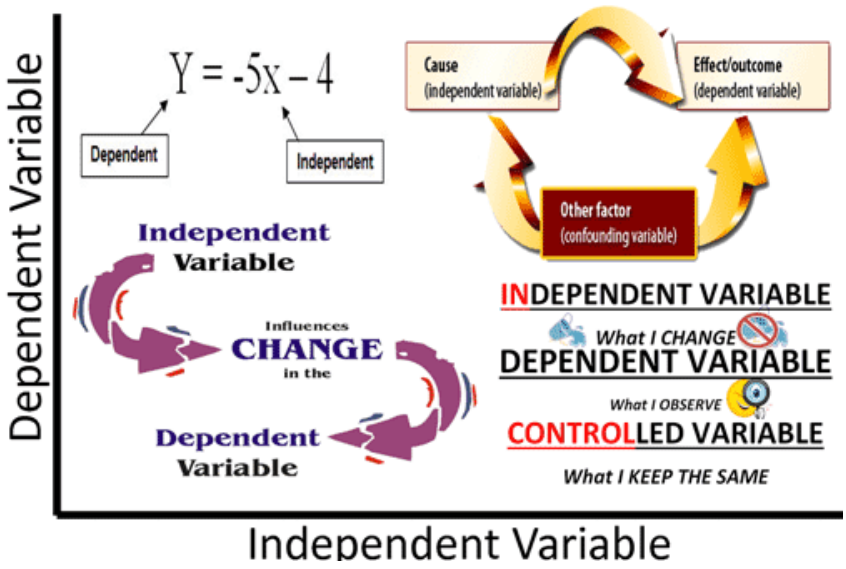
Mengenal Model-Model dan Algoritma Sains Data

1). Variabel target (*dependent*) dan prediktor (*independent*)

Sebelum membahas lebih jauh tentang model data science, akan dibahas terlebih dahulu beberapa istilah penting. Terkadang istilah-istilah untuk suatu hal tertentu berbeda-beda dan sering membingungkan. Istilah-istilah tersebut, walau pada suatu konteks tertentu maknanya bisa berbeda, namun secara umum dalam konteks yang lain bisa berarti kurang lebih sama. Sebagai contoh istilah **variabel**, *attribute, field, covariate,* peubah, indikator, dan *feature* memiliki arti yang kurang lebih sama di analisa data.



Gambar 1. *Perbedaan Variabel target dan Prediktor* .

- Variable *Dependent*/*response/regressand/predicated/explained/experimental/responding/outcome/output/* **Target**: adalah satu atau lebih variabel yang **dipengaruhi** oleh satu atau lebih variabel yang lain. Contoh: Variabel *gaji* dipengaruhi oleh variabel *lama kerja*, pangkat serta jabatan seorang pegawai.
- Variable *Independent*/*regressor/controlled/manipulated/explanatory/experimental/response/outcome/indikator/covariate*/**Prediktor** : adalah satu atau lebih

variabel yang **mempengaruhi** satu atau lebih variabel yang lain. Contoh: Variabel *kecepatan* mempengaruhi *waktu tempuh* perjalanan.

- Variabel **Kontrol**/*scientific constant*: adalah variabel/elemen yang nilainya tetap (konstan), biasanya pada suatu eksperimen untuk menguji hubungan antara variabel target dan prediktor.
  Contoh: Penggunaan Placebo (obat palsu) pada penelitian/eksperimen efek suatu obat tertentu.
- Variable **Confounding**/*confounding factor/confound/confounder*: Biasa juga disebut sebagai "variabel ketiga" atau "variabel mediator", yaitu suatu (extra*) variabel yang mempengaruhi hubungan antara variabel dependent dan independent.
  Contoh: Pada penelitian tentang dampak olahraga (prediktor) terhadap berat badan (target), maka variabel lain seperti pola makan dan usia juga akan mempengaruhi.

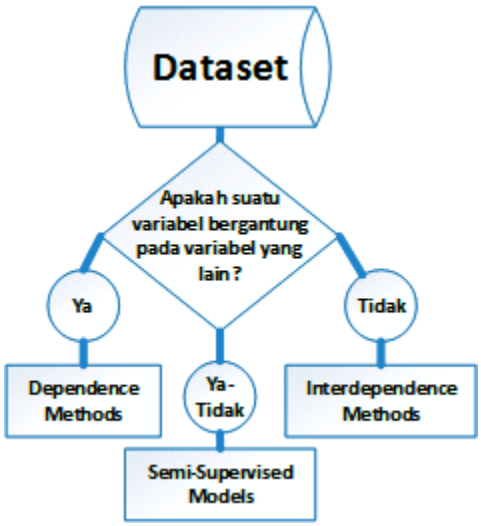
2). Metode Interdependence (~unsupervised learning) dan Dependence (Supervised Learning)

Secara garis besar terdapat dua macam kelas model statistika/data science:

1. **Interdependence Methods**: Model-model data science dimana tidak ada dugaan suatu variable dipengaruhi/memengaruhi variabel yang lain (tidak ada konsep target/prediktor).
   Contoh: Analisis cluster (pengelompokan), *principal component analysis* (PCA).
2. **Dependence Methods**: Model-model statistika dimana sebagian variabel diduga memengaruhi/dipengaruhi variabel yang lain.
   Contoh: Regresi, klasifikasi *(SVM/Decision Tree/neural network)*.

Namun demikian akhir-akhir ini terdapat cukup banyak penelitian dimana domain permasalahannya terletak di antara keduanya (Semi-Supervised Learning-akan dibahas lebih

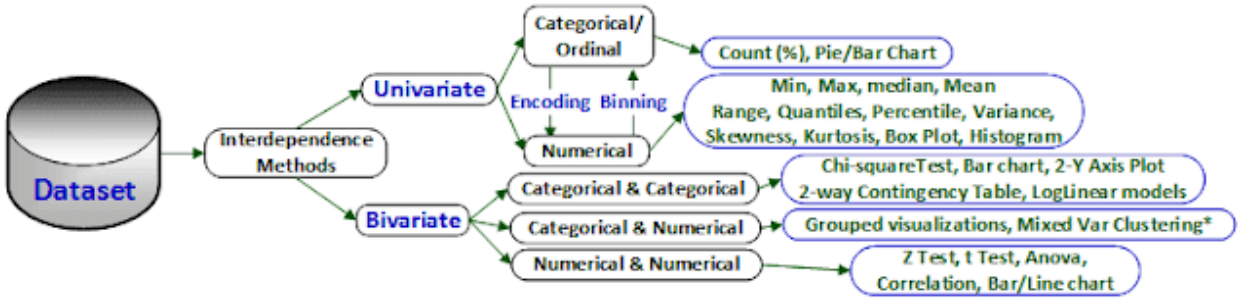
lanjut). Secara sederhana Gambar 2 menjelaskan konsep model *interdependence, dependence,* dan *semi-supervised*.



Gambar 2. (Inter)Dependence Methods Flowchart.

3). Univariate/Bivariate Data – Interdependence

Pada beberapa kasus sederhana kita hanya ingin menganalisa satu variabel (**univariate**) atau dua variabel (**bivariate**) saja yang tidak saling bergantung satu sama lain. Analisa yang bisa dilakukan pada kondisi seperti ini bergantung dari tipe variabelnya (Gambar 3).



Gambar 3. Univariate-Bivariate Interdepndence Analysis.

- **Univariate kategorik** (nominal/ordinal): Pada kasus ini tidak terlalu banyak analisa yang dapat dilakukan. Diantara hal yang dapat dilakukan adalah menghitung jumlah kemunculan (frekuensi) atau persentase. Sedangkan visualisasi yang bisa digunakan diantaranya adalah Pie-Chart atau Bar Chart.

- **Univariate numerik** (Interval/Ratio): Pada saat datanya berupa angka (numerik/metric), analisa atau perhitungan yang dapat dilakukan lebih banyak, misalnya: rata-rata/mean, median, percentile, minimum, maksimum, variansi. Histogram dan Box Plot dapat digunakan sebagai visualisasi.

  Average bermakna ukuran pusat data, dapat berupa *mean*, *median*, atau modus bergantung pada konteksnya.
- **Tambahan**: Bergantung masalahnya, ada beberapa hal lain yang bisa dilakukan saat kita menganalisa sebuah variable.

  * Binomial Test: Suatu tes yang digunakan untuk menguji apakah suatu barisan kejadian mengikuti suatu distribusi tertentu. Contoh sederhana: menguji munculnya angka pada koin atau suatu angka pada dadu untuk memeriksa apakah dadu/koin tersebut adil.

  * One Sample t-test: Digunakan untuk memeriksa apakah rata-rata sampel kita berasal dari suatu populasi dengan mean yang kita ketahui. Analogi untuk data ordinal/interval dan uji One Sample Median.

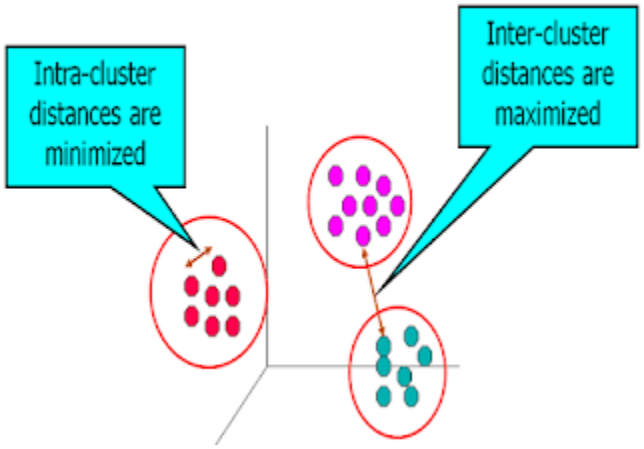
  * Uji Chi-square goodness-of-fit : Digunakan pada data kategorik untuk memeriksa apakah data yang diambil dengan suatu *teknik sampling* (simple random sampling) memiliki distribusi yang konsisten dengan populasi.
- **Encoding**/*continuization*: adalah suatu proses perubahan suatu variable kategorik menjadi bentuk biner {0,1} atau angka. Walau berubah menjadi angka, makna variabelnya tetap kategorik/non-metric. **Contoh**: {Pria,Wanita}==>{1,0} atau {TK,SD,SMP,SMU, D1,D2,D3,D4,S1,S2,S3}==>{0, 1, … , 10}.
- **Binning**/*discretization*: adalah kebalikan *encoding*, yaitu suatu proses perubahan variabel numerik menjadi kategorik. Ada 2 macam proses Binning: *supervised* dan *unsupervised*. Pada proses *binning* biasanya banyak informasi yang hilang.

Contoh: Umur seseorang dirubah menjadi kategori: balita, anak-anak, remaja, dewasa, dan manula.

- Bivariate **kategorik- kategorik**: Saat kedua variabel kategorik terdapat beberapa uji statistik dan visualisasi yang dapat digunakan seperti yang telah dijelaskan di Gambar 3 (chi-square test, 2-way contingency table, loglinear model, 2-Y charts, dll).
- Bivariate **numerik-numerik**: Saat kedua variable numerik lebih banyak pilihan analisa yang bisa dilakukan, seperti Anova (Analysis of Variance) untuk memeriksa beda rata-rata antar group atau korelasi untuk memeriksa hubungan (linear) antar kedua variabel yang ada. Kesalahan yang paling sering terjadi pada analisis korelasi adalah lupa bahwa secara umum (awalnya/tanpa transformasi) korelasi hanya memeriksa hubungan linear dan **tidak** menyatakan hubungan sebab-akibat. Sehingga nilai korelasi yang kecil tidak bermakna tidak terdapat hubungan antar keduanya (bisa jadi hubungannya bukan linear: kuadratik, exponential, atau yang lainnya). Hubungan sebab-akibat baru dapat disimpulkan setelah analisa lebih lanjut oleh ahli di bidangnya (*domain knowledge*).
- Bivariate **kategorik-numerik**: Saat kedua variable kategorik dan numerik tidak banyak "metode *interdependence*" yang bisa dilakukan. Namun demikian artikel berikut memberikan beberapa contoh visualisasi yang sangat menarik. Hal lain yang bisa dilakukan dalam cakupan *interdependence* adalah *clustering*untuk tipe variable beragam *(mixed data types clustering)*. Ada beberapa metode yang bisa digunakan, salah satunya diterangkan dalam paper berikut. Perlu diingat, bahwa di bagian ini pertimbangannya hanya pada metode interdependence, analisa yang bisa dilakukan lebih banyak lagi ketika variabelnya diasumsikan saling bergantung satu sama lain (*dependence*).

4). Clustering Analysis (Unsupervised Learning)

Clustering/pengelompokan data memiliki tujuan umum untuk mengelompokkan/*grouping* data sedemikian sehingga objek-objek pada suatu cluster *similar* (serupa) satu sama lain dan objek antar cluster berbeda (Gambar 4). Aplikasi clustering sering digunakan untuk menemukan suatu informasi/*pattern* yang tersembunyi (*latent/hidden*) di data. Tidak hanya itu clustering juga biasa digunakan untuk mendeteksi *outlier* atau anomali yang ada di data. Tentu saja masih banyak lagi aplikasi analisa clustering, mulai dari *spam detection*, *image processing*, riset pemasaran (*market research*), dan masih banyak lagi. Analisa Cluster termasuk analisa/metode yang paling tua dan paling banyak aplikasi serta penelitiannya.



Gambar 4. Clustering Analysis

- **Tipe-tipe clustering**: Sebenarnya masih banyak lagi clustering dengan pendekatan lain. Di antaranya adalah clustering dengan menggunakan ranking, faktorisasi matriks atau tensor, dan Co-Clustering.
- **Co-Clustering**/*BiClustering/two-mode clustering/block clustering:* Adalah teknik pengelompokkan data, dimana *instances*/observasi/baris dikelompokkan dan dalam waktu yang sama juga dilakukan pengelompokkan berdasarkan variabel. Pada data terstruktur ini berarti baris dan kolom

dikelompokkan secara bersamaan. Aplikasi *co-clustering* banyak dimanfaatkan di bidang *Bioinformatics*.

- **Evaluasi**: Makna "*cluster*" sebenarnya tidak terdefinisi dengan baik (*well-defined*). Inilah yang sebenarnya membuat clustering lebih '*tricky*' ketimbang model lain seperti klasifikasi atau regresi.

  [1]. **Evaluasi Internal**: Evaluasi akan hasil clustering berdasarkan data dan hasil cluster. Sayangnya evaluasi internal ini *bias* terhadap metric dan algoritma apa yang kita gunakan. Sebagai contoh, ukuran Silhouette coefficients cenderung lebih menguntungkan algoritma yang menggunakan centroid seperti k-means. Contoh lain ukuran internal popular lain: Davies–Bouldin index & Dunn index.

  [2]. **Evaluasi Eksternal**: Kelemahan evaluasi internal cluster selain bias terhadap pemilihan metric dan algoritma adalah "*evaluasi internal terbaik belum tentu merupakan hasil yang paling bermanfaat di aplikasi nyatanya*". Terkadang (kalau tidak seringnya) cluster yang terbentuk tidak seperti yang diharapkan *user*. Evaluasi eksternal mengatasi hal ini dengan membandingkan hasil cluster dengan suatu "*Gold standard/Ground Truth*" yaitu suatu hasil pengelompokkan yang dilakukan (biasanya) secara manual oleh para ahli. Ukuran evaluasi external yang paling popular adalah NMI *(Normalized Mutual Information)* dan *F-Score*.

  Perhitungan *F-Score* dan NMI di *clustering* berbeda dengan *F-Score* dan NMI pada model klasifikasi. Pada *clustering* digunakan *pairwise F-Score* dan NMI.

  [3]. **Evaluasi Aplikasi**: Kelemahan evaluasi internal dan eksternal *clustering*adalah komputasi yang cukup tinggi. Saat datanya besar (*e.g. Big Data*) nyaris tidak mungkin (minimal sangat *costly*) mengevaluasi hasil cluster dengan cara di atas. Akhir-akhir ini evaluasi dengan mengaplikasikan hasil clusternya lebih diminati (e.g. spam detection).
- Jarak/*metric*/**Distance**~**Similarity** : Cukup banyak metode *cluster*menggunakan konsep jarak (*distance*) dan *similarity*.

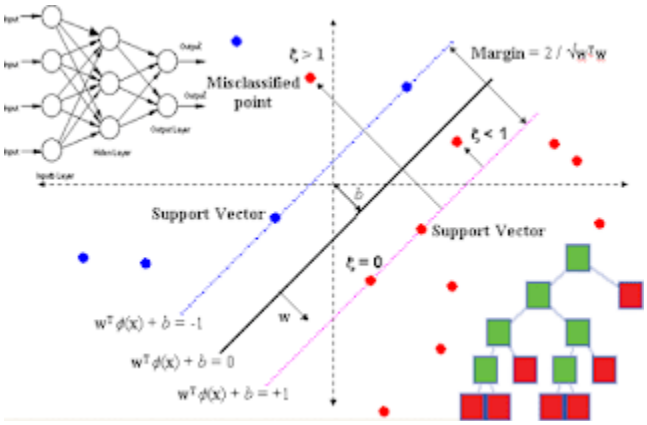
Jarak/*distance*sebenarnya "*metric*" di topology (Matematika). Sebuah fungsi dikatakan sebagai fungsi jarak jika ia memenuhi tiga syarat: *non-negative*, *symmetric*, dan ketidaksamaan segitiga.

Sehingga, fungsi "*cosine*" yang sering digunakan pada *clustering* (terutama saat datanya teks) sebenarnya **bukanlah** jarak, tapi lebih cocok disebut sebagai ukuran kedekatan (*similarity*). Karena *cosine* tidak memenuhi sifat ketidaksamaan segitiga.

- *Curse of dimensionality* (cod): Di *clustering* (terutama data teks) istilah ini cukup terkenal. cod merupakan anomali yang terjadi akibat variabel (dimensi) yang terlalu banyak (tinggi) di data. Saat dimensi semakin besar, konsep "dekat" dan "jauh" semakin tidak bermakna. Penjelasan lebih lanjut akan diberikan di artikel selanjutnya.
- **Clustering untuk Big Data**: Tantangan analisis cluster untuk data yang besar tidak hanya komputasi yang cepat (mendekati real-time) untuk mendapatkan analisa yang bermanfaat untuk pengambilan keputusan (*value*), atau *curse of dimensionality*, tapi juga karena sifat "*velocity*" di Big Data, maka diperlukan metode clustering yang mampu mendeteksi munculnya cluster baru di data yang belum ada sebelumnya (mirip konsep drift).

5). Classification Models (Supervised Learning)

Klasifikasi adalah permasalahan meng-kategorisasikan sekelompok observasi baru ke sekumpulan kategori (kelas) yang ada sebelumnya. Klasifikasi digunakan jika variabel target bertipe kategorik dan prediktornya satu atau lebih variabel numerik dan-atau kategorik. Terdapat cukup banyak model klasifikasi yang dapat digunakan, mulai dari yang klasik seperti Linear Discriminant Analysis (LDA) dan regresi logistik, lalu ke *moderate* seperti SVM (support vector machines), decision tree dan neural network (jaringan syaraf tiruan), sampai yang lebih terkini seperti random forest, XGboost dan deep learning.



Gambar 5. Classification Methods.

- **Aplikasi**: Aplikasi klasifikasi sangat banyak, mulai dari deteksi wajah di *mobile phone*/kamera yang kita gunakan, deteksi kanker, pengenalan suara (e.g. Cortana/SIRI), dan masih banyak lagi.
- **Linear-Non Linear Classifier**: Sebagian model klasifikasi awalnya di peruntukkan untuk memisahkan data secara linear (e.g. SVM atau LDA). Namun demikian model non-linear dapat digunakan dengan mentransformasi data atau menggunakan fungsi kernel.
- **Kernel**: Fungsi kernel (di machine learning) sering digunakan untuk mentransformasi data ke dimensi yang lebih tinggi dengan harapan di dimensi yang baru tersebut (biasa disebut feature space) data dapat dipisahkan secara linear. Teknik ini (biasa disebut kernel trick), memiliki resiko *curse of dimensionality* seperti yang telah dijelaskan sebelumnya. Makna "kernel" di Statistik, Machine Learning, Computer Science (OS), Matematika, dan biologi jauh berbeda. Ketika berbicara mengenai "kernel" yakinkan selalu dalam konteks yang jelas.
- **Data Training, Test, & Validasi**: Ketika melakukan analisa dengan model klasifikasi biasanya data dibagi menjadi 2: Training data, yaitu data yang digunakan untuk membentuk model terbaik, lalu data test yang digunakan untuk meng-evaluasi model. Namun terkadang ada sebagian *practitioner* yang

membagi datanya menjadi 3 (+*validation set*). Mengapa? karena biasanya proses pemodelan di training data dan evaluasi di test data dilakukan berulang-ulang hingga didapat model yang optimal. Sehingga satu-satunya mekanisme untuk meng-evaluasi tahap akhir yang objektif untuk melihat seberapa baik modelnya adalah melalui data yang ada di validation set.

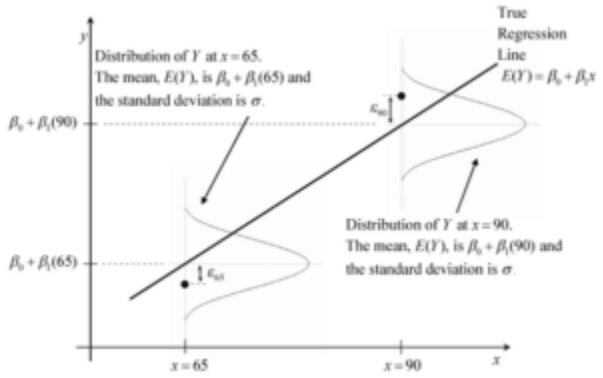
Error empiris (error pada training set) sebenarnya tidak terlalu penting. Mengapa? Karena pemodelan/algoritma hampir seluruh model klasifikasi adalah dengan menyelesaikan suatu optimasi yang meminimalkan error ini. "**Namun**" model klasifikasi yang dihasilkan akan digunakan di data yang tidak ada di training data. Dengan kata lain, dengan bahasa statistika, kita memodelkan dengan sampel, namun modelnya sendiri akan digunakan di populasi (diluar sampel).

- **OverFitting**: Overfitting terjadi jika modelnya terlalu fit (cocok/baik) ke training data, namun tidak ke test data. Dengan kata lain memiliki kemampuan generalisasi ke populasi yang buruk. Overfitting biasanya terjadi karena modelnya terlalu kompleks (e.g. memiliki terlalu banyak parameter).
- **Evaluasi**: Model klasifikasi memiliki beberapa metric evaluasi seperti precision, recall, FScore, NMI, dan ROC. Evaluasi di model klasifikasi biasanya lebih *straight-forward*.

6). Regression Models

Model Regresi digunakan saat kita ingin menganalisa hubungan antara variabel target bertipe numerik dengan satu atau beberapa variable prediktor bertipe kategorik dan-atau numerik.

Regresi termasuk model yang paling dasar ketika seseorang pertama kali belajar pengolahan data (statistika).

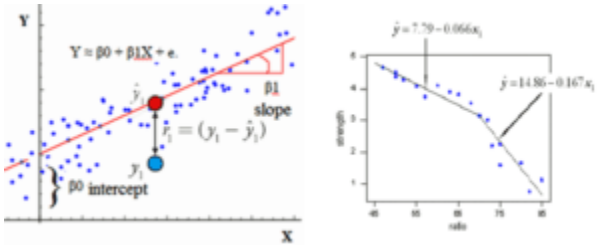


Awalnya regresi biasanya digunakan untuk melihat hubungan linear antar variabel target dan prediktor (Gambar 6). Namun beberapa teknik seperti transformasi data (kernel) atau piecewise regression dapat digunakan untuk menganalisa pola non-linear.

- **Aplikasi**: Regresi memiliki penerapan di banyak bidang, baik ekonomi, bisnis, sosial, psikologi, bahkan ke sains seperti biologi, fisika, maupun kimia.
- **Model-model regresi**: Seperti yang diberikan di gambar 1 terdapat cukup banyak model regresi (tidak hanya Regresi Linear Berganda/sederhana yang biasanya dibahas di buku pendahuluan Statistika): Ordinary Least Square (OLS), Regression tree, bahkan Neural Network/SVM untuk regresi, dan masih banyak lagi.
- **Interpolasi dan Ekstrapolasi**: Regresi bagi Matematikawan sangat tidak asing, di matematika konsep interpolasi/curve-fitting sangat mirip dengan analisa regresi di statistika. Perbedaannya adalah di matematika interpolasi hanya melihat error deterministik datanya, namun tidak menelaah lebih jauh distribusi data (dan generalisasi). Jika interpolasi hanya melihat prediksi dalam domain data prediktor, ekstrapolasi berusaha memprediksi diluar batasan domain prediktor (sampel).

  Kesalahan paling umum pengguna analisa regresi adalah melakukan ekstrapolasi berlebihan. Analisa regresi tidak di-desain untuk melakukan prediksi diluar domain

variabel prediktor. Jika ekstrapolasi hendak dilakukan maka sederetan asumsi yang sangat ketat harus dipenuhi.

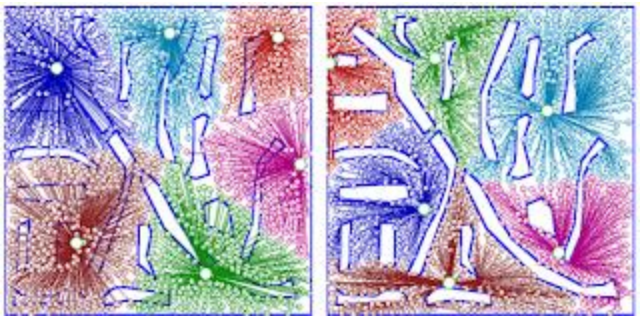


***Gambar 7**. Regression Analysis.*

7). Semi-Supervised Learning

Seperti yang telah ditunjukkan di Gambar 3, ada kalanya data yang kita miliki memiliki variabel target dan prediktor, namun terdapat data yang hilang (*missing values*) di beberapa observasi di variabel targetnya. Atau pada kasus lain, misal kita ingin melakukan clustering, namun dengan suatu kendala/batasan tertentu (*constraint*), atau kita menginginkan clustering yang sesuai dengan suatu informasi awal (prior) yang kita miliki. Pada kasus-kasus seperti ini semi-supervised learning digunakan.

- **Aplikasi**: Aplikasi semi-supervised learning banyak sekali, saya akan memberikan satu ilustrasi untuk memperjelas. Gambar 8 (kiri) adalah contoh yang cukup terkenal dari penelitian Tung et al (2000), ketika melakukan clustering analysis mesin-mesin ATM bank di suatu kota. Namun di kota tersebut terdapat beberapa hambatan (e.g. sungai), sehingga solusi clustering (unsupervised/biasa) memiliki solusi yang menghitung jarak melewati hambatan tadi. Dengan constrained clustering (gambar 8 – kanan) kita bisa mendapatkan hasil clustering yang lebih feasible (sesuai/layak).

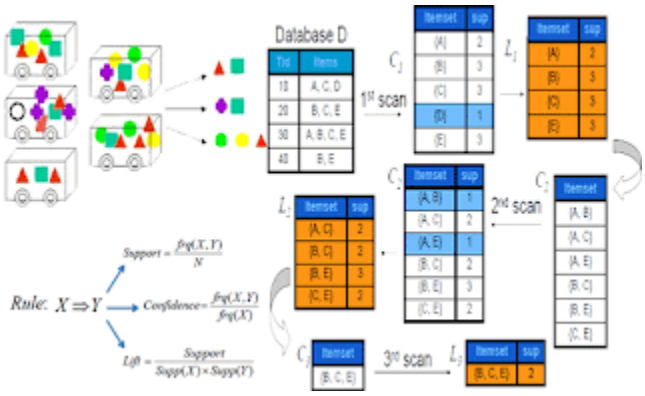


Gambar 8. Salah satu aplikasi semi-supervised clustering yang cukup terkenal Tung et al, 2000

Semi-Supervised Clustering: Semi-supervised clustering bisa menjadi solusi permasalahan umum clustering yang sudah dibahas sebelumnya. Dengan teknik yang ada di semi-supervised clustering kita bisa mendapatkan solusi clustering yang tidak hanya feasible, namun juga sesuai dengan yang kita inginkan/harapkan. Secara umum terdapat beberapa pendekatan semi-supervised clustering yang akan saya bahas dengan rinci di post yang lain (Gambar 9): Semi-supervised clustering dapat dilakukan dengan menggunakan label (topic), fungsi jarak tertentu, constraint (kendala), hybrid, dan pemilihan parameter cluster.

8). Association Rule/ Market Basket Analysis

Model Association Rule (biasa juga disebut sebagai Market Basket Analysis) bisa digunakan untuk mengoptimalkan tata letak barang-barang yang ada di suatu swalayan atau menentukan program promo yang tepat. Salah satu aplikasi association rule (AR) adalah dengan mengolah data pembelian konsumen lalu menghitung beberapa statistik darinya seperti support, confidence, dan lift untuk menentukan pola belanja konsumen. Ketika algoritma AR dijalankan, ia akan menghitung kombinasi item yang dibentuk dari item-item yang ada, hal ini menjadi kendala ketika jenis itemnya cukup banyak. Selain masalah komputasi, menemukan aturan (rule) yang secara statistik signifikan menjadi cukup menantang pada keadaan seperti ini.



Gambar 9. Market Basket Analysis / Association Rule.

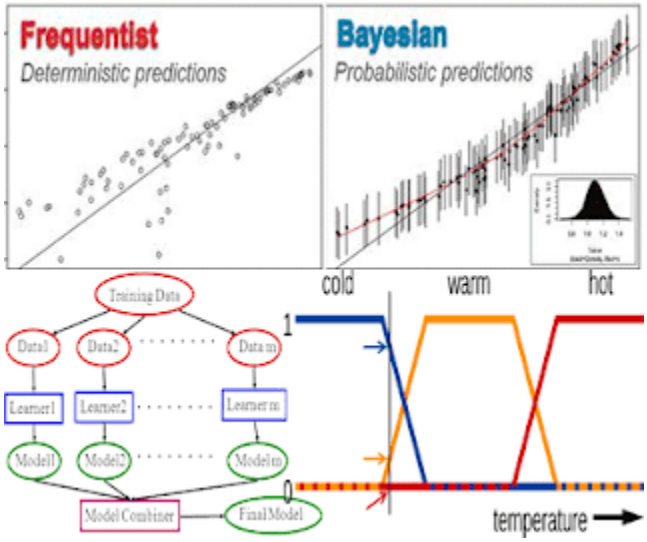
9). Bayesian, Ensemble, dan Fuzzy

Kalau di perhatikan di Gambar 1, ketiga metode ini dikaitkan dengan klasifikasi, clustering, dan regresi. Mengapa? Karena baik pendekatan Bayesian, Ensemble, maupun Fuzzy logic dapat digunakan dalam ketiga permasalahan tersebut.

- **Bayes**: Thomas Bayes (1701–1761) bisa jadi merupakan salah satu statistikawan paling berpengaruh sepanjang sejarah. Teorema Bayes (Bayes Rule) adalah salah satu teorema fenomenal yang menjelaskan probabilitas bersyarat suatu kejadian. Bayesian probability ini menjadi dasar penting Bayesian inference yang merupakan dasar penting Bayesian Statistics (e.g. Gambar 11-bagian atas). Model Bayes memiliki keunggulan karena hasilnya merupakan suatu distribusi probabilitas, sehingga pengambil keputusan memiliki keleluasaan dan keyakinan yang lebih baik, ketimbang model frequentist (model-model yang kita bahas sebelumnya).
- **Ensemble**: Beberapa tahun belakangan ini model-model ensemble menjadi salah satu topik hangat penelitian (termasuk di dalamnya random forest). Model ensemble pada dasarnya adalah perpaduan beberapa model (Gambar 10-kiri-bawah). Salah satu ensemble yang paling mudah adalah concensus model. Menggunakan teknik ensemble kita dapat meningkatkan akurasi

(biasanya tidak banyak), namun komplekasitas komputasinya meningkat sangat besar, sehingga tidak cocok untuk data yang besar (e.g. Big Data).

- **Fuzzy**: Fuzzy logic men-generalisir konsep binary logika (True/False-{0,1}) ke interval kontinu [0,1]. Dengan logika fuzzy suatu kebenaran bisa bernilai 3/4 True dan 1/4 False. Konsep fuzzy bisa digunakan untuk memodelkan suatu kategori yang tidak tegas, misal konsep dingin, hangat, dan panas (Gambar 10 kanan-bawah). Dari konsep fuzzy ini, model Fuzzy Clustering, Fuzzy Classification, dan Fuzzy Regression dapat dikembangkan.

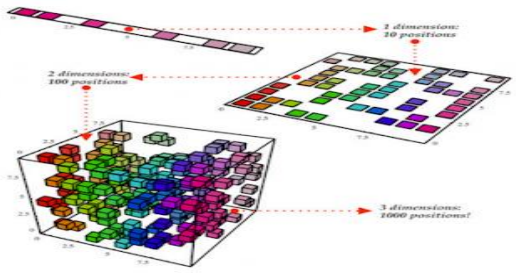


Gambar 10. Bayesian Model (Atas), Konsep Ensemble (kiri bawah), & Fuzzy Logic (kanan bawah).

10). Dimensionality Reduction/Feature Selection

Ada kalanya kita disuguhkan seonggok data dengan begitu banyak variabel dan sebuah hipotesis (dugaan) atau sebuah variabel target. Tentu saja tahap pertama yang perlu kita lakukan adalah memilah variabel mana yang merupakan prediktor yang baik bagi target kita, sisanya biasanya hanyalah "noise" bagi modelnya. Atau pada keadaan lain kita memiliki terlalu banyak variabel yang membuat komputasi menjadi terlalu tinggi. Skenario-skenario ini adalah saat-saat dimana pengurangan dimensi (variabel) dibutuhkan (Gambar 11).

- **Teknik reduksi dimensi**: Principal Component Analysis/SVD, Latent Dirichlet Allocation, Random Indexing, dll.
- **Aplikasi**: Menghilangkan noise, mempercepat komputasi atau mengurangi space penyimpanan, visualisasi, menghilangkan korelasi antar variabel prediktor (regresi) atau keperluan interpretasi model.
- **Feature Selection**/variable selection/attribute selection/variable subset selection: adalah suatu proses pemilihan sebagian variabel dari keseluruhan variabel yang ada. Ada tiga metode utama dalam pemilihan variabel ini yang akan dibahas di lain kesempatan: **Filter, Wrapper, dan Embedded**.
- **Feature Extraction**: Sebelumnya kita memiliki data terstruktur dalam bentuk tabel (numerik/kategorik) yang sudah "siap saji". Namun ketika dihadapkan dengan data tidak terstruktur seperti teks, gambar, suara atau video maka kita perlu melakukan suatu proses transformasi data tersebut menjadi suatu bentuk yang lebih terstruktur. Proses ini mengacu pada masalah "feature extraction", yaitu penentuan variable penentu dari suatu objek. Sebagai contoh misal kita memiliki data tweet dari beberapa user Twitter, setiap kata pada tweet tersebut bisa digunakan sebagai feature (variabel) dengan domain frekuensi kemunculan kata. Namun tentu saja frekuensi kata bukan satu-satunya pilihan, masih banyak cara feature bisa dipilih mulai dari binary, tf-idf, topic, top-sig, dsb.



Gambar 11. Dimensionality Reduction