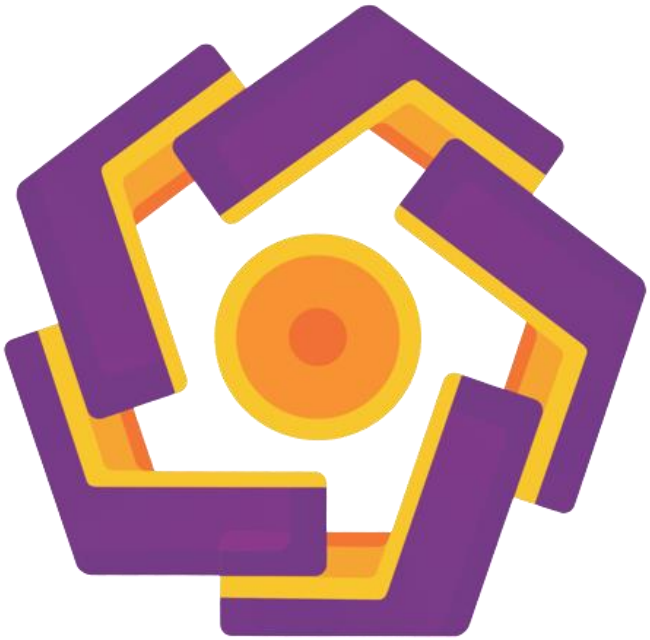
TUGAS MAHASISWA – I

RANGKUMAN DARI BERBAGAI SUMBER
KONSEP DAN CONTOH SESUAI DENGAN KATA KUNCI



ANOSA PUTRI RUISE
21.55.2175

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PEMBELAJARAN JARAK JAUH
UNIVERSITAS AMIKOM YOGYAKARTA
2022

TUGAS 1

MERANGKUM KONSEP DAN CONTOH DARI BERBAGAI SUMBER

1. Business Analityc, Data Analytic, dan Data Science

a. Business Analytic

Dari sisi pembagian tugas, biasanya organisasi akan memandang implementasi TIK ini menjadi dua sisi: fungsi spesialis TIK (atau sering dikenal sebagai "orang IT") sebagai pihak yang bertanggung jawab dari sejak pemilihan strategi/solusi TIK yang sesuai dengan kebutuhan organisai hingga implementasinya, dan fungsi pengguna (atau enduser) sebagai pihak yang akan menggunakan solusi TIK yang telah tersedia untuk digunakan sebagai perangkat bantu penyelesaian pekerjaan dalam setiap fungsi organisasi (McLeod dan Schell 2001).

Pembagian tersebut dari waktu ke waktu semakin cair, sehingga sampai dengan tingkat tertentu end-user karena berbagai alasan yang melatarinya diberikan juga peluang untuk mengembangkan sendiri aplikasi TIK yang sesuai dengan kepentingan fungsi bisnisnya. Hal ini sering disebut sebagai end user computing (EUC). Alasan penerapan EUC ini, menurut McLeod dan Schell (2001) antara lain adalah keterbatasan spesialis informasi untuk memenui semua kebutuhan end user sehingga akhirnya end user harus berusaha memenuhi sendiri kebutuhan mereka. Sampai dengan hari ini EUC ini semakin berkembang.

Analytics merupakan konsep, pemikiran dan praktik/implementasi yang bermanfaat untuk melakukan analisis terhadap suatu permasalahan IT dalam sebuah organisasi, biasanya analisis tersebut akan mengarah dan menekankan pada tujuan atau proses/algoritma dari target perusahaan.

b. Data Analytic

Data analytic adalah cara dan proses dalam mengolah berbagai data dan informasi, data analytic tidak terlepas dari big data yang dibutuhkan untuk ditransformasikan menjadi sebuah informasi yang berguna bagi suatu organisasi maupun perusahaan. Hal utama yang harus dilakukan dalam data analytics antara lain mengumpulkan, mengelola, menyimpan, menganalisis, membuat laporan, menayangkan serta mengamankan koleksi informasi yang sudah dihasilkan.

c. Big Data

Big Data awalnya disebut dalam sebuah artikel ilmiah dengan judul "Aplication-Controlled Demand Paging For Out-Of-Core Visualization, yang ditulis oleh Michael Cox dan David Ellsworth pada tahun 1997.

Diantara definisi Big data pernah dikemukakan oleh Gartner yang menyimpulkan bahwa bigdata adalah aset informasi yang sangat besar dan bergerak sangat cepat, serta amat bervariasi yang membutuhkan cara baru pemrosesan untuk memperbaiki pembuatan keputusan, menemukan pemahaman dan mengoptimalkan proses Karakteristik pada bigdata terletak pada 3V yaitu volume, velositas dan varietas dan dalam perjalanannya tumbuh satu V lagi yaitu value yang terkait dengan nilai dan kegunaan data yang tersedia. Dari definisi tesebut maka data terkait erat dengan informasi yang saat ini tersedia begitu banyak. Besarnya

informasi yang tersedi hingga dalam jumlah yang tidak kita bayangkan akhir-akhir ini merupakan sesuatu keuntungan bagi kita yang hidup di era informasi, namun demikian juga memiliki sisi perlu kita cermati mengingat jumlah yang sangat besar sehingga diperlukan proses seleksi terhadap data yang memang berguna.

Jika demikian maka big data ini merupakan suatu situasi nyata yang kita hadapi dan membutuhkan perhatian dan kepedulian kita untuk mengelolanya. Bukan pada ukuran jumlahnya yang besar, tetapi lebih pada kegunaan bagi kehidupan kita baik di Lembaga maupun untuk kebutuhan pribadi (Albertus 2015) (Narendra, 2015).

2. Data, Informasi dan Pengetahuan

a. Data

Data adalah hasil observasi langsung terhadap suatu kejadian/fakta yang mewakili suatu objek atau konsep dalam suatu kejadian yang nyata. Menurut Ralston dan Reilly (Chamidi, 2004: 314), data didefinisikan sebagai fakta atau apa yang dikatakan sebagai hasil dari suatu observasi terhadap fenomena alam. Sebagai hasil observasi langsung terhadap kejadian atau fakta dari fenomena di alam nyata, data bisa berupa tulisan atau gambar yang dilengkapi dengan nilai tertentu. Contohnya, daftar hadir siswa semester 1 Ilmu Perpustakaan dan kearsipan adalah data. Daftar tersebut masih merupakan bentuk mentah karena belum memberikan informasi apa-apa.

b. Informasi

Banyak orang mengatakan bahwa informasi adalah segala yang kita komunikasikan, seperti yang disampaikan oleh seseorang lewat bahasa lisan, surat kabar, video, dan lain-lain. Ungkapan ini— karena seringnya dipakai—Fox (1983) yang dikutip Pendit (1992:64) mengategorikannya sebagai the ordinary notion of information. Dalam ungkapan ini, terkandung pengertian bahwa tidak ada informasi kalau tidak ada yang membawanya. Di antara yang membawa informasi ini, yang paling sering dibicarakan adalah bahasa manusia melalui komunikasi antarmanusia. Meskipun tidak selalu manusia yang membawa informasi, komunikasi bisa juga berarti asap, DNA, aliran listrik, atau gambar. Dengan demikian, informasi di sini bisa dianggap sebagai pesan atau makna yang terkandung dalam sebuah pesan. Padahal, dalam kenyataan sehari-hari, sering kita harus membedakan informasi yang dikandung suatu kalimat atau yang tertulis dalam kalimat tersebut. Misalnya, si A mengatakan, “Pintar kamu,” kepada si B. Belum tentu yang dimaksud si A bahwa si B benar-benar pintar, tetapi ada makna lain. Jadi, ada makna yang terkandung dalam informasi tersebut.

Kesimpulannya adalah ada 3 makna berbeda dari kata informasi, pertama adaah sebagai suatu proses yang merujuk pada suatu kejadian menjadi terinformasi, yang kedua adalah bermakna sebagai pengetahuan yang mengacu pada kegiatan di dunia yang tak terhingga , yang ketiga adalah sebagai suatu penyajian yang nyata dari sebuah pengetahuan yang dapat dilihat dari symbol-simbol dan dapat ditanggap oleh pancaindera manusia dan dapat dipertukarkan.

Terdapat berbagai macam jenis informasi yang dikemukakan oleh Soetaminah (1991), diantaranya adalah:

i. Informasi untuk kegiatan politik, yang dapat digunakan oleh politikus untuk kepentingan politiknya
ii. Informasi untuk kegiatan pemerintahan yang digunakan untuk menyusun rencana, membuat keputusan dan kebijakan pemerintahan
iii. Informasi kegiatan social yang digunakan untuk membuat program kerja
iv. Informasi untuk dunia militer yang dapat digunakan untuk melakukan perubahan system persenjataan
v. Informasi untuk penelitian untuk mencari informasi tentang penelitian yang pernah dilakukan oleh peneliti lain

Beberapa karakteristik informasi yang pernah dikemukakan oleh Wulandari (2007) diantaranya:

i. Luas Informasi
ii. Kepadatan Informasi
iii. Frekuensi Informasi
iv. Waktu Informasi
v. Sumber Infrmasi

c. Pengetahuan

Pengetahuan adalah sesuatu yang digunakan manusia untuk memahami dunia, yang dapat diubah-ubah berdasarkan informasi yang diterima. Pengetahuan si A bisa berbeda dengan pengetahuan si B, berdasarkan informasi yang sama. Dengan demikian, informasi dan data merupakan sarana baku untuk menunjang dan meningkatkan kegiatan bidang ilmu pengetahuan, kebudayaan, dan teknologi.

Perbedaan konsep data, informasi, dan pengetahuan dijelaskan oleh Teskey (Pendit, 1992: 80—81) seperti berikut. Data adalah hasil dari observasi langsung terhadap suatu kejadian. Ia merupakan entitas (entity) yang dilengkapi dengan nilai tertentu. Entitas ini merupakan perlambangan yang mewakili objek atau konsep dalam dunia nyata. Data ini bisa disimpan dalam bentuk lebih konkret, misalnya dalam bentuk tertulis, grafis, elektronik, dan sebagainya. Sementara itu, informasi adalah kumpulan data yang terstruktur untuk memperlihatkan hubungan-hubungan entitas di atas. Pengetahuan adalah model yang digunakan manusia untuk memahami dunia dan yang dapat diubah-ubah oleh informasi yang diterima pikiran manusia.

Hubungan antara informasi dan pengetahuan lebih menekankan pada pengertian informasi dan pengetahuan sebagai sebuah proses yang bersambungan.

Hubungan informasi data dan pengetahuan dijelaskan oleh Sulistyo-Basuki (2011). Menurutnya, informasi dimulai dengan sebuah peristiwa (event), misalnya gunung meletus, bencana banjir, anak menangis, dan pegawai menerima gaji. Peristiwa itu direpresentasikan dalam bentuk simbol. Simbol ini dapat berupa teks, angka, suara, gambar, gabungan dua jenis simbol atau lebih, serta gabungan yang diatur dengan peraturan dan formulasi sehingga menjadi data. Data tersebut, bila

diterima oleh pancaindra manusia, hal itu berubah menjadi informasi. Bila informasi ini ditransfer ke manusia lain, hal itu berubah menjadi pengetahuan (knowledge). Manusia yang memperoleh pengetahuan akan menjadi (lebih) bijak (wise) daripada sebelumnya

3. Statistika Deskriptif

Suryoatmono (2004:18) menyatakan: Statistika Deskriptif adalah statistika yang menggunakan data pada suatu kelompok untuk menjelaskan atau menarik kesimpulan mengenai kelompok itu saja

a. Ukuran Lokasi: mode, mean, median, dll
b. Ukuran Variabilitas: varians, deviasi standar, range, dll
c. Ukuran Bentuk: skewness, kurtosis, plot boks

Pangestu Subagyo (2003:1) menyatakan: Yang dimaksud sebagai statistika deskriptif adalah bagian statistika mengenai pengumpulan data, penyajian, penentuan nilai-nilai statistika, pembuatan diagram atau gambar mengenai sesuatu hal, disini data yang disajikan dalam bentuk yang lebih mudah dipahami atau dibaca. Sudjana (1996:7) menjelaskan: Fase statistika dimana hanya berusaha melukiskan atau mengalisa kelompok yang diberikan tanpa membuat atau menarik kesimpulan tentang populasi atau kelompok yang lebih besar dinamakan statistika deskriptif.

Dalam materi statistika tidak terlepas dari data yang dapat dipercaya kebenarannya. Yang mana keseluruhan dari data tersebut disebut sebagai populasi.

Analisis deskriptif adalah bentuk analisis data penelitian untuk menguji generalisasi hasil penelitian yang didasarkan atas satu sampel. Analisis deskriptif ini dilakukan melalui pengujian hipotesis deskriptif. Hasil analisisnya adalah apakah hipotesis penelitian dapat digeneralisasikan atau tidak. Jika hipotesis nol (H0) diterima, berarti hasil penelitian dapat digeneralisasikan. Analisis deskriptif ini menggunakan satu variabel atau lebih tapi bersifat mandiri, karena itu analisis ini tidak berbentuk perbandingan atau hubungan.

Jenis teknik statistik yang digunakan untuk menguji hipotesis deskriptif harus sesuai dengan jenis data atau variable berdasarkan skala pengukurannya, yaitu nominal, ordinal , atau interval/rasio. Untuk menguji data nominal, digunakan dua cara yaitu :

Statistika Deskriptif adalah bagian statistika mengenai pengumpulan data, penyajian, penentuan nilai-nilai statistika, pembuatan diagram atau gambar mengenai suatu hal.

Jenis teknik statistik yang digunakan untuk menguji hipotesis deskriptif harus sesuai dengan jenis data atau variabel berdasarkan skala pengukurannya, yaitu nominal, ordinal , atau interval/rasio.

4. Mean, Modus, Standar Deviasi

a. Mean

Mean adalah ukuran yang sering disebut dengan istilah rata-rata, yang bias dicari dengan perhitungan jumlah nilai data dibagi banyak data observasi. Mengingat gugus data yang diamati bisa diperoleh dari populasi atau dari sampel, maka dibedakan antara rata-rata populasi dengan rata-rata sampel. Rata-rata populasi dilambangkan dengan μ (miyu), sedangkan ratarata sampel dilambangkan dengan x (x bar).

Rumus rata-rata hitung adalah sebagai berikut:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

No.	Nama	Nilai
1	Anto	50
2	Bayu	55
3	Cica	60
4	Deny	65
5	Elan	70
6	Fahri	75
7	Gina	80
8	Hana	85
9	Indy	90
Rata-rata Hitung		***70***

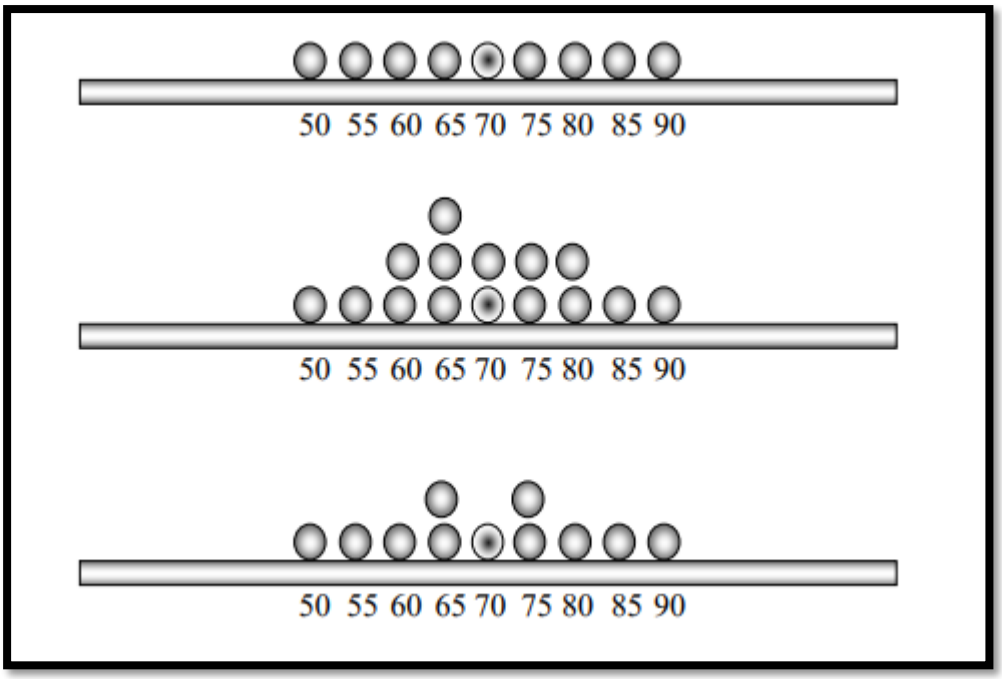
Rata-rata hitung pada baris paling bawah pada tabel di samping (= 70), di dapat dengan cara menjumlahkan nilai kesepuluh mahasiswa (50 + 55 + 60 + ... + 90) kemudian hasilnya dibagi dengan 9 (yaitu jumlah observasi). Pemahaman makna rata-rata hitung (selanjutnya kita sebut dengan : rata-rata), adalah sebagai berikut:

Dari ketiga gambaran di atas, kita lihat bahwa nilai rata-rata sangat tergantung pada besaran tiap-tiap data, termasuk apabila dalam data terdapat nilai ekstreem, yaitu nilai yang sangat kecil atau sangat besar dan jauh berbeda dari kelompok data.

b. Modus

Adalah nilai yang mempunyai frekuensi terbanyak dalam kumpulan data. Ukuran ini biasanya digunakan untuk mengetahui tingkat seringnya terjadi suatu peristiwa. (ukuran ini (sebenarnya) cocok digunakan untuk data berskala nominal.

Pada data yang tidak dikelompokkan, modus diperoleh dengan menghitung frekuensi dari masing-masing niai pengamatan, dan kemudian dicari nilai pengamatan yang mempunyai frekuensi observasi paling banyak (nilai data yang paling sering muncul). Penggambaran contoh data di atas akan kita ubah menjadi sebagai berikut:



i. Gambar 1, tidak terdapat modus karena semua data mempunyai frekuensi muncul yang sama.
ii. Gambar 2, modus sebesar 65 karena nilai ini yang paling sering muncul.
iii. Gambar 3, modus jatuh pada nilai 65 dan 75. Modus bisa muncul di lebih dari satu tempat, dan kita sebut bimodal.

c. Standar Deviasi

Standar deviasi disebut juga simpangan baku. Seperti halnya varians, standar deviasi juga merupakan suatu ukuran dispersi atau variasi. Standar deviasi merupakan ukuran dispersi yang paling banyak dipakai. Hal ini mungkin karena standar deviasi mempunyai satuan ukuran yang sama dengan satuan ukuran data asalnya. Misalnya, bila satuan data asalnya adalah cm, maka satuan stdar deviasinya juga cm. Sebaliknya, varians memiliki satuan kuadrat dari data asalnya (misalnya cm^2). Simbol standar deviasi untuk populasi adalah σ dan untuk sampel adalah s. Berikut terdapat empat (4) rumus dalam standar deviasi, diantaranya:

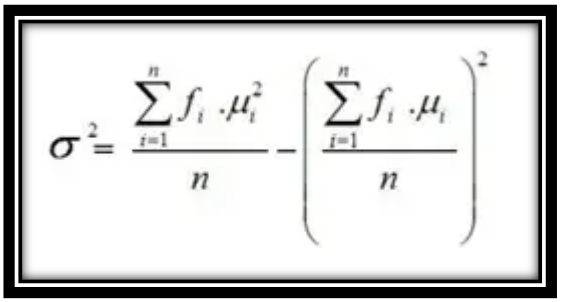
i. Rumus Standar Deviasi Data Tunggal

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

ii. Rumus Standar Deviasi Data Populasi

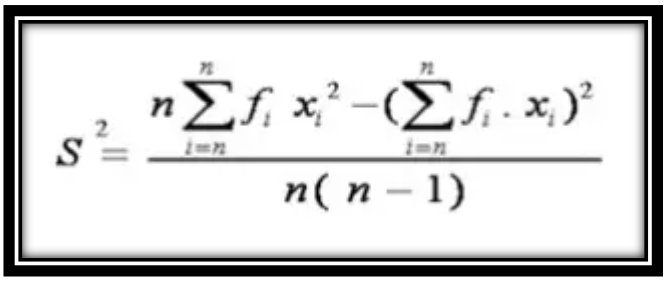
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

iii. Rumus Standar Deviasi Data Kelompok untuk Sampel



$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i \cdot \mu_i^2}{n} - \left(\frac{\sum_{i=1}^{n} f_i \cdot \mu_i}{n} \right)^2$$

iv. Rumus Standar Deviasi Data Kelompok untuk Populasi



$$S^2 = \frac{n\sum_{i=n}^{n} f_i \; x_i^{\,2} - (\sum_{i=n}^{n} f_i \cdot x_i)^2}{n(n-1)}$$

5. Perbedaan Data Science dan Artificial Intelligence

a. Data Science

adalah suatu disiplin ilmu yang khusus mempelajari data, khususnya data kuantitatif (data numeric), baik yang terstruktur ataupun yang tidak terstruktur. Berbagai subjek dalam data science meliputi semua proses data, mulai dari pengumpulan data, analisis data, pengolahan data, manajemen data, kearsipan, pengelompokan data, penyajian data, distribusi data, dan cara mengubah data menjadi kesatuan informasi yang dapat dipahami semua orang.

Ilmu-ilmu yang menjadi penunjang utama dalam ilmu data terdiri dari matematika, statistika, ilmu komputer, sistem informasi, manajemen, ilmu informasi, termasuk juga ilmu komunikasi dan kepustakaan. Bahkan ilmu ekonomi, terutama ilmu bisnis, juga berperan penting dalam ilmu data. Orang yang berkerja dengan Data Science, disebut data scientist. Data Scientist itu sudah ada yang sering disebut dengan statistikawan.

Oleh karena itu, tidak mengherankan jika data scientist sekarang lebih sering menciptakan algoritmaalgoritma di dalam program komputer agar data yang masuk dapat langsung diolah sendiri oleh komputer tersebut. Data Science merupakan ilmu yang sangat cepat berkembang dan di kembangkan oleh organisasi yang menggunakan konsep DDDDM.

Data Science memiliki siklus hidup atau stage dalam implementasinya. Siklus hidup yang dijalani dalam data science terdiri dari 5 stage yaitu:

i. Capture
   Siklus ini berhubungan dengan pengumpulan data
ii. Maintain
   Siklus ini berhubungan dengan data warehouse, data staging, data cleansing, arsitektur data.

iii. Process
Siklus ini berhubungan dengan data pemrosesan data, data mining, data modeling, data summarizing.

iv. Analyze
Siklus ini berhubungan dengan analisa data, data predictive.

v. Communicate
Siklus ini berhubungan dengan visualisasi data, data reporting, pengambilan keputusan

b. Artificial Intelligence (AI)

Artificial Intelligence (Kecerdasan Buatan) dapat menjadi pelengkap dari para manusia untuk dapat mengurangi tingkat pengambilan keputusan yang berdasarkan keyakinan pribadi (Bullock, 2019).

Penggunaan kecerdasan buatan dalam literatur telah banyak diimplementasikan salah satunya terhadap penggunaan kecerdasan buatan dalam manajemen sumber daya manusia, seperti proses penggalian informasi dari kandidat seleksi pegawai (Kaczmarek, Kowalkiewicz, & Piskorski, 2005); pemilihan pegawai berprestasi dengan menggunakan teknik penggalian data (Chien & Chen, 2008); dan pemetaan pengembangan pegawai menggunakan teknik agen intelijen (Giotopoulos, Alexakos, & Beligiannis, 2005). Penelitian tersebut berfokus pada level individu dan penggunaannya terbatas pada sektor swasta tetapi belum ada penggunaan secara spesifik mengenai penggunaan kecerdasan buatan pada proses pengawasan manajemen di sektor publik, oleh karena itu dilakukanlah penelitian ini untuk menutup gap tersebut.

Penelitian kecerdasan buatan di Indonesia sudah dimulai sejak tahun 1987, yang kala itu digunakan oleh BPPT untuk proyek sistem mesin penerjemah multi bahasa yang disponsori oleh pemerintah Jepang. Dari penelitian tersebut dilahirkan beberapa penelitian lainnya diantaranya proyek Universal Networking Language (UNL), ASEAN-MT, dll. Selain proyek penelitian kecerdasan buatan tersebut juga dilakukan komersialisasi dengan membuat sebuah produk yang mampu membuat notulensi rapat secara cepat dengan mencatat segala bentuk pembicaraan, produk tersebut dinamakan “Perisalah”. Selanjutnya penelitian tersebut dilanjutkan ke arah speech-to-speech dengan melakukan integrasi teknologi pengenalan wicara, mesin translasi dan pembangkit wicara (textto-speech synthesizer) (Riza, Nugroho, & Gunarso, 2020).

6. Perbedaan Data Mining dan Machine Learning

a. Data Mining

Data mining dianggap sebagai proses mengekstraksi informasi yang berguna dari sejumlah data yang besar. Data mining digunakan untuk menemukan pola baru, akurat, dan berguna dalam data, mencari makna dan informasi yang relevan untuk organisasi atau individu yang membutuhkannya. Data mining adalah salah satu tools yang digunakan oleh manusia saat ini.

Pada dasarnya, data mining punya empat fungsi utama, yaitu:

i. Prediksi – Setelah menemukan suatu pola dari kumpulan data, pola tersebut digunakan untuk memprediksi hasil yang terjadi di periode berikutnya.
ii. Deskripsi – Fungsinya untuk memahami karakteristik utama dari suatu data.
iii. Asosiasi – Berfungsi untuk menemukan hubungan antar data. Sehingga kamu bisa tahu apakah karakteristik suatu data dapat mempengaruhi data lainnya.
iv. Klasifikasi – Fungsinya untuk memberikan atribut tertentu pada suatu data. Jadi, datanya lebih mudah untuk diinterpretasikan.

Cara mempraktikan data mining:

i. Memahami Tujuan
ii. Mengumpulkan Data
iii. Menyiapkan Data
iv. Proses Modelling
v. Evaluasi Pola
vi. Penyajian Data

b. Machine Learning

adalah mesin yang dikembangkan untuk bisa belajar dengan sendirinya tanpa arahan dari penggunanya. Pembelajaran mesin dikembangkan berdasarkan disiplin ilmu lainnya seperti statistika, matematika dan data mining sehingga mesin dapat belajar dengan menganalisa data tanpa perlu di program ulang atau diperintah.

Dalam hal ini machine learning memiliki kemampuan untuk memperoleh data yang ada dengan perintah ia sendiri. ML juga dapat mempelajari data yang ada dan data yang ia peroleh sehingga bisa melakukan tugas tertentu. Tugas yang dapat dilakukan oleh ML pun sangat beragam, tergantung dari apa yang ia pelajari.

Istilah machine learning pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan dasar-dasar machine learning dan konsepnya. Sejak saat itu ML banyak yang mengembangkan. Salah satu contoh dari penerapan ML yang cukup terkenal adalah Deep Blue yang dibuat oleh IBM pada tahun 1996.

Deep Blue merupakan machine learning yang dikembangkan agar bisa belajar dan bermain catur. Deep Blue juga telah diuji coba dengan bermain catur melawan juara catur profesional dan Deep Blue berhasil memenangkan pertandingan catur tersebut.

Peran machine learning banyak membantu manusia dalam berbagai bidang. Bahkan saat ini penerapan ML dapat dengan mudah kamu temukan dalam kehidupan sehari-hari. Misalnya saat kamu menggunakan fitur face unlock untuk membuka perangkat smartphone kamu, atau saat kamu menjelajah di internet atau media sosial kamu akan sering disuguhkan dengan beberapa iklan. Iklan-iklan yang

dimunculkan juga merupakan hasil pengolahan ML yang akan memberikan iklan sesuai dengan pribadi kamu.

Sebenarnya masih banyak contoh dari penerapan machine learning yang sering kamu jumpai. Lalu pertanyaanya, bagaimana ML dapat belajar? ML bisa belajar dan menganalisa data berdasarkan data yang diberikan saat awal pengembangan dan data saat ML sudah digunakan. ML akan bekerja sesuai dengan teknik atau metode yang digunakan saat pengembangan. Apa saja tekniknya? Yuk kita simak bersama.

Ada beberapa teknik yang dimiliki oleh *machine learning*, namun secara luas ML memiliki dua teknik dasar belajar, yaitu *supervised* dan *unsupervised*.

i. Supervised Learning

Teknik supervised learning merupakan teknik yang bisa kamu terapkan pada pembelajaran mesin yang bisa menerima informasi yang sudah ada pada data dengan memberikan label tertentu. Diharapkan teknik ini bisa memberikan target terhadap output yang dilakukan dengan membandingkan pengalaman belajar di masa lalu.

ii. Unsupervised Learning

Teknik unsupervised learning merupakan teknik yang bisa kamu terapkan pada machine learning yang digunakan pada data yang tidak memiliki informasi yang bisa diterapkan secara langsung. Diharapkan teknik ini dapat membantu menemukan struktur atau pola tersembunyi pada data yang tidak memiliki label.

Sedikit berbeda dengan supervised learning, kamu tidak memiliki data apapun yang akan dijadikan acuan sebelumnya. Misalkan kamu belum pernah sekalipun membeli film sama sekali, akan tetapi pada suatu waktu, kamu membeli sejumlah film dan ingin membaginya ke dalam beberapa kategori agar mudah untuk ditemukan.

7. Algoritma Klasifikasi

Klasifikasi adalah tipe analisis data yang dapat membantu orang menentukan kelas label dari sampel yang ingin di klasifikasi. Klasifikasi merupakan Metode supervised learning, metode yang mencoba menemukan hubungan antara atribut masukan dan atribut target. Tujuan klasifikasi untuk meningkatkan kehandalan hasil yang diperoleh dari data.

Algoritma C4.5 Salah satu metode klasifikasi yang digunakan. Melibatkan konstruksi pohon keputusan, koleksi node keputusan. Setiap cabang kemudian mengarah ke node lain baik keputusan atau ke node daun untuk mengakhiri (Larose, 2005). C4.5 adalah algoritma yang mempunyai input berupa training samples berupa data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya dan samples yang merupakan field - field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data.

Decission Tree atau Pohon Keputusan adalah struktur sederhana yang dapat digunakan sebagai pengklasifikasi. Referensi penting dalam pengerjaan aslinya adalah Classification and Regression Tree oleh Breiman et al.

Pada pohon keputusan, masing-masing node internal (non-leaf) merepresentasikan sebuah variabel atribut (atribut prediksi atau fitur) dan masing-masing cabang merepresentasikan satu keadaan dari variabel ini. Masing-masing dari tiga daun (leaf) menspesifikasikan nilai yang diharapkan dari kelas variabel (variabel yang akan di prediksi). Aspek penting dari prosedur untuk membangun pohon keputusan adalah pemisahan kriteria (split criterion) termasuk kriteria untuk membuat cabang dan kriteria terakhir (stop criterion), kriteria yang digunakan untuk menghentikan pencabangan. Pohon keputusan dibuat menggunakan himpunan dari data yang digunakan sebagai data pembelajaran (training dataset). Himpunan yang berbeda yang disebut test dataset digunakan untuk melakukan pengujian untuk mengecek model. Pohon keputusan menawarkan banyak keuntungan, antara lain:

i. Fleksibilitas untuk berbagai tugas data mining, seperti klasifikasi, regresi, clustering dan seleksi fitur.
ii. Cukup jelas dan mudah diikuti (ketika dipadatkan).
iii. Fleksibilitas dalam menangani berbagai input data: nominal, numerik dan tekstual.
iv. Adaptasi di dataset pengolahan yang mungkin memiliki kesalahan atau nilai-nilai yang hilang.
v. Kinerja prediktif tinggi untuk upaya komputasi yang relatif kecil. 6. Tersedia dalam berbagai paket data mining melalui berbagai platform 7. Berguna untuk dataset besar (dalam kerangka ensemble).

8. K-Nearest Neighbour

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masingmasing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran. Sebuah titik pada ruang ini ditandai kelas c jika kelas c merupakan klasifikasi yang paling banyak ditemui pada k buah tetangga terdekat titk tersebut. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Euclidean dengan rumus seperti dibawah

$$distance = \sqrt{\sum_{i=1}^{n} (X^i_{training} - X_{testing})^2}$$

dengan

$X^i_{training}$: data training ke-*i*,
$X_{testing}$: data testing,
i : record (baris) ke-*i* dari tabel,
n : jumlah data training.

Pada fase pembelajaran, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi dari data pembelajaran. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk data test (yang klasifikasinya tidak diketahui). Jarak dari vektor yang baru ini terhadap seluruh vektor data pembelajaran dihitung, dan sejumlah K buah yang paling dekat diambil. Titik yang baru klasifikasinya diprediksikan termasuk pada klasifikasi terbanyak dari titik-titik tersebut. Nilai K yang terbaik untuk algoritma ini tergantung pada data. Secara umum, nilai K yang tinggi akan mengurangi efek noise pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur. Nilai K yang bagus dapat dipilih dengan optimasi parameter, misalnya dengan menggunakan cross-validation. Kasus khusus di mana klasifikasi diprediksikan berdasarkan data pembelajaran yang paling dekat (dengan kata lain, K = 1) disebut algoritma nearest neighbour.

Ketepatan algoritma KNN sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Algoritma KNN memiliki beberapa kelebihan yaitu ketangguhan terhadap training data yang memiliki banyak noise dan efektif apabila training data-nya besar. Sedangkan, kelemahan KNN adalah KNN perlu menentukan nilai dari parameter K (jumlah dari tetangga terdekat), training berdasarkan jarak tidak jelas mengenai jenis jarak apa yang harus digunakan dan atribut mana yang harus digunakan untuk mendapatkan hasil terbaik, dan biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap query instance pada keseluruhan training sample.

K-Nearest Neighbour (KNN) adalah suatu metode yang menggunakan algoritma supervised dimana hasil dari query instance yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Tujuan dari algoritma ini adalah mengklasifikasikan obyek baru bedasarkan atribut dan training sample. Classifier tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik query, akan ditemukan sejumlah K obyek atau (titik training) yang paling dekat dengan titik query. Klasifikasi menggunakan voting terbanyak diantara klasifikasi dari K obyek. Algoritma KNN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari query instance yang baru.

X_1 = Asam Durabilitas (detik)	X_2 = Kekuatan (Kg / meter persegi)	Klasifikasi
7	7	Buruk
7	4	Buruk
3	4	Baik
1	4	Baik

Sebagai kasus, misalnya saat ini pabrik kertas telah menghasilkan jaringan baru yang lulus uji laboratorium dengan X1=3 dan X2=7. Untuk menebak klasifikasi jaringan baru ini maka dilakukan perhitungan dengan menggunakan algoritma KNN.

Adapun langkah–langkah untuk menghitung K tetangga terdekat dengan algoritma KNN adalah sebagai berikut.

a. Tentukan parameter K (jumlah tetangga terdekat). Misalkan K = 3

b. Hitung jarak antara permintaan (data testing) dan contoh-contoh latihan semua (data training). Data training yang akan dihitung kedekatannya mempunyai koordinat (3,7).

X_1 = Asam Durabilitas (detik)	X_2 = Kekuatan (Kg / meter persegi)	Square Jarak ke contoh permintaan (3, 7)
7	7	$(7-3)^2 + (7-7)^2 = 16$
7	4	$(7-3)^2 + (4-7)^2 = 25$
3	4	$(3-3)^2 + (4-7)^2 = 9$
1	4	$(1-3)^2 + (4-7)^2 = 13$

c. Urutkan jarak dan menentukan tetangga terdekat berdasarkan jarak terdekat ke-K.

X_1 = Asam Durabilitas (detik)	X_2 = Kekuatan (Kg / meter persegi)	Square Jarak ke contoh permintaan (3, 7)	Peringkat jarak minimum	Apakah termasuk dalam-tetangga terdekat 3?
7	7	$(7-3)^2 + (7-7)^2 = 16$	3	Ya
7	4	$(7-3)^2 + (4-7)^2 = 25$	4	Tidak
3	4	$(3-3)^2 + (4-7)^2 = 9$	1	Ya
1	4	$(1-3)^2 + (4-7)^2 = 13$	2	Ya

d. Kumpulkan kategori Y dari baris tetangga terdekat. Pada baris kedua kategori tetangga terdekat (Y) tidak dimasukkan karena data tersebut peringkatnya lebih dari 3 tetangga terdekat.

7	7	$(7-3)^2 + (7-7)^2 = 16$	3	Ya	Buruk
7	4	$(7-3)^2 + (4-7)^2 = 25$	4	Tidak	-
3	4	$(3-3)^2 + (4-7)^2 = 9$	1	Ya	Baik
1	4	$(1-3)^2 + (4-7)^2 = 13$	2	Ya	Baik

e. Gunakan mayoritas sederhana dari kategori tetangga terdekat sebagai nilai prediksi contoh query

Dari tabel di atas diperoleh dua kertas tisu baru berkualitas baik dan satu kertas tisu baru berkualitas buruk. Karena tetangga terdekat yang didapat lebih banyak yang berkualitas baik, maka dapat disimpulkan bahwa kertas tisu baru yang lulus uji laboratorium dengan X1=3 dan X2=7 adalah termasuk dalam kategori baik.

9. Algoritma C4.5

kenal dan digunakan untuk klasifikasi data yang memiliki atribut-atribut numerik dan kategorial. Hasil dari proses klasifikasi yang berupa aturan-aturan dapat digunakan untuk memprediksi nilai atribut bertipe diskret dari record yang baru. Algortima C4.5 sendiri merupakan pengembangan dari algortima ID3, dimana pengembangan dilakukan dalam hal, bisa mengatasi missing data, bisa mengatasi data kontinu dan pruning.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

a. Pilih atribut sebagai akar.
b. Buat cabang untuk tiap-tiap nilai.
c. Bagi kasus dalam cabang.
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunaka rumus seperti yang tertera dalam persamaan:

Gain (S.A) = Entropy (S) - Zi; =i Entropy (Si)

Dimana : S : himpunan kasus A : (Elisa, 2017)atribut N : jumlah partisi atribut A \Si\ : jumlah kasus pada partisi ke-i |Sr | : jumlah kasus dalam S

10. Naive Bayes

Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan Naive dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi Naive Bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya.

Persamaan dari teorema Bayes adalah :

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Keterangan :

a. X : Data dengan class yang belum diketahui
b. H : Hipotesis data X merupakan suatu class spesifik
c. $P(H|X)$: Probabilitas hipotesis H berdasar kondisi$X$ (posteriori probability)
d. $P(H)$: Probabilitas hipotesis H (prior probability)
e. $P(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H
f. $P(X)$: Probabilitas X

Untuk menjelaskan teorema Naive Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, teorema bayes di atas disesuaikan sebagai berikut :

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots . F_n|C)}{P(F_1 . \dots F_n)}$$

Dimana Variabel C merepresentasikan kelas, sementara variabel F1 ... Fnmerepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (Posterior) adalah peluang munculnya kelas C (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik – karakteristik sampel pada kelas C (disebut juga likelihood), dibagi dengan peluang kemunculan karakteristik – karakteristik sampel secara global (disebut juga evidence). Karena itu, rumus diatas dapat pula ditulis secara sederhana sebagai berikut :

$$Posterior = \frac{Prior \times likelihood}{evidence}$$

Nilai Evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai – nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan.Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan($C|F1, \dots , Fn$) menggunakan aturan perkalian sebagai berikut :

$$\begin{aligned} P(C|F_1, \dots, F_n) &= P(C)\, P(F_1, .., F_n|C) \\ &= P(C)P(F_1|C)P(F_2, \dots, F_n|C, F_1) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \dots, F_n|C, F_1, F_2) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2), P(F_4, \dots, F_n|C, F_1, F_2, F_3) \\ &= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2) \dots P(F_n|C, F_1, F_2, F_3, .., F_{n-1}) \end{aligned}$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor – faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (naif), bahwa masing – masing petunjuk ($\boldsymbol{F1}$, $\boldsymbol{F2}$ … $\boldsymbol{Fn}$) saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut :

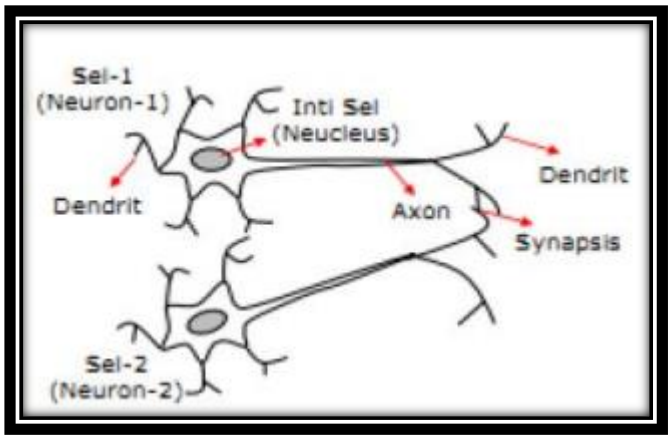
$$P(P_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

Untuk $i \neq j$, sehingga
$$P(F_i|C, F_j) = P(F_i|C)$$

Dari persamaan diatas dapat disimpulkan bahwa asumsi independensi naif tersebut membuat syarat peluang menjadi sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan.

11. Jaringan Syaraf Tiruan

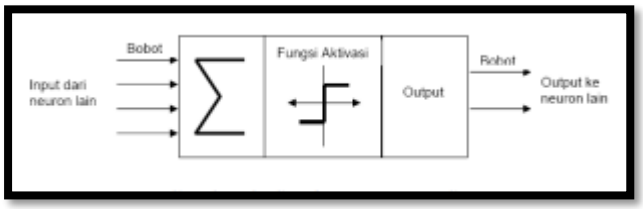
Jaringan Syaraf Tiruan (JST) merupakan suatu sistem pemrosesan informasi yang mempunyai karakteristik menyerupai jaringan syaraf biologis (JSB) Jaringan Syaraf Tiruan tercipta sebagai suatu generalisasi model matematis dari pemahaman manusia (human cognition) (Maharani Dessy Wuryandari, 2012)



Seperti halnya otak manusia, jaringan syaraf juga terdiri dari beberapa neuron, dan terdapat hubungan antara neuron-neuron tersebut. menunjukkan struktur neuron yang mana Neuron-neuron akan mentransformasikan informasi yang diterima melalui sambungan keluarannya menuju ke neuron-neuron yang lain.

Pada jaringan syaraf hubungan ini dikenal dengan nama bobot. Informasi tersebut tersimpan pada suatu nilai tertentu pada bobot tersebut.

keduanya atau mungkin lebih untuk mendapatkan redundansi data. Ini diproses oleh suatu fungsi perambatan yang akan menjumlahkan nilai-nilai semua bobot yang akan datang. Hasil penjumlahan ini kemudian dibandingkan dengan suatu Informasi yang disebut dengan masukkan dikirim ke neurondengan bobot kedatangan tertentu. Masukkan nilai ambang (threshold) tertentu melalui fungsi aktivasi setiap neuron.



12. Deep Learning

Deep learning merupakan subbidang *machine learning* yang algoritmanya terinspirasi dari struktur otak manusia. Saat ini, teknik *deep learning* sangat populer di

kalangan praktisi data dan menarik perhatian banyak pihak. Hal ini karena teknologi *deep learning* telah diterapkan dalam berbagai produk berteknologi tinggi seperti self-driving car. Selain itu, ia juga ada di balik produk dan layanan yang kita gunakan sehari-hari. Contohnya antara lain, asisten digital, Google Translate, dan *voice-activated device* (perangkat cerdas yang bisa diaktifkan dengan suara).

D*eep learning* terdiri dari beberapa jaringan saraf tiruan yang saling berhubungan. Berikut ini adalah beberapa algoritmanya:

a. *Convolutional Neural Network* (CNN)

CNN terdiri dari banyak *layer* untuk memproses dan mengekstrak fitur dari data. Ia biasanya digunakan untuk memproses gambar dan mendeteksi objek. Saat ini, CNN banyak digunakan untuk mengidentifikasi citra satelit, citra medis, dan mendeteksi anomali.

b. *Recurrent Neural Network* (RNN)

Recurrent Neural Networks (RNN) merupakan salah satu bentuk arsitektur *Artificial Neural Network*s (ANN) yang dirancang khusus untuk memproses data yang bersambung/ berurutan (*sequential data*). RNN biasanya digunakan untuk menyelesaikan permasalahan data historis atau time series, contohnya data ramalan cuaca. Selain itu, RNN juga dapat diimplementasikan pada bidang *natural language understanding* (pemahaman bahasa alami), misalnya translasi bahasa.

c. *Long Short Term Memory Network* (LTSM)

LSTM merupakan tipe *Recurrent Neural Network* yang dapat mempelajari data historis atau time series. Ia merupakan algoritma deep learning yang kompleks dan dapat mempelajari informasi jangka panjang dengan sangat baik. LSTM sangat powerful untuk menyelesaikan berbagai permasalahan kompleks seperti speech recognition, speech to text application, komposisi musik, dan pengembangan di bidang farmasi.

d. ***Self Organizing Maps* (SOM)**

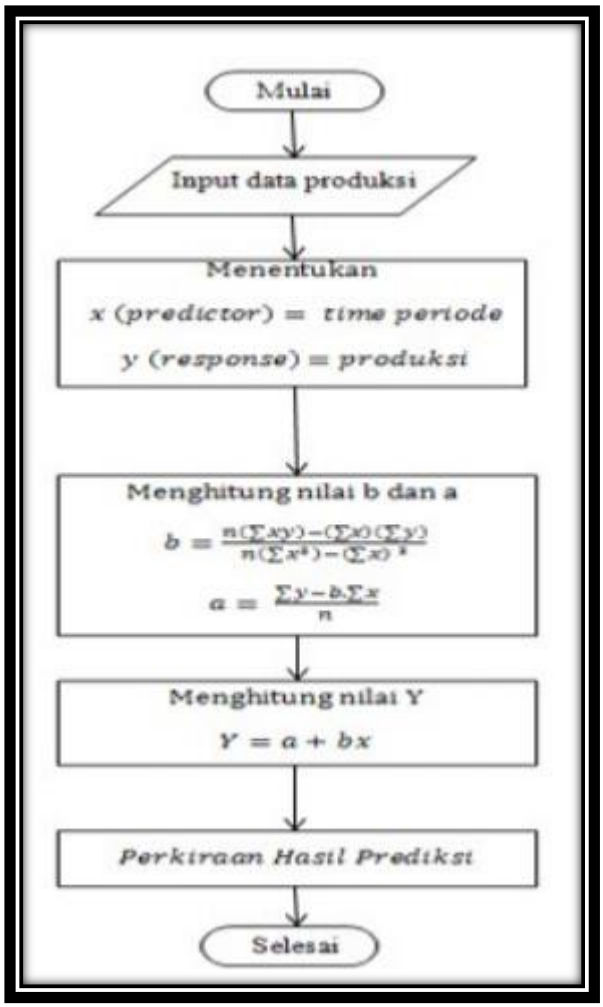
Jenis terakhir adalah *self organizing maps* atau SOM. Algoritma ini mampu membuat visualisasi data secara mandiri. SOM diciptakan untuk membantu penggunanya dalam memahami data dan informasi berdimensi tinggi.

Berikut adalah beberapa penerapannya:

a. Pengenalan gambar
b. Pengenalan suara
c. *Natural language processing*
d. Deteksi anomaly

13. Metodi Regresi

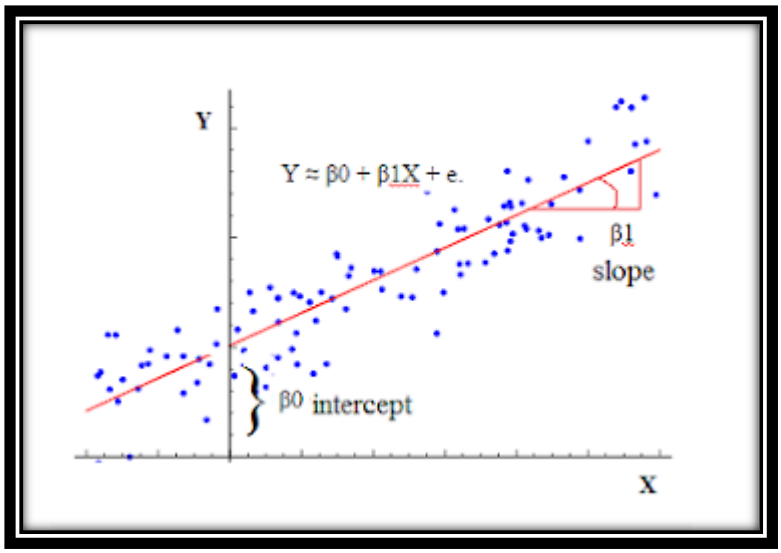
Regresi linear sederhana adalah metode statistik yang berfungsi untuk menguji sejauh mana hubungan sebab akibat antar variabel faktor penyebab (X) terhadap variabel akibatnya. Faktor penyebab pada umunya dilambangkan dengan x atau disebut juga dengan prediktor, sedangkan variabel akibat dilambangkan dengan y atau disebut juga dengan respon. Penyelesain menggunakan regresi linear sederhana



Metode regresi linear adalah alat statistik yang dipergunakan untuk mengetahui pengaruh antara satu atau beberapa variabel terhadap satu buah variabel. Manfaat dari regresi linear diantaranya analisis regresi lebih akurat dalam melakukan analisis korelasi, karena analisis itu kesulitan dalam menunjukan tingkat perubahan suatu variabel terhadap variabel lainnya (slop) dapat ditentukan.

14. Regresi Linier

digunakan untuk melihat bentuk hubungan antar variabel melalui suatu persamaan. Terdapat tiga jenis regresi yang digunakan sesuai dengan tujuan analisis yaitu Regresi Linier Sederhana, Regresi Linier Berganda, dan Regresi non Linear. Hubungannya bisa berupa hubungan sebab akibat selain itu juga dapat mengukur seberapa besar suatu variabel mempengaruhi variabel lain dan dapat digunakan untuk melakukan peramalan nilai suatu variabel berdasarkan variabel lain.



Dalam regresi linear sederhana hubungan variabel tersebut dapat dituliskan dalam bentuk model persamaan linear:

$$\hat{y} = a + bx$$

Dalam prosedur regresi hal pertama yang harus dilakukan adalah melakukan **identifikasi model** dengan menggunakan Scatter plot (diagram pencar) yang berguna untuk mengidentifikasi model hubungan antara variabel X dan Y. Bila pencaran titik-titik pada plot ini menunjukkan adanya suatu kecenderungan (trend) yang linier, maka model regresi linier layak digunakan. Setelah itu dapat dilakukan estimasi terhadap parameter model.

15. Algoritma Clustering

Clustering atau klasterisasi adalah metode pengelompokan data. Menurut Tan, 2006 clustering adalah sebuah proses untuk mengelompokan data ke dalam beberapa cluster atau kelompok sehingga data dalam satu cluster memiliki tingkat kemiripan yang maksimum dan data antar cluster memiliki kemiripan yang minimum.

Clustering merupakan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan cluster. Objek yang di dalam cluster memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan cluster yang lain. Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma clustering. Oleh karena itu, clustering sangat berguna dan bisa menemukan group atau kelompokyang tidak dikenal dalam data. Clustering banyak digunakan dalam berbagai aplikasi seperti misalnya pada business inteligence, pengenalan pola citra, web search, bidang ilmu biologi, dan untuk keamanan (security). Di dalam business inteligence, clustering bisa mengatur banyak customer ke dalam banyaknya kelompok. Contohnya mengelompokan customer ke dalam beberapa cluster dengan kesamaan karakteristik yang kuat. Clustering juga dikenal sebagai data segmentasi karena clustering mempartisi banyak data set ke dalam banyak group berdasarkan kesamaannya. Selain itu clustering juga bisa sebagai outlier detection.

a. *Clustering* merupakan metode segmentasi data yang sangat berguna dalam prediksi dan analisa masalah bisnis tertentu. Misalnya Segmentasi pasar, marketing dan pemetaan zonasi wilayah.
b. Identifikasi obyek dalam bidang berbagai bidang seperti computer vision dan image processing

Hasil *clustering* yang baik akan menghasilkan tingkat kesamaan yang tinggi dalam satu kelas dan tingkat kesamaan yang rendah antar kelas. Kesamaan yang dimaksud merupakan pengukuran secaranumeric terhadap dua buah objek. Nilai kesamaan antar kedua objek akan semakin tinggi jika kedua objek yang dibandingkan memiliki kemiripan yang tinggi. Begitu juga dengan sebaliknya. Kualitas hasil *clustering* sangat bergantung pada metode yang dipakai. Dalam *clustering* dikenal empat tipe data. Keempat tipe data pada tersebut ialah:

a. Variabel berskala interval
b. Variabel biner
c. Variabel nominal, ordinal, dan rasio
d. Variabel dengan tipe lainnya.

Metode *clustering* juga harus dapat mengukur kemampuannya sendiri dalam usaha untuk menemukan suatu pola tersembunyi pada data yang sedang diteliti. Terdapat berbagai metode yang dapat digunakan untuk mengukur nilai kesamaan antar objek-objek yang dibandingkan. Salah satunya ialah dengan *weighted Euclidean Distance*. Euclidean distance menghitung jarak dua buah point dengan mengetahui nilai dari masing-masing atribut pada kedua poin tersebut. Berikut formula yang digunakan untuk menghitung jarak dengan Euclidean distance:

$$Distance(p,q) = \left(\sum_{k}^{n} \mu k \, |Pk - qk|r\right) 1/r$$

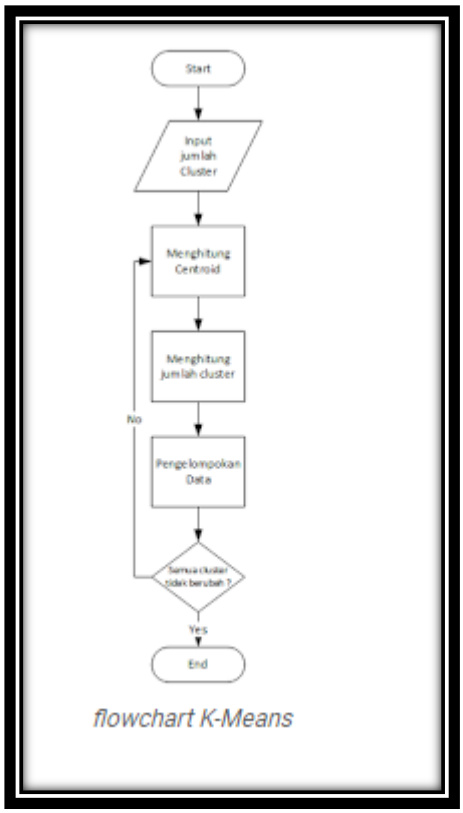
16. K-Means Clustering

K-Means yaitu salah satu dari metode pengelompokan data nonhierarki (sekatan) yang dapat mempartisi data kedalam bentuk dua kelompok ataupun lebih. Metoda tersebut akan mempartisi data kedalam suatu kelompok dimana data yang berkarakteristik sama akan dimasukkan kedalam satu kelompok sama sedangkan data yang memiliki karakteristik yang berbeda akan dikelompokan kedalam kelompok lainnya. Tujuan dari pengelompokan yaitu untuk meminimalkan dari fungsi objektif yang diset dalam proses pengelompokan, pada umumnya akan berusaha meminimalkan variasi didalam suatu kelompok dan memaksimalkan variasi antar kelompok.
Secara sederhana algoritma K-Means dimulai dari tahap berikut :

a. Pilih K buah titik centroid.
b. Menghitung jarak data dengan centroid.

c. Update nilai titik centroid.
d. Ulangi langkah 2 dan 3 sampai nilai dari titik centroid tidak lagi berubah.

Kita coba gambarkan dalam sebuah flowchart, agar kita lebih mudah memahami algoritma K-Means. Berikut adalah gambaran flowchartnya :



flowchart K-Means

Jadi dari flowchart diatas, kita memiliki input dan 3 buah proses. Yaitu pertama adalah proses menghitung centroid, kemudian proses kedua menghitung data yang akan dikelompokkan dengan centroid, kemudian proses ketiga adalah mengelompokkan data berdasarkan jarak terdekat (minimum distance). Dan kita membuat perulangan dengan kondisi "apakah posisi centroid tetap dan tidak ada perubahan terhadap datanya ?" apabila ya maka kita selesai melakukan pengelompokan. Tapi apabila masih ada perubahan centroid maka kita update kembali nilai centroid melalui proses pertama.

17. Algoritma Asosiasion Rule

Dengan menggunakan algoritma assosiation rule, konsumen dapat mengetahui dengan cepat produk mana yang sering dibeli dan mungkin dibutuhkan. Saat ini proses bisnis yang ada di toko Delta Jaya Motor masih konvesional, seperti penulisan nota transaksi penjualan menggunakan kwitansi yang ditulis tangan dan belum tersedianya layanan informasi teknologi untuk menampilkan sparepart dan stok produk. Oleh karena itu, toko Delta Jaya Motor memerlukan suatu sarana informasi dan layanan penjualan berbasis website yang bisa memenuhi kebutuhan dan memudahkan perusahaan dalam menjalankan kegiatan bisnis. Pemanfaatan teknologi informasi dapat memudahkan transaksi serta memberikan informasi tentang produk-produk unggulan

agar penjualan toko semakin meningkat (Afif, Swedia, & Cahyanti, 2019). Langkah-langkah penyelesaiannya adalah sebagai berikut :

a. Mencari semua hubungan antar item berdasarkan frekuensi pembelian;
b. Menentukan aturan asosiasi yang memenuhi kriteria support dan confidence berdasarkan frekuansi pembelian antar item.

18. Algoritma Apriori

Menurut (Erwin, 2009) dalam (Saefudin & DN, 2019) Algoritma Apriori adalah salah satu algoritma yang melakukan pencarian frequensi itemset dengan menggunakan teknik association rule. Algoritma Apriori menggunakan pengetahuan frekuensi atribut yang telah diketahui sebelumnya untuk memproses informasi selanjutnya. Pada algoritmaApriori menentukan kandidat yang mungkin muncul dengan cara memperhatikan minimum support dan minimum confidence. Support adalah nilai pengunjung atau persentase kombinasi sebuah item dalam database. Dijelaskan dalam jurnal tersebut bahwa Algoritma Apriori tersebut digunakan untuk menentukan jenis ikan yang paling diminati oleh konsumen, sehingga dapat membantu UD. Mumu Jaya Pandeglang mengetahui ikan yang paling banyak diminati untuk menentukan persediaan stoknya.

DAFTAR PUSTAKA

Bustami. (2014). PENERAPAN ALGORITMA NAIVE BAYES UNTUK MENGKLASIFIKASI DATA NASABAH ASURANSI. *JURNAL INFORMATIKA Vol. 8, No. 1, Januari 2014*, 15.

Darono, A. (2016, Oktober 27). PEMBELAJARAN BUSINESS ANALYTICS DAN BIG DATA DALAM. p. 15.

Dra. Sri Ati, M., Prof. Dr. Nurdien , H. Kistanto, M.A., & Amin Taufik,S.Sos. . (n.d.). Pengantar Konsep Informasi, Data,.

Elisa, E. (2017). Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti. *JOIN | Volume 2 No. 1 | Juni 2017*, 6.

HENDRIAN, S. (n.d.). ALGORITMA KLASIFIKASI DATA MINING UNTUK MEMPREDIKSI . *Faktor Exacta 11 (3): 266-274, 2018*, 9.

Mohamad, A. A., R., D. M., I, P. E., & Septian, R. W. (2020). Penggabungan Teknologi Untuk Analisa Data Berbasis Data Science. *ISBN: 978-602-52720-7-3*, 5.

Narendra, A. P. (2015, Desember). Big Data, Data Analyst, and Improving the Competence of Librarian. *Volume 1, Nomor 2, Juli – Desember 2015*, p. 11.

Nasution, L. M. (2017). STATISTIK DESKRIPTIF . *Jurnal Hikmah, Volume 14, No. 1, Januari – Juni 2017, ISSN :1829-8419*, 7.

NILAI RINGKASAN DATA. (n.d.). *bab 3 : Nilai Rigkasan Data*, 16.

Sudarsono, A. (2016). JARINGAN SYARAF TIRUAN UNTUK MEMPREDIKSI LAJU PERTUMBUHAN PENDUDUK MENGGUNAKAN METODE BACPROPAGATION (STUDI KASUS DI KOTA BENGKULU) . *Jurnal Media Infotama Vol. 12 No. 1, Februari 2016*, 9.

∗, W. Y. (2012). Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah . *Jurnal Matematika, Statistik dan Komputasi*, 12.