DATA MINING PROCESS

Data Preparation

- Data preparation is a self-service activity that converts disparate, raw, messy data into a clean and consistent view. The process includes searching, cleaning, transforming, organizing and collecting data. Preparing data is critical but time-intensive; data teams spend up to 80% of their time converting raw data into high quality, analysis-ready output. (Source: https://www.ibm.com/my-en/analytics/data-preparation)

- Data adalah informasi yang berhasil di catat/ di rekam.
- Data dibedakan menjadi data terstruktur dan data tidak terstruktur (gambar, teks, suara).
- Data juga dapat dibagi menjadi data kualitatif/kategori dan data kuantitatif.

Data Preparation

Pada tahap data preparation, beberapa hal yang perlu dilakukan sebelum tahap analisis lebih jauh: (proses ini disebut tahap data pre-processing)

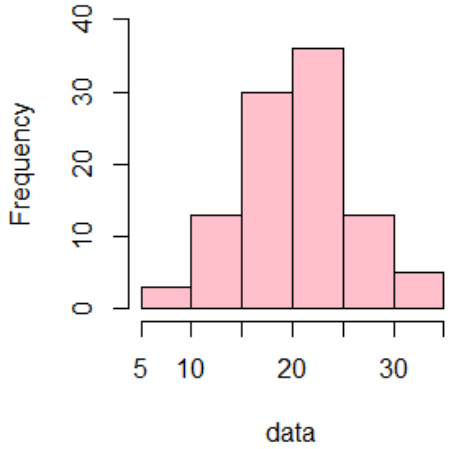
- Cek dimensi / ukuran data (banyaknya variabel dan observasi)
- Cek apakah perlu dilakukan reduksi dimensi
- Cek kelengkapan data (missing value)
- Cek adakah data pencilan (outlier)
- Cek adakah duplikasi data, salah penamaan data, data salah tempat, salah entry data.
- Cek apakah perlu dilakukan transformasi / normalisasi pada data

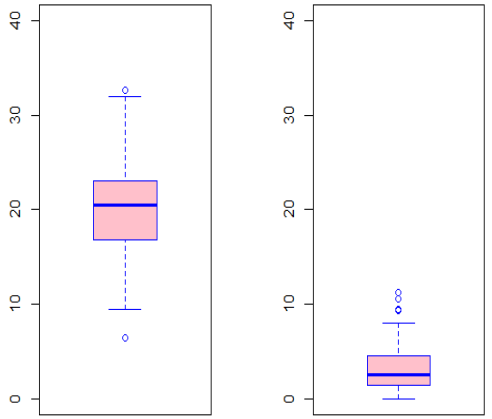
Untuk memudahkan pekerjaan ini, bisa dilakukan dengan membuat plot (visualisasi data).

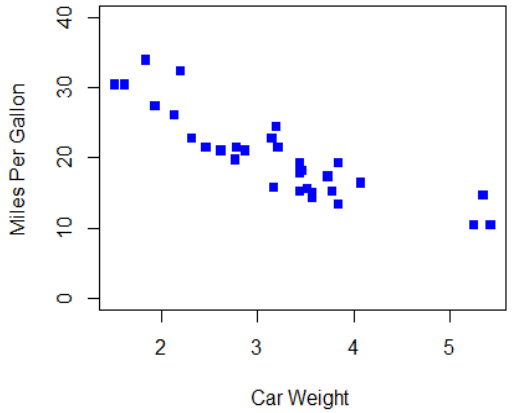
Penyajian Data secara Visual

Kontinyu

- Histogram
- Box Plot
- Scatter plot

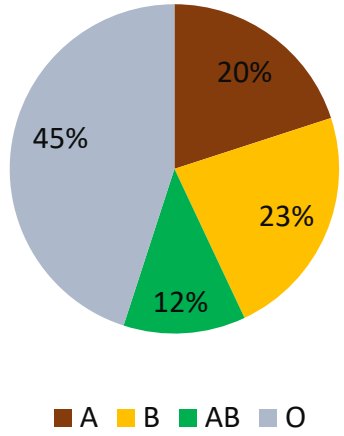


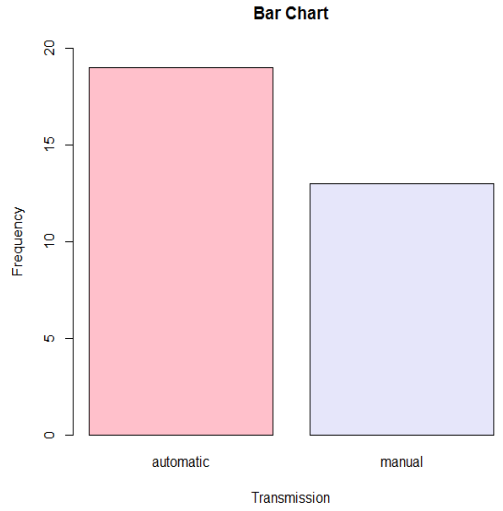




Kategorik

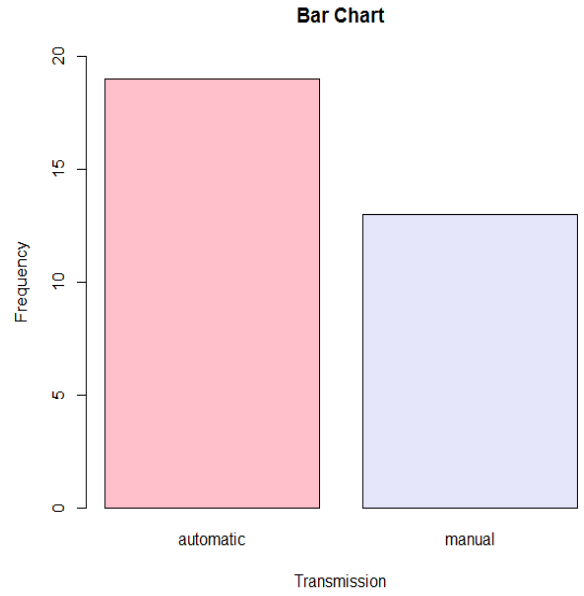
- Bar Chart
- Pie Chart





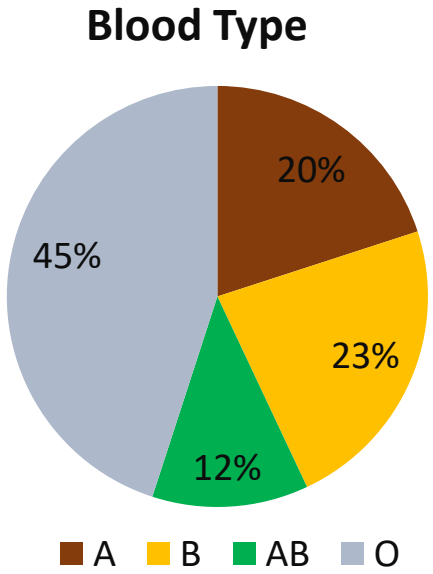
Bar Chart

- Menggunakan persegi panjang horizontal atau vertikal sering digunakan untuk menampilkan data kualitatif/kategorikal dan tidak ada penekanan persentase total yang diwakili setiap kategori.
- Panjang batang menunjukkan frekuensi data.
- Skala pengukurannya nominal atau ordinal.

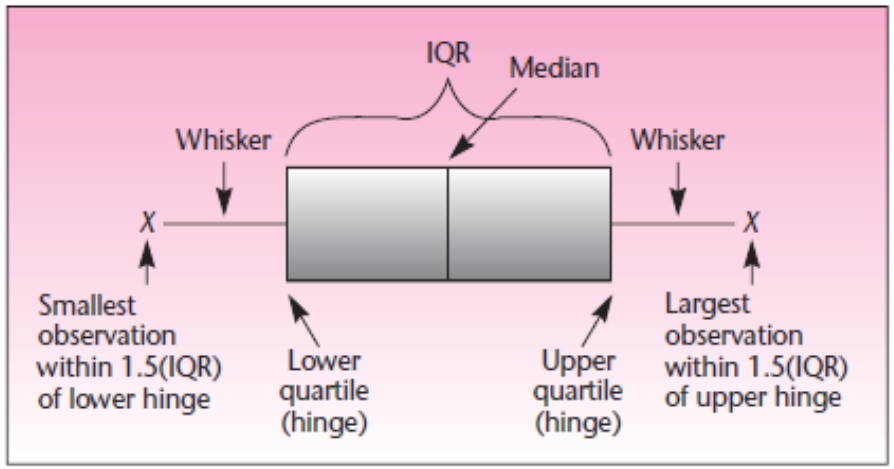


Pie Chart

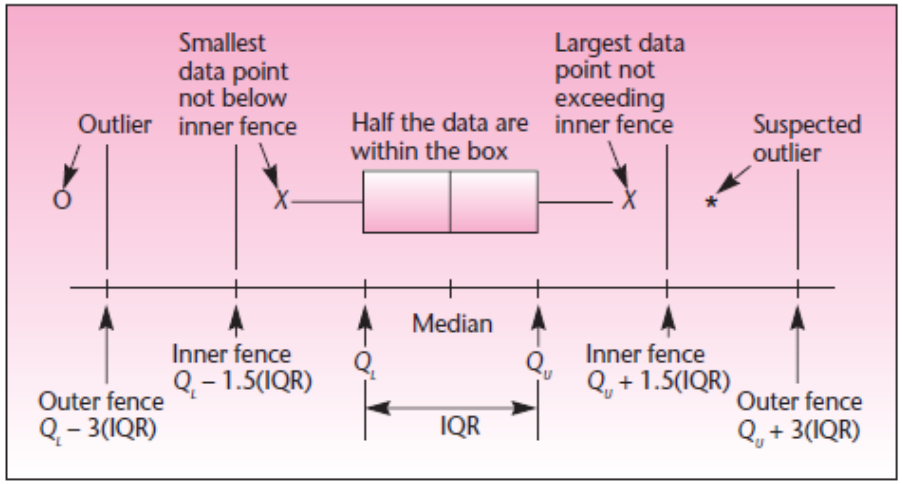
- Tampilan deskriptif sederhana dari data kualitatif yang menjumlahkan total tertentu.
- Luas lingkaran mewakili 100% (dari semua kategori)
- Ukuran setiap irisan adalah persentase dari total yang diwakili oleh kategori.
- Skala pengukuran bisa nominal atau ordinal.

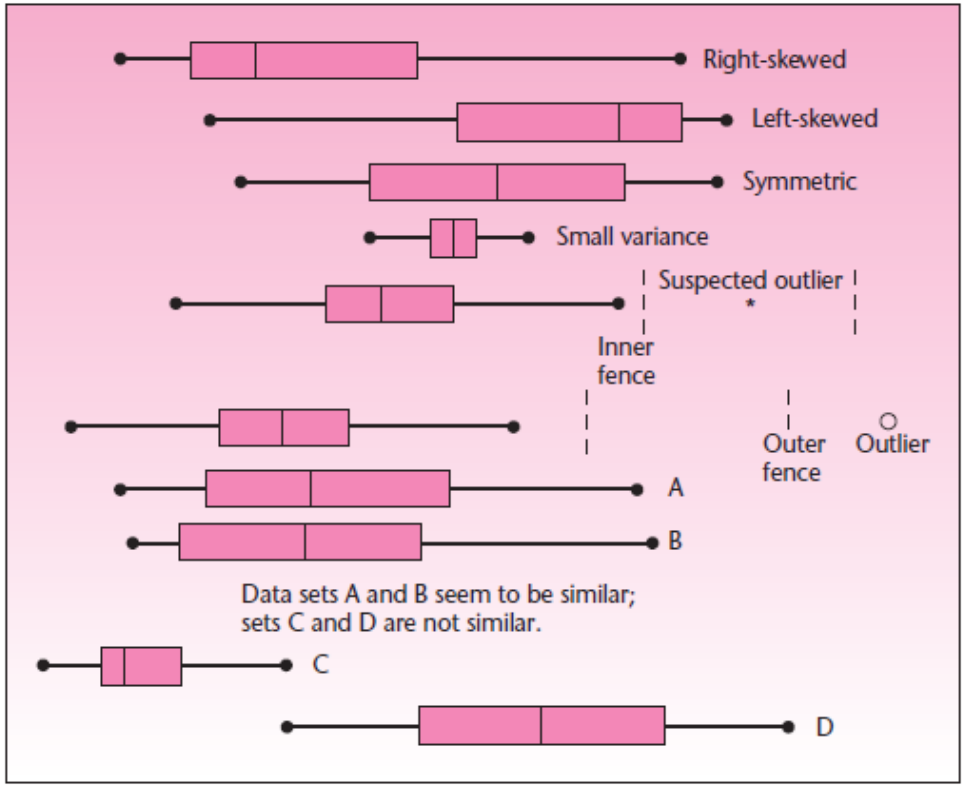


Box Plot



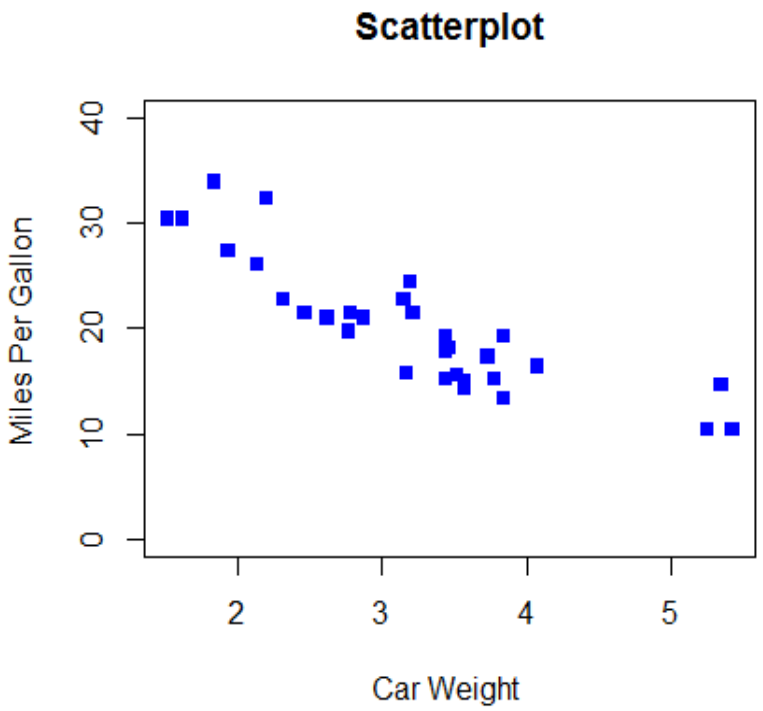
Komponen Box Plot





Aczel dan Sounderpandian (2008)

Scatter Plot

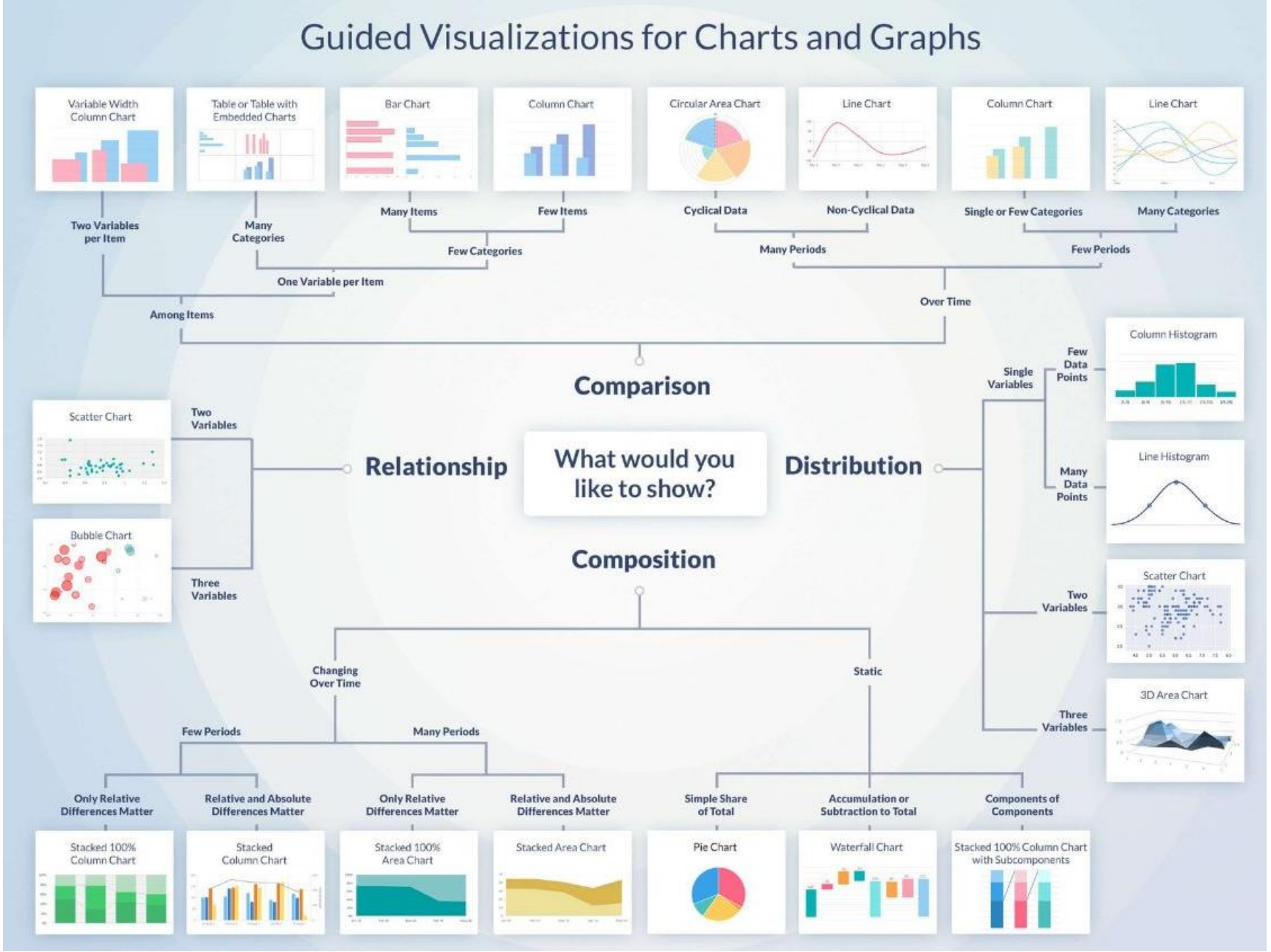


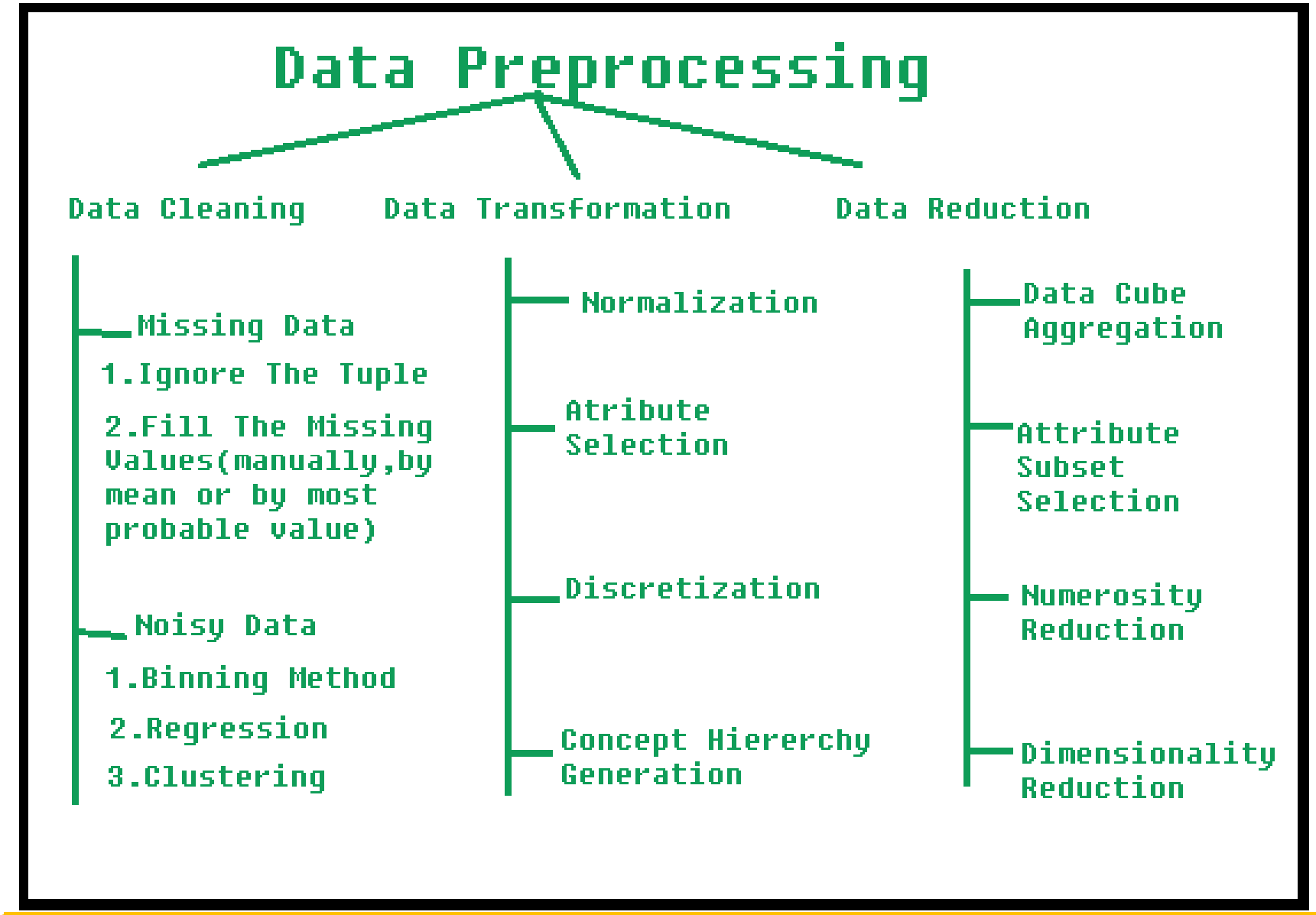

```
> cor(wt,mpg)
[1] -0.8676594
```

Data mtcars terdiri dari 11 variabel, 32 observasi

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

Scatter plot digunakan untuk mengidentifikasi dan menunjukkan hubungan di antara pasangan kumpulan data. Plot terdiri dari titik-titik yang tersebar mewakili pengamatan. Sebuah titik diberi tanda pada plot pada koordinat (x, y).


Guided Visualizations for Charts and Graphs
What would you like to show?
Comparison
Relationship
Distribution
Composition
Variable Width Column Chart
Table or Table with Embedded Charts
Bar Chart
Column Chart
Circular Area Chart
Line Chart
Column Chart
Line Chart
Two Variables per Item
Many Categories
Many Items
Few Items
Few Categories
One Variable per Item
Among Items
Cyclical Data
Non-Cyclical Data
Many Periods
Single or Few Categories
Many Categories
Few Periods
Over Time
Scatter Chart
Bubble Chart
Two Variables
Three Variables
Column Histogram
Line Histogram
Scatter Chart
3D Area Chart
Single Variables
Few Data Points
Many Data Points
Two Variables
Three Variables
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Stacked 100% Column Chart
Stacked Column Chart
Stacked 100% Area Chart
Stacked Area Chart
Pie Chart
Waterfall Chart
Stacked 100% Column Chart with Subcomponents



Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Source: https://www.geeksforgeeks.org/

Data Cleaning

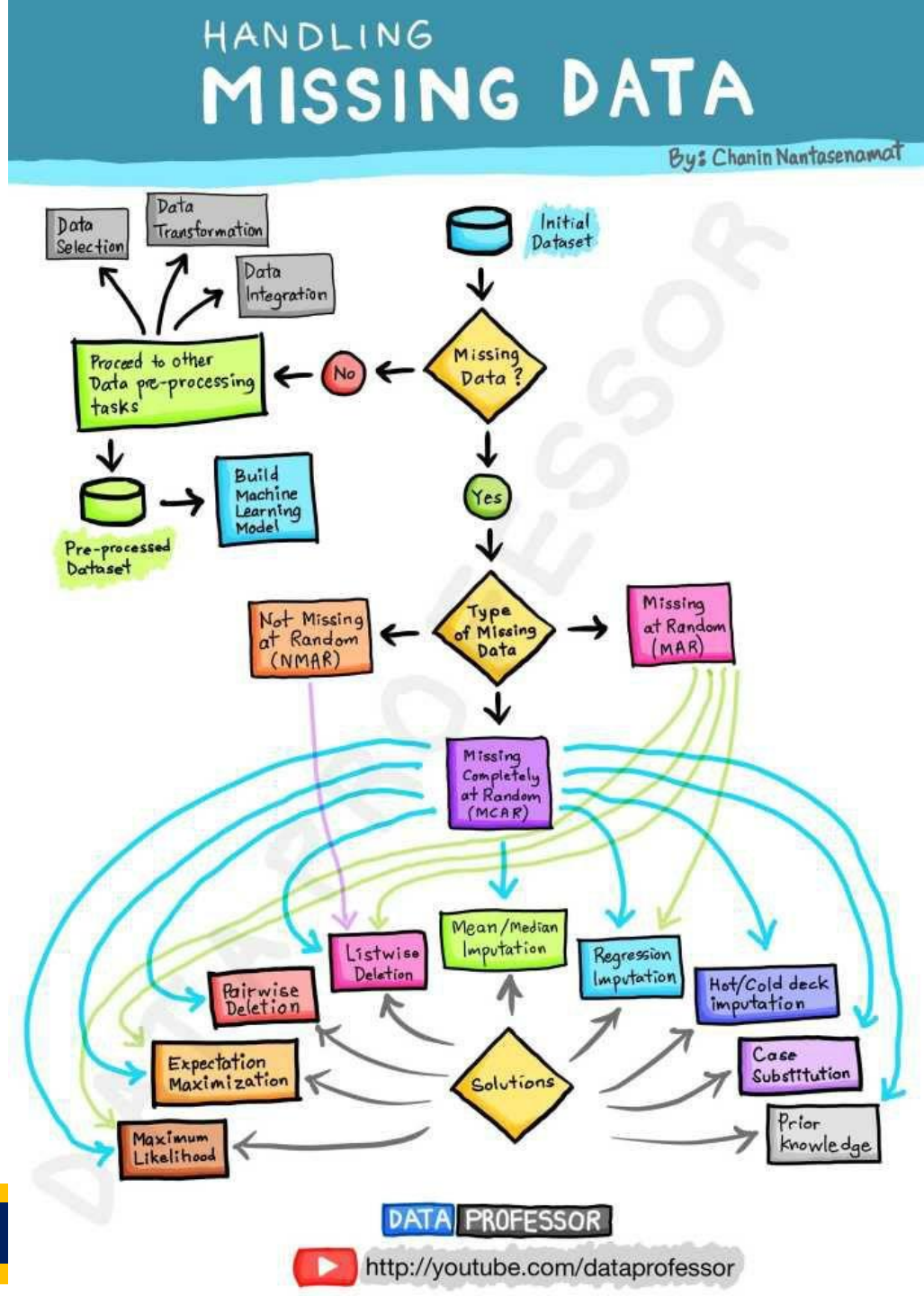
Missing data bisa terlihat dari datanya hilang, NULL, NA, 0, dll. Untuk mengatasi missing data bisa dilakukan imputasi.

Data Transformation

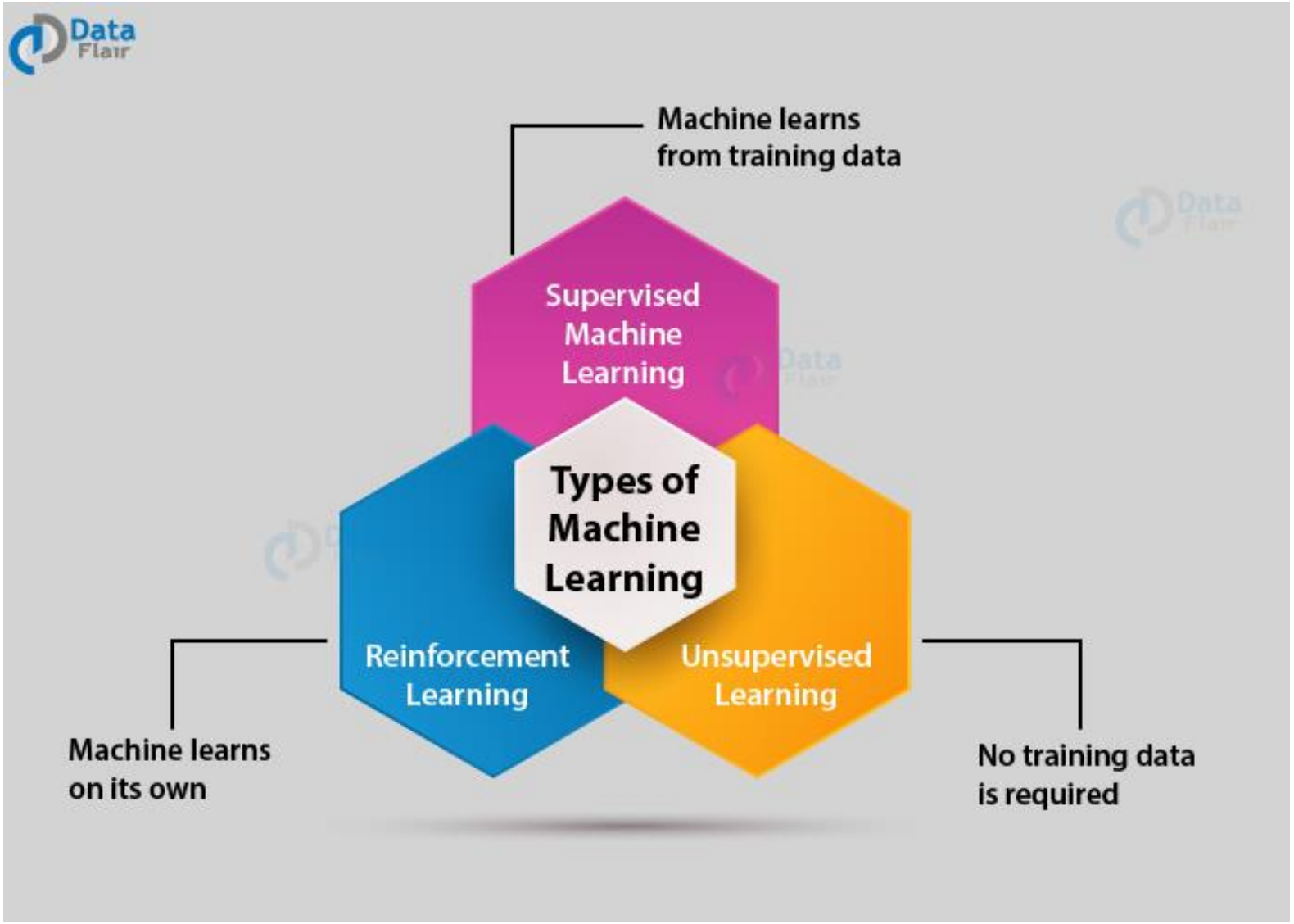
Transformasi data biasanya diperlukan apabila dalam suatu dataset memiliki beberapa variabel yang satuan/unitnya berbeda.

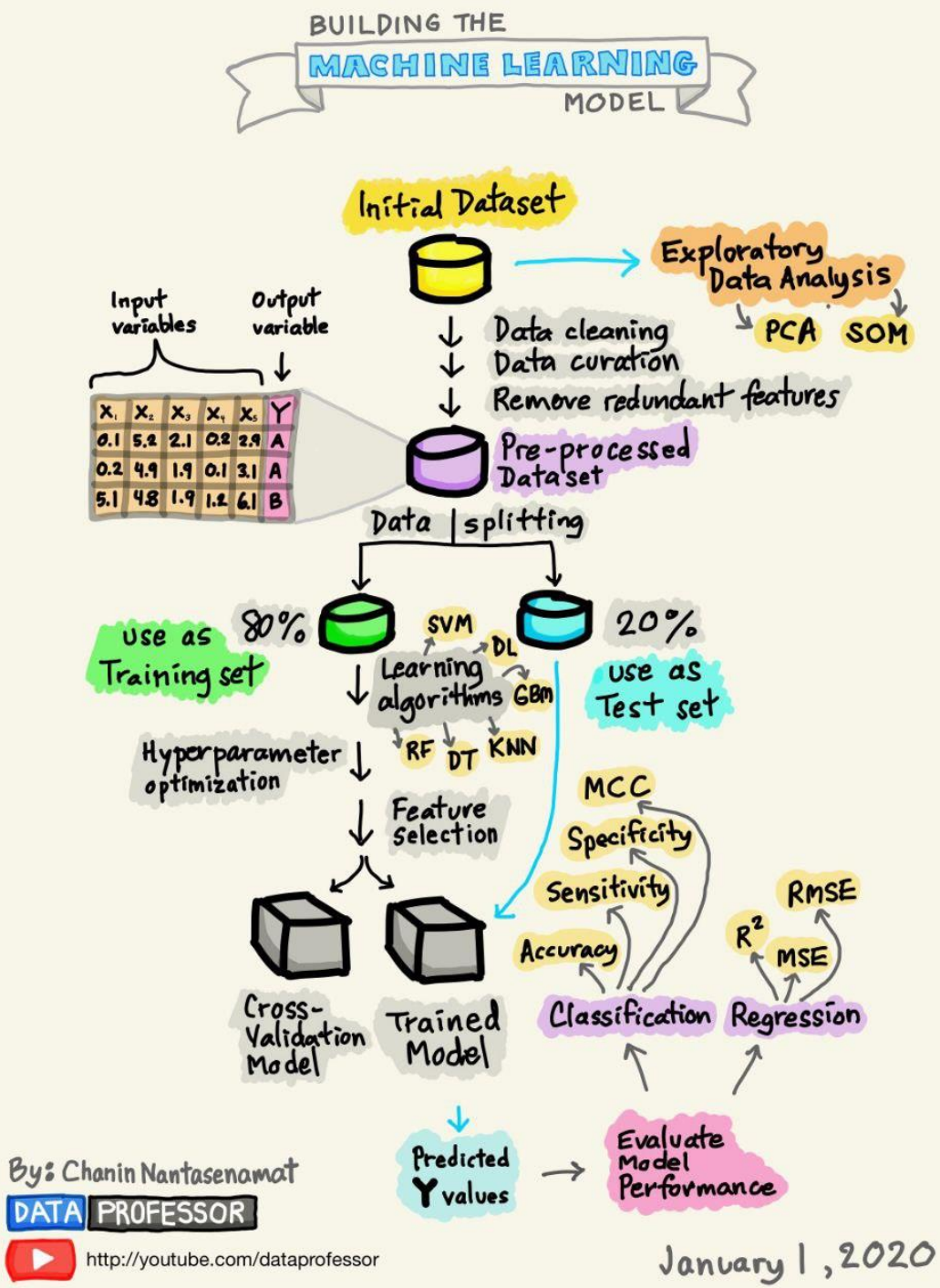
Dimensionality Reduction

Dimensionality reduction diperlukan apabila jumlah variabel/features dalam dataset sangat banyak. Cara yang bisa dilakukan diantaranya menggunakan PCA, Stepwise deletion, forward construction, backward elimination, lda (linear discriminant analysis), dll.



Data Modeling





- An "*algorithm*" in machine learning is a procedure that is run on data to create a machine learning "*model*."
- Machine learning algorithms perform "*pattern recognition*." Algorithms "*learn*" from data, or are "*fit*" on a dataset.

- A "*model*" in machine learning is the output of a machine learning algorithm run on data.
- A model represents what was learned by a machine learning algorithm.

Beberapa algoritma dalam Machine Learning:

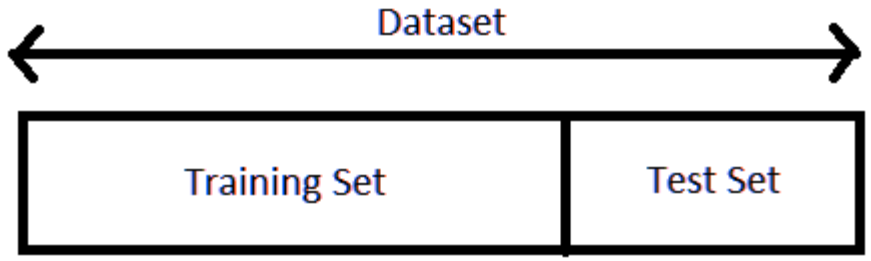
- Linear Regression
- Logistic Regression
- Decision Tree
- Artificial Neural Network
- k-Nearest Neighbors
- k-Means

Name	Type	Description	Advantages	Disadvantages
Linear Regression		-The best fit line through all data points	-Easy to understand -you can clearly see what the biggest drivers odf the model are.	-sometimes to simple to capture cpmöex relationships between variables, -Tendency für the model to overfit.
Logistic Regression		-The adoption for linear regression to problembs of classification	-Easy to understand	-sometimes to simple to capture cpmöex relationships between variables, -Tendency für the model to overfit.
Decision Tree		-A graph that uses branching method to match all possible outcomes of a decision	-Easy to understand and implement.	-Not often use of ist own for prediction because it's also often too simple and not powerful enough for complex data.
Random Forest		- Takes the average of many decision trees. Each tree is weaker than the full decision tree, but combining them we get better overall performance.	-A sort of „wisdom of the crowd", Tend to result in very high quality results. -Fast to train	-Can be slow to output predictions relative to other algorithms. -Not easy to understand predictions.
Gradient Boosting		-Uses even weaker decision trees that increasingly focused on „hard examples"	-High-performing	-A small change in the future set or training set can create radical changes in the model. -Not easy to understand predictions.
Neural Networks		-Mimics the behaviour of the brain. NNs are interconnected Neurons that pass messages to each other. Deep Learning uses severak layers of NNs to put one after the other.	-Can handle extremely complex tasks. No other alsgorithm comes close in image recognition.	-very very slow to train. Because they have so many layers. Require a lot of power. -Almost impossible to understand predictions.

Use the model

Data Splitting

- The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.
- Pembagian training dan testing pada data biasanya dilakukan dengan perbandingan 80/20 atau 70/30.



- Train set: berisi data yang akan dimasukkan ke dalam model. Secara sederhana, model kita akan belajar dari data ini.
- Test set: berisi data untuk menguji model yang telah dilatih. Dari data testing ini dapat diketahui seberapa efisien model yang terbentuk atau seberapa baik model dapat melakukan prediksi.
- K-fold cross validation juga dapat digunakan untuk membagi data training dan testing terutama jika ukuran datanya tidak terlalu banyak.

Model Evaluation

Confusion Matrix

	Predicted: NO	Predicted: YES
Actual: NO	True Negative	False Positive
Actual: YES	False Negative	True Positive

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

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{F1- score} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

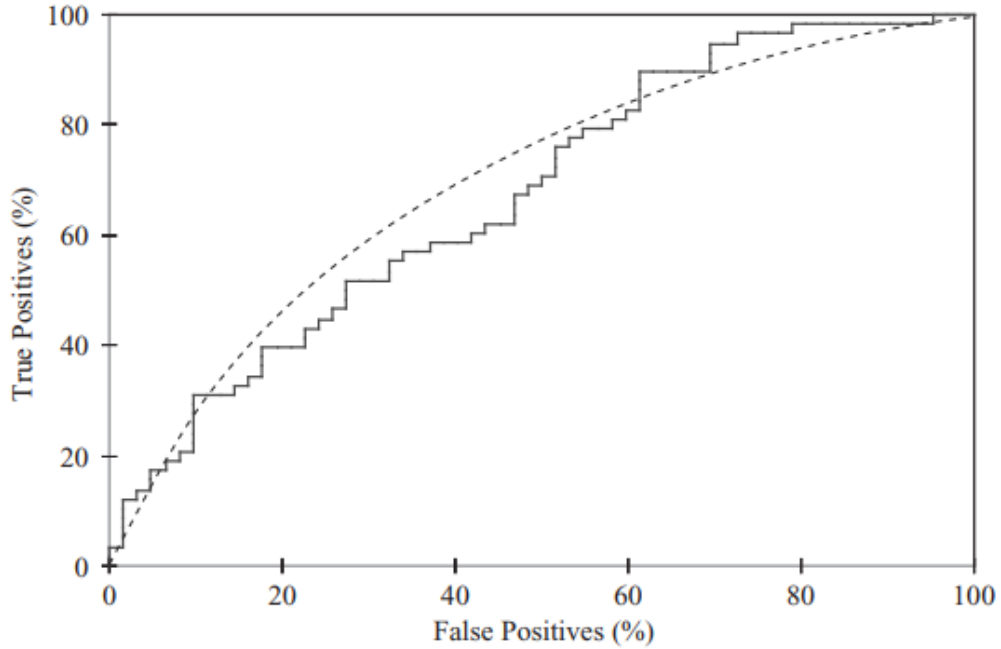
Accuracy : the proportion of the total number of predictions that were correct.
Positive Predictive Value or Precision : the proportion of positive cases that were correctly identified.
Sensitivity or Recall : the proportion of actual positive cases which are correctly identified.
Specificity : the proportion of actual negative cases which are correctly identified.
F1-Score : the harmonic mean of precision and recall values for a classification problem.

Model Evaluation

ROC Curve

ROC stands for receiver operating characteristic and the graph is plotted against TP and FP for various threshold values. As TP increases FP also increases.

To summarize ROC curves in a single quantity, people sometimes use the area under the curve (AUC) because, roughly speaking, the larger the area the better the model.

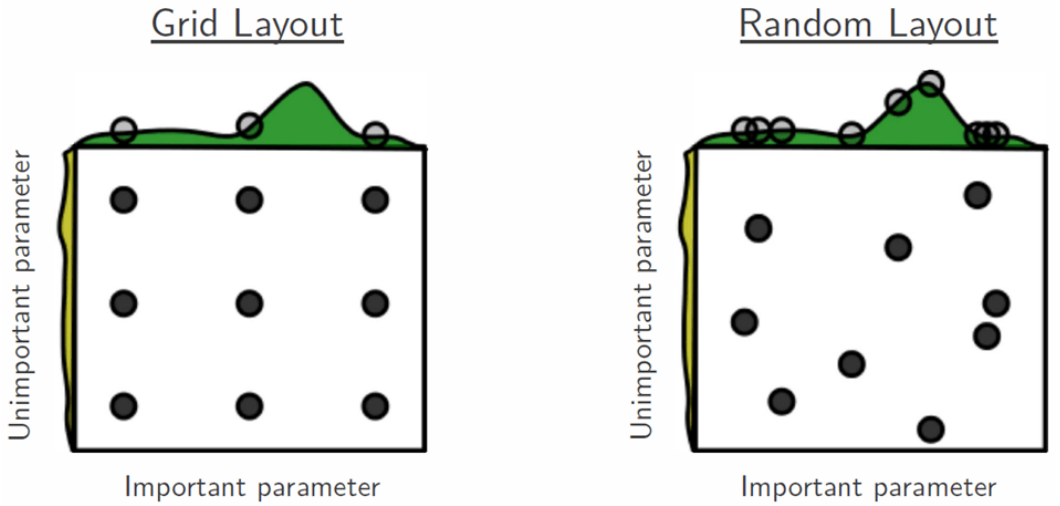


Selain itu, evaluasi model juga dapat menggunakan nilai RMSE, AIC, BIC, R^2, adj-R^2, dll.

Hyperparameter Tuning

Model hyperparameters are often referred to as model parameters which can make things confusing. A good rule of thumb to overcome this confusion is as follows: "*If you have to specify a model parameter manually, then it is probably a model hyperparameter.*" Each model has its own sets of parameters that need to be tuned to get optimal output. For every model, our goal is to **minimize the error** or say to have predictions. Some examples of model hyperparameters include:**as close as possible to actual values**

- How many trees should I include in my random forest?
- The C and sigma hyperparameters for support vector machines?
- How many neurons should I have in my neural network layer?
- How many layers should I have in my neural network?
- What should I set my learning rate to for gradient descent?
- Etc.



Source: https://www.datacamp.com/

Terima Kasih