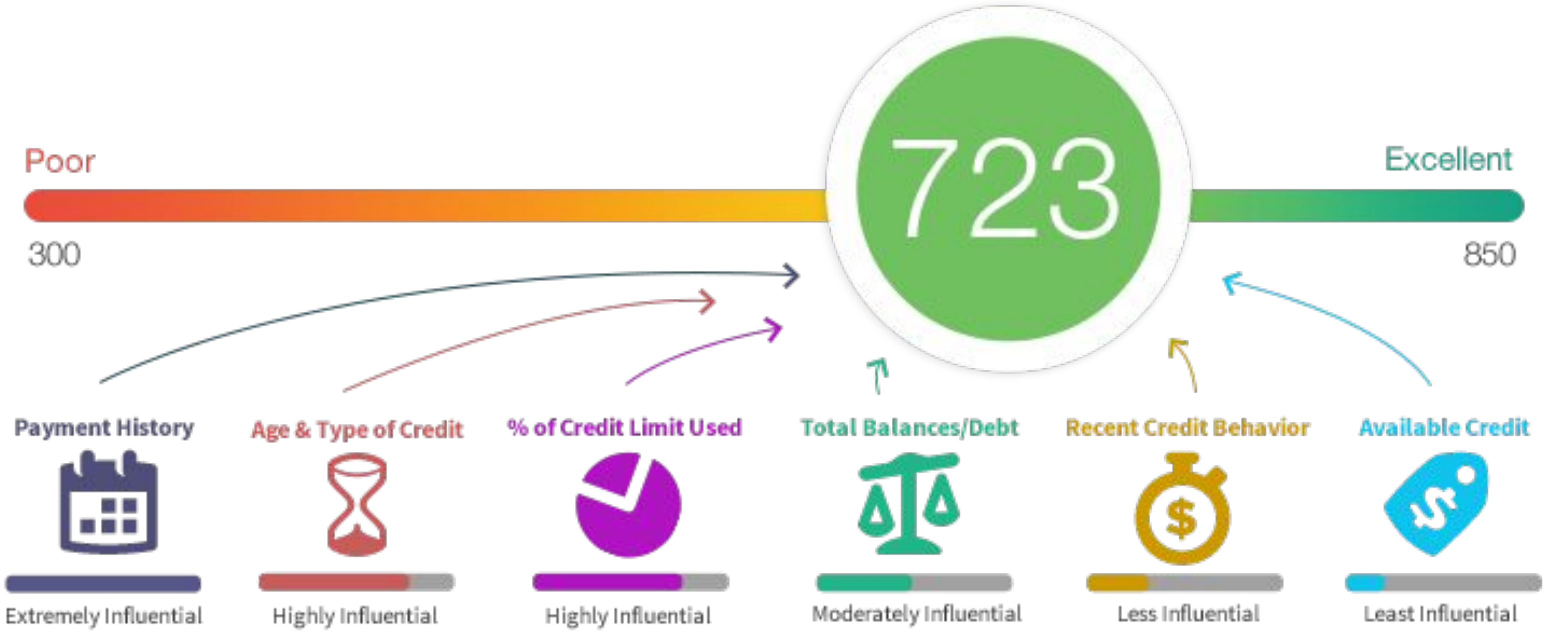


Google Search





Credit Scoring



Bot


cleverbot
36781 people talking
Gak tau.
Nama saya ysabel yang pelakon tu.
Maem sek yo.
Kepala hotak hang.
Hmm.
What are you thinking? share!
say to cleverbot...
think about it
think for me
thoughts so far


SimSimi
Talk
Teach
Quotes
Hi~!
simi, lo kenal sm kaka gw g'?
kenal, jomblo ngenes itu kan, kasian bgt tuh dia, gak prnh pacaran :P
SimSimi learned 14,025,934 sentences.
Bahasa Indc
Send

Study Case

Data Science Indonesia

Data Science Indonesia (DSI), adalah sebuah Komunitas yang didirikan pada bulan Mei 2015 yang terdiri dari sekumpulan ilmuwan, seniman dan pembelajar yang ingin membangun budaya Data Driven di Indonesia dengan menginspirasi, mengajarkan serta menawarkan nilai dari sebuah data melalui pendekatan Data Science

Visi Kami:

•Bersama masyarakat menciptakan ekosistem inovasi berbasis data untuk meningkatkan kesejahteraan masyarakat

Misi :

•Menjadi mitra bagi sektor publik maupun swasta untuk mengeskplorasi pendekatan data science dalam mencari solusi atas tantangan yang ada

•Meningkatkan pengetahuan dan kesadaran masyarakat terhadap data science

•Menjadi wadah bagi masyarakat untuk berjejaring dalam konteks pemanfaatan data science

DATA SCIENCE INDONESIA

http://datascience.or.id/daftar-anggota-dsi/

Contact:

contact@datascience.or.id