MODUL KULIAH
DATA MINING

PROGRAM STUDI : Teknik Informatika
KODE MATA KULIAH : SK219205
MATA KULIAH : Data Mining
SEMESTER : 4
SKS : 2 SKS
DOSEN PENGAMPU : Herdiesel Santoso, S.T., S.Kom., M.Cs.

STMIK EL RAHMA YOGYAKARTA			
PENYUSUN	**REVISI KE**	**TGL. PENYUSUNAN**	**DIPERIKSA OLEH**
Herdiesel Santoso, S.T., S.Kom., M.Cs.	I	10 Januari 2020	Ketua Program Studi Teknik Informatika

HALAMAN PENGESAHAN

PRAKATA

Alhamdulillahi Rabbil Alamin penulis ucapkan sebagai tanda syukur yang dalam kepada Allah SWT atas limpahan rahmat, karunia, serta petunjuk-Nya sehingga penulis dapat menyusun Diktat Data Mining yang merupakan bagian dari media pembelajaran kuliah Data Mining bagi dosen dan mahasiswa. Diktat ini diharapkan dapat dimanfaatkan oleh mahasiswa dan dapat menjadi acuan belajar yang lebih baik.

Penulis menyadari bahwa penyusunan Diktat Data Mining ini tidak akan terwujud tanpa adanya bantuan, bimbingan, dan dorongan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, pada kesempatan ini penulis mengucapkan terima kasih kepada:

1. Eko Riswanto, S.T., M.Cs., selaku Ketua STMIK El Rahma Yogyakarta.
2. Momon Muzakkar, S.T., M.Eng., selaku Wakil Ketua Bidang Akademik STMIK El Rahma Yogyakarta.
3. Wahyu Widodo, S.Kom.,M.Kom selaku Ketua Program Studi Teknik Informatika STMIK El Rahma Yogyakarta.
4. Keluarga dan Rekan-rekan sesama Dosen yang telah memberikan kondisi yang positif untuk penyusunan Diktat Data Mining ini.
5. Mahasiswa Kami : Rizky, Aminur, Huda, Tri Widayanti, Yulistiana, Linda Pratiwi, Risa, Emasetyawati, Loo Cantik, Eva Rianti yang sudah membantu menyiapkan materi dalam Diktat Ini.
6. Mahasiswa Kelas Data mining STMIK El Rahma yang sudah membaca dan mendiskusikan materi dalam Diktat ini.

Penulis juga menyadari bahwa dalam Diktat Data Mining ini masih terdapat banyak kekurangan. Untuk itu, perlu adanya saran, kritik yang konstruktif, maupun revisi untuk Diktat Data Mining berikutnya. Semoga Diktat Data Mining ini dapat memberikan manfaat bagi pihak yang membacanya.

Yogyakarta, 14 Februari 2020

Herdiesel Santoso S.T., S.Kom., M.Cs

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

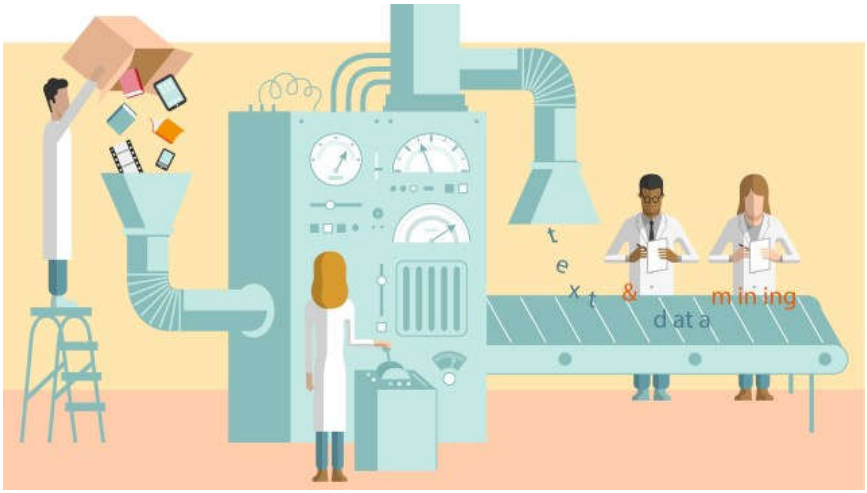
LAMPIRAN

BAB I
Pengantar Datamining

1.1 Pengertian Data Mining

Saat ini terjadi fenomena yaitu berupa data yang melimpah, setiap hari banyak orang yang berurusan dengan data yang bersumber dari berbagai jenis observasi dan pengukuran. Misalnya data yang menjelaskan karakteristik spesies makhluk hidup, data yang menggambarkan ciri-ciri fenomena alam, data yang berasal dari ringkasan hasil eksperimen ilmu pengetahuan, dan data yang mencatat perfoma suatu mesin.

Teknologi database saat ini memungkinkan untuk menyimpan sejumlah data dalam jumlah yang sangat besar dan terakumulasi. Disinilah awal timbulnya persoalan ledakan data (jumlah data yang tiba-tiba begitu sangat besar). Data perlu disimpan, tapi yang lebih penting dari itu adalah proses penemuan pengetahuan (knowledge) dari data yang disimpan. Oleh karenanya data yang tersimpan dalam sebuah gudang data yang disebut dengan data warehouses perlu dianalisa.



Gambar 1.1 Proses Data mining

Penggalian data (bahasa Inggris: data mining) adalah disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar. Suatu pola dikatakan menarik apabila pola tersebut tidak sepele, implisit, tidak diketahui sebelumnya, dan berguna. Proses penggalian data tersebut dilakukan secara semi otomatik, menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstrasi dan menidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat

yang tersimpan di dalam database besar. Ada banyak nama lain dari data mining seperti: *Knowledge discovery databases* (KDD), *knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence.* Data Mining diperlukan saat data yang tersedia terlalu banyak (misalnya data yang diperoleh dari sistem basis data perusahaan, *e-commerce*, data saham, dan data bioinformatika), tapi tidak tahu pola apa yang bisa didapatkan.

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu – ilmu lain, seperti database system, data warehousing, statistik, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing. Karakteristik *data mining* sebagai berikut:

1. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. *Data mining* biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dapat dipercaya.
3. *Data mining* berguna untuk membuat keputusan kritis.

1.2 Pentingnya Data Mining

Beberapa faktor yang mendukung perlunya dilakukan data mining adalah :

1. Data telah mencapai jumlah dan ukuran yang sangat besar

Hasil dan proses data mining merupakan suatu informasi yang akan mendasari tindakan tertentu sehingga tingkat kebenaran informasi tersebut menjadi sangat signifikan, dan makin besar serta makin banyak data yang digunakan maka akan semakin valid hasilnya. Perkembangan data dalam hal jumlah dan ukuran telah mencapai kecepatan yang sangat cepat, sehingga ukuran basis data yang dimiliki oleh sebuah perusahaan bisa mencapai kisaran gigabyte atau bahkan terabyte.

2. Telah dilakukan proses data warehousing

Untuk mencapai hasil yang memuaskan, maka sumber data yang digunakan dalam proses data mining seringkali merupakan data gabungan dari banyak departemen, daerah operasi bahkan dari sumber-sumber lain seperti data kependudukan. Oleh karena itu maka disarankan perlunya proses data warehousing untuk menjaga konsistensi, memberikan prespektif yang lebih baik terhadap data dan menjaga integritas data.

3. Kemampuan Komputasi yang semakin terjangkau

Pada dasarnya proses data mining memerlukan komputasi dan sumberdaya data yang sangat besar dalam pengolahanya. Penurunan harga yang cukup cepat terhadap perangkat keras komputer serta semakin tingginya kinerja yang berhasil dicapai oleh perangkat komputer maupun teknologi pengolahan data seperti teknologi pararel proses, menjadikan proses data mining suduh cukup layak untuk dilakukan secara komersial.

4. Persaingan bisnis yang semakin ketat

Tekanan persaingan bisnis yang semakin ketat mendorong perusahaan-perusahaan untuk selalu berinovasi agar mampu meningkatkan daya saingnya dipasar global. Beberapa tren yang berkembang saat ini adalah

a. Setiap bisnis adalah bisnis pelayanan
b. Adanya fenomena kustomisasi produk oleh masyarakat
c. Informasi adalah produk

1.3 Tujuan Data Mining

Tujuan dari data mining (Hoffer, Prescott, dan McFadden, 2007) adalah:

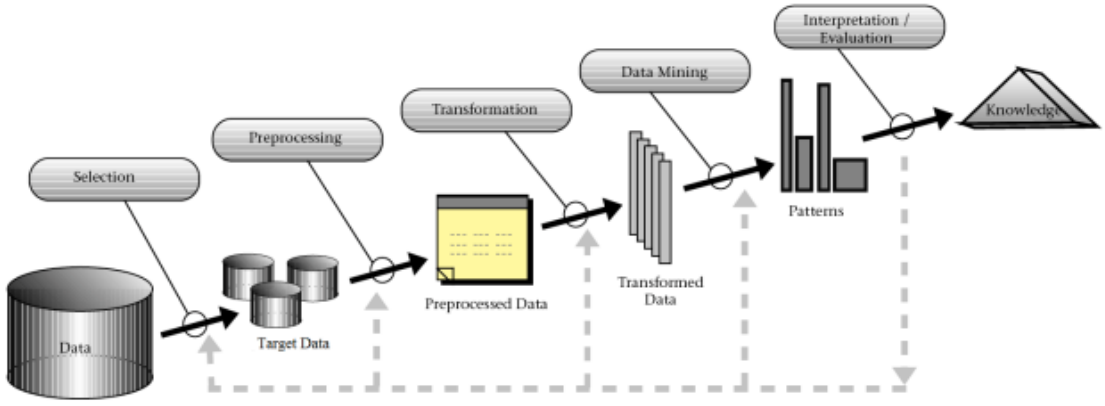
a. Explanatory
Untuk menjelaskan beberapa kondisi penelitian, seperti mengapa penjualan truk pick-up meningkat di Colorado.

b. Confirmatory
Untuk mempertegas hipotesis, seperti halnya dua kali pendapatan keluarga lebih suka dipakai untuk membeli peralatan keluarga dibandingkan dengan satu kali pendapatan keluarga.

c. Exploratory
Untuk menganalisa data yang memiliki hubungan yang baru. Misalnya, pola apa yang cocok untuk kasus penggelapan kartu kredit.

1.4 Tahapan Data Mining

Istilah data mining dan knowledge discovery in databases (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses pentggalian informasi tersembunyi dalam suatau basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Proses KDD secara garis besar dapat dijelaskan sebagai berikut:



Gambar 1.2 Tahapan Data mining

a. Seleksi Data

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional. Datamining bisa menggunakan sekumpulan dataset. Jenis dataset ada dua: Private dan Public.

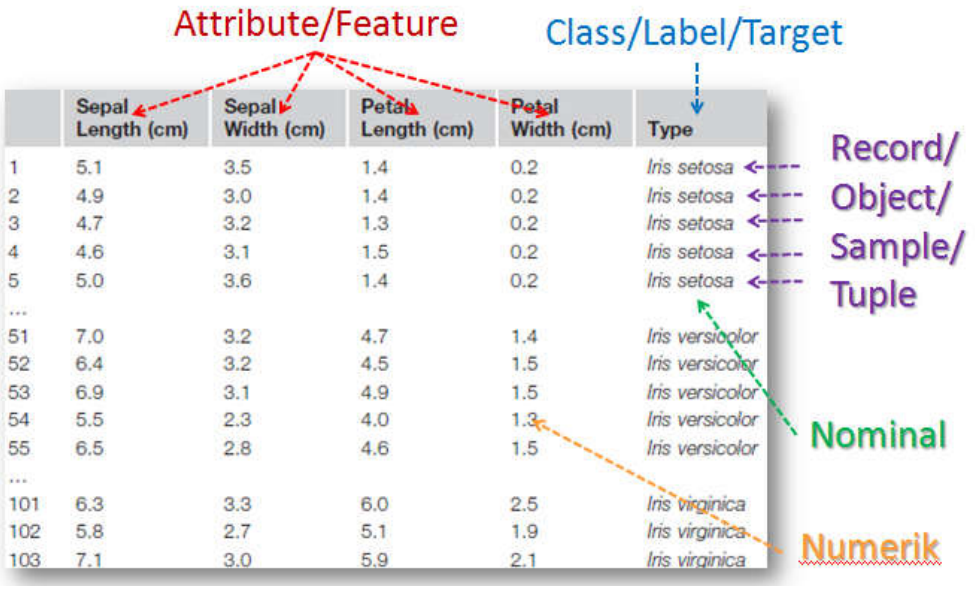
1. Private Dataset adalah data set dapat diambil dari organisasi yang kita jadikan obyek penelitian. Contohnya : Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc.
2. Public Datase adalah data set dapat diambil dari repositori pubik yang disepakati oleh para peneliti data mining. Contohnya :
   - UCI Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html)
   - ACM KDD Cup (http://www.sigkdd.org/kddcup/)
   - PredictionIO (http://docs.prediction.io/datacollection/sample/)

Trend penelitian data mining saat ini adalah menguji metode yang dikembangkan oleh peneliti dengan public dataset, sehingga penelitian dapat bersifat: comparable, repeatable dan verifiable

b. Pre-processing/ Cleaning (pemilihan data)

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada

dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal. Data terdiri dari bermacam-macam atribut harus Atribut adalah faktor atau parameter yang menyebabkan class/label/target terjadi.



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
...					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*

Gambar 1.3 Contoh Dataset (Himpunan Data)

c. Transformasi

Koding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

d. Data mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan. Metode datamining terdiri dari :

1. Estimation (Estimasi):
    - Linear Regression, Neural Network, Support Vector Machine, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
    - Linear Regression, Neural Network, Support Vector Machine, etc
3. Classification (Klasifikasi):
    - Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, Logistic Regression, etc
4. Clustering (Klastering):
    - K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc

5. Association (Asosiasi):
    - FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

e. Interpretasi / Evaluasi

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut dengan interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya. Metode-metode yang digunakan untuk melakukan estimasi adalah :

1. Estimation:
    - Error: Root Mean Square Error (RMSE), MSE, MAPE, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
    - Error: Root Mean Square Error (RMSE) , MSE, MAPE, etc
3. Classification:
    - Confusion Matrix: Accuracy
    - ROC Curve: Area Under Curve (AUC)
4. Clustering:
    - Internal Evaluation: Davies–Bouldin index, Dunn index,
    - External Evaluation: Rand measure, F-measure, Jaccard index, Fowlkes–Mallows index, Confusion matrix
5. Association:
    - Lift Charts: Lift Ratio
    - Precision and Recall (F-measure)

1.5 Arsitektur Data Mining

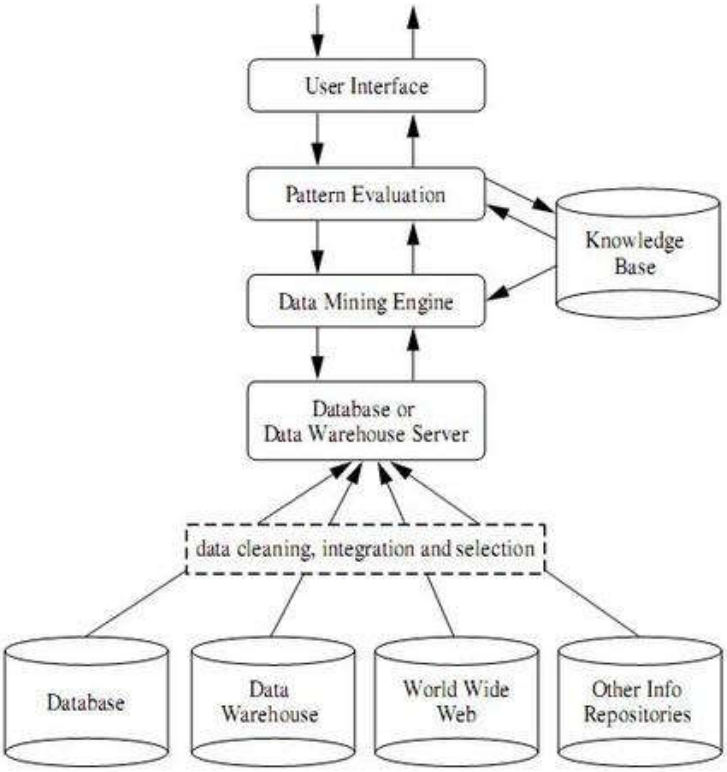
Data mining merupakan proses pencarian pengetahuan yang menarik dari data berukuran besar yang disimpan dalam basis data, data warehouse atau tempat penyimpanan informasi lainnya. Dengan demikian arsitektur sistem data mining memiliki komponen-komponen utama (Han dan Kamber, 2006) yaitu:

a. Database, data warehouse, World Wide Web, atau tempat penyimpanan informasi lainnya: bisa berbentuk satu atau banyak database, data warehouse, spreadsheet, ataupun tempat penyimpanan informasi lainnya. Data Cleaning, Data Integration dan Data Selection dapat dijalankan pada data tersebut.
b. Database dan data warehouse server. Komponen ini bertanggung jawab dalam pengambilan data yang relevan, berdasarkan permintaan pengguna.
c. Knowledge Based. Komponen ini merupakan domain knowledge yang digunakan untuk memandu pencarian atau mengevaluasi pola-pola yang

dihasilkan. Pengetahuan tersebut meliput hirarki konsep yang digunakan untuk mengorganisasikan atribut atau nilai atribut kedalam level abstraksi yang berbeda. Pengetahuan tersebut juga dapat berupa kepercayaan pengguna (user belief), yang dapat digunakan untuk menentukan kemenarikan pola yang diperoleh.

d. Data mining engine. Bagian ini merupakan komponen penting dalam arsitektur sistem data mining. Komponen ini terdiri dari modul-modul fungsional seperti karakterisasi, asosiasi, klasifikasi, dan analisis cluster.

e. Ghrapical user interface (GUI). Modul ini berkomunikasi dengan pengguna dan data mining. Melalui komponen ini, pengguna berinteraksi dengan sistem menggunakan query.



Gambar 1.4 Asitektur Data mining

1.6 Diskusi Pengantar Data mining

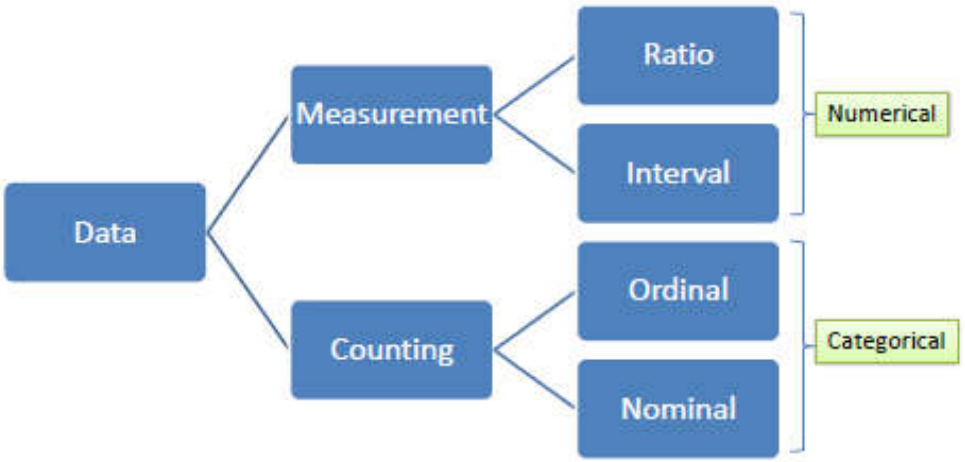
Diskusikas Dengan Kelompok Kalian :

1. Jelaskan dengan kalimat sendiri apa yang dimaksud dengan data mining?
2. Apa perbedaan basis data, data *warehouse* dan data mining?
3. Berikan contoh dengan mengambil kasus disekitarmu penerapan data mining?

BAB II
Data dan Karakterisitik Metode Datamining

2.1 Pengertian Data

Data adalah keterangan atau fakta mengenai suatu persoalan baik yang berbentuk ciri khas, kategori, atau sifat, maupun berbentuk bilangan atau angka-angka. Data mentah adalah data yang baru dikumpulkan dan belum pernah mengalami proses pengolahan apapun. Cara umum yang digunakan untuk mengklasifikasikan data adalah ditentukan oleh empat macam level pengukuruan atribut, yaitu level nominal, ordinal, interval dan rasio.



Gambar 2.1 Jenis Atribut Data

1. Data nominal atau data yang hanya dapat dibedakan.

Data nominal adalah data yang diberikan pada objek atau kategori yang tidak menggambarkan kedudukan objek atau kategori tersebut terhadap objek atau kategori lainnya, tetapi hanya sekedar label atau kode saja. Misalnya tentang jenis olah raga yakni tenis, basket dan renang. Kemudian masing-masing anggota set di atas kita berikan angka, misalnya tenis (1), basket (2) dan renang (3). Jelas kelihatan bahwa angka yang diberikan tidak menunjukkan bahwa tingkat olah raga basket lebih tinggi dari tenis ataupun tingkat renang lebih tinggi dari tenis. Angka tersebut tidak memberikan arti apa-apa jika ditambahkan. Angka yang diberikan hanya berfungsi sebagai label saja. Begitu juga tentang suku, yakni Dayak, Bugis dan Badui, tentang jenis kelamin yakni laki-laki dan perempuan, tentang agama yakni islam,kristen, katolik, hindu, budha, tentang negara dll.

2. Data Ordinal atau data yang dapat dibedakan dan diurutkan.

Data ordinal adalah data yang penomoran objek atau kategorinya disusun menurut besarnya, yaitu dari tingkat terendah ke tingkat tertinggi atau sebaliknya dengan jarak atau rentang yang tidak harus sama. Data ini memiliki ciri seperti ciri data nominal ditambah satu ciri lagi, yaitu kategori data dapat disusun atau diurutkan berdasarkan urutan logis dan sesuai dengan besarnya karakteristik yang dimiliki. Jika kita memiliki sebuah set objek yang dinomori, dari 1 sampai n, misalnya peringkat 1, 2, 3, 4, 5 dan seterusnya, bila dinyatakan dalam skala, maka jarak antara data yang satu dengan lainnya tidak sama. Ia akan memiliki urutan mulai dari yang paling tinggi sampai paling rendah. Atau paling baik sampai ke yang paling buruk (Nazir,2003).

Nilai akhir pada KHS mahasiswa merupakan konversi angka ke huruf :

- nilai A adalah dari 80-100
- nilai B adalah dari 65-79
- nilai C adalah dari 55-64
- nilai D adalah dari 45-54
- nilai E adalah dari 0-44

Nilai-nilai ini dapat diurutkan, misalnya nilai A lebih baik dari nilai B, tetapi seberapa besar selisih antara A dan B tidak dapat ditentukan. Jelasnya A-B tidak bermakna.

Misalnya dalam skala Likert mulai dari sangat setuju, setuju, ragu-ragu, tidak setuju sampai sangat tidak setuju. Atau jawaban pertanyaan tentang kecenderungan masyarakat untuk menghadiri rapat umum pemilihan kepala daerah, mulai dari tidak pernah absen menghadiri, dengan kode 5, kadang-kadang saja menghadiri, dengan kode 4, kurang menghadiri, dengan kode 3, tidak pernah menghadiri, dengan kode 2 sampai tidak ingin menghadiri sama sekali, dengan kode 1.

3. Data Interval atau data yang dapat dibedakan, diurutkan dan dapat dikuantitatifkan.

Data interval adalah data di mana objek/kategori dapat diurutkan berdasarkan suatu atribut yang memberikan informasi tentang interval antara tiap objek/kategori sama. Besarnya interval dapat ditambah atau dikurangi. Data ini memeiliki ciri sama dengan ciri pada data ordinal ditambah satu ciri lagi, yaitu urutan kategori data mempunyai jarak yang sama.

Contoh :

TEMPERATUR: suhu processor yang berjalan pada sebuah laptop 20 derajat celsius dan 40 derajat Celsius merupakan contoh data dalam level interval. Nilai-nilai ini dapat diurutkan dan selisihnya dapat ditentukan dengan jelas, dalam contoh ini selisihnya adalah 20 derajat celsius. Tetapi secara alami tidak ada titik

nol dimana suhu atau temperatur ini dimulai. Suhu 0 derajat tidak berarti tidak ada panas. Tidaklah benar mengatakan bahwa suhu processor merk X 40 derajat celsius panasnya 2 kali lipat dari suhu processor merk Y 20 derajat Celsius.

4. Data Rasio atau data yang dapat dibedakan, diurutkan, dapat dikuantitatifkan dan memiliki rasio bermakna.

Data rasio adalah data yang memiliki sifat-sifat data nominal, data ordinal, dan data interval, dilengkapi dengan kepemilikan nilai atau titik nol absolut/mutlak dengan makna empirik. Data rasio memiliki sifat; dapat dibedakan, diurutkan, punya jarak, dan punya nol mutlak.

Contoh:

a. HARGA: harga-harga RAM merupakan data level rasio dimana harga 0 rupiah menunjukkan tidak ada harga alias gratis.
b. BOBOT: berat laptop merupakan data level rasio dimana berat 0 kg menyatakan tidak ada bobot.

Ringkasan jenis atribut data dapat dilihat pada tabel 2.1 :

Tabel 2.1 Ringkasan Jenis Atribut Data

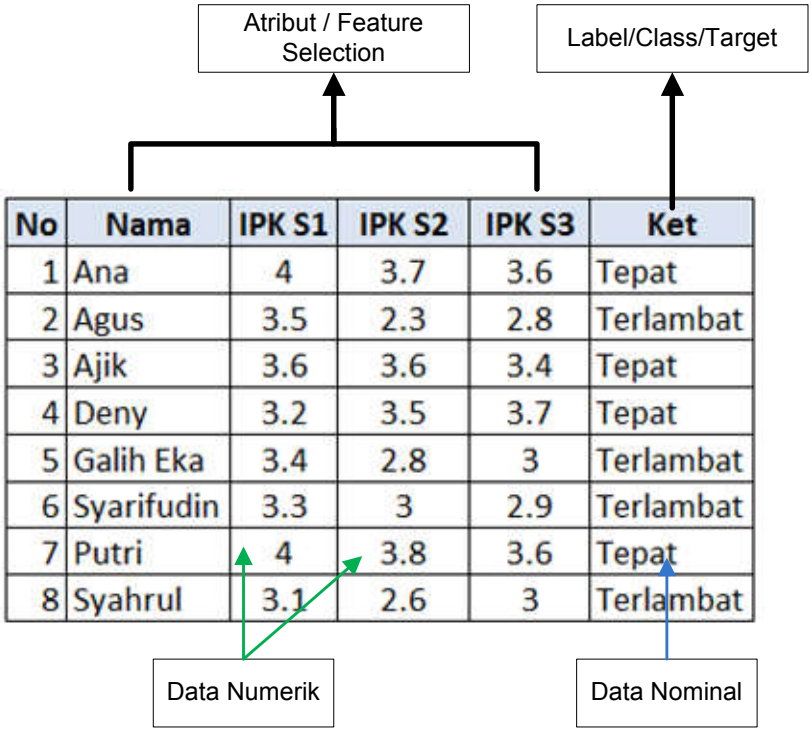
Jenis Atribut	Deskripsi	Contoh	Operasi
Ratio (Mutlak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Mempunyai titik nol yang absolut (*, /)	• Umur • Berat badan • Tinggi badan • Jumlah uang	geometric mean, harmonic mean, percent variation
Interval (Jarak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Tidak mempunyai titik nol yang absolut (+, -)	• Suhu 0°c-100°c, • Umur 20-30 tahun	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ordinal (Peringkat)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Tetapi diantara data tersebut terdapat hubungan atau berurutan (<, >)	• Tingkat kepuasan pelanggan (puas, sedang, tidak puas)	median, percentiles, rank correlation, run tests, sign tests
Nominal (Label)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Menunjukkan beberapa object yang berbeda (=, ≠) 33	• Kode pos • Jenis kelamin • Nomer id karyawan • Nama kota	mode, entropy, contingency correlation, χ^2 test

Pada Data Mining secara garis besar dikategorikan menjadi 2(dua) tipe yaitu:

1. Numeric merupakan tipe data yang bisa di kalkulasi (data rasio dan interval)

2. Nominal merupakan tipe data yang tidak bisa di kalkulasi baik tambah, kurang, kali maupun bagi.

Attribut adalah deskripsi data yang bisa mengidentifikasikan entitas Field adalah lokasi penyimpanan Record adalah kumpulan dari berbagai field yang saling berhubungan.
Class / Label / Target bisa disebut sebagai atribut keputusan.
Untuk contoh pemanfaatan tipe data dapat terlihat pada tabel di bawah ini:

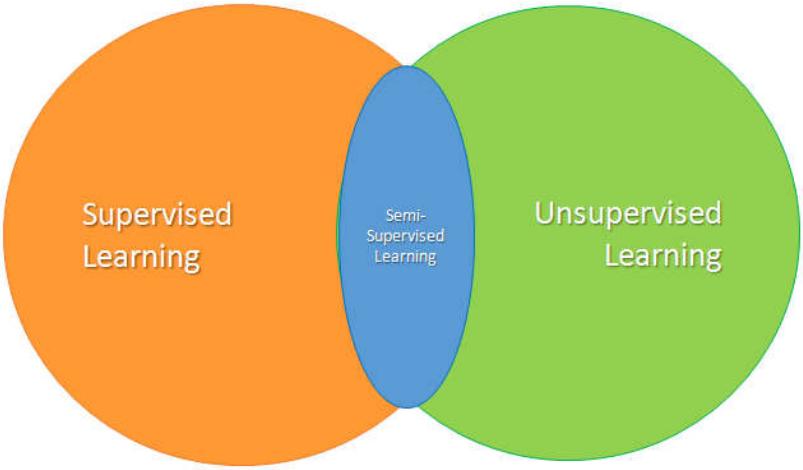


No	Nama	IPK S1	IPK S2	IPK S3	Ket
1	Ana	4	3.7	3.6	Tepat
2	Agus	3.5	2.3	2.8	Terlambat
3	Ajik	3.6	3.6	3.4	Tepat
4	Deny	3.2	3.5	3.7	Tepat
5	Galih Eka	3.4	2.8	3	Terlambat
6	Syarifudin	3.3	3	2.9	Terlambat
7	Putri	4	3.8	3.6	Tepat
8	Syahrul	3.1	2.6	3	Terlambat

Gambar 2.2 Contoh Dataset (Himpunan Data)

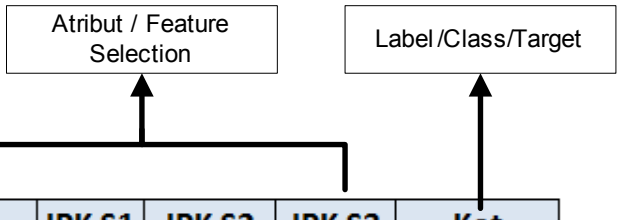
2.2 Metode Datamining

Data mining model dibuat berdasarkan salah satu dari dua jenis pembelajaran supervised dan unsupervised.



Gambar 2.3 Data mining model

Supervised learning disebut juga pembelajaran dengan guru. Cirinya adalah data set memiliki target/label/class. Algoritma melakukan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai dari variable predictor. Sebagian besar algoritma data mining (estimation, prediction/forecasting, classification) adalah supervised learning.

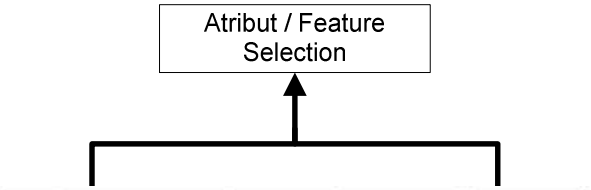


No	Nama	IPK S1	IPK S2	IPK S3	Ket
1	Ana	4	3.7	3.6	Tepat
2	Agus	3.5	2.3	2.8	Terlambat
3	Ajik	3.6	3.6	3.4	Tepat
4	Deny	3.2	3.5	3.7	Tepat
5	Galih Eka	3.4	2.8	3	Terlambat
6	Syarifudin	3.3	3	2.9	Terlambat
7	Putri	4	3.8	3.6	Tepat
8	Syahrul	3.1	2.6	3	Terlambat

Gambar 2.4 Contoh Dataset dengan Label/Class/Target

Unsupervised Learning adalah algoritma data mining mencari pola dari semua variable (atribut). Cirinya adalah variable (atribut) yang menjadi target/label/class

tidak ditentukan (tidak ada). Algoritma clustering adalah algoritma unsupervised learning.



No	Nama	IPK S1	IPK S2	IPK S3
1	Ana	4	3.7	3.6
2	Agus	3.5	2.3	2.8
3	Ajik	3.6	3.6	3.4
4	Deny	3.2	3.5	3.7
5	Galih Eka	3.4	2.8	3
6	Syarifudin	3.3	3	2.9
7	Putri	4	3.8	3.6
8	Syahrul	3.1	2.6	3

Gambar 2.5 Contoh Dataset tanpa Label/Class/Target

Semi-supervised learning adalah metode data mining yang menggunakan data dengan label dan tidak berlabel sekaligus dalam proses pembelajarannya. Data yang memiliki kelas digunakan untuk membentuk model (pengetahuan), data tanpa label digunakan untuk membuat batasan antara kelas.

Metode data mining dapat diklasifikasikan berdasarkan fungsi yang dilakukan atau berdasarkan jenis aplikasi yang menggunakannya:

1. Estimasi (supervised)
2. Prediksi (supervised)
3. Klasifikasi (supervised)
4. Association Rules (unsupervised)
5. Clustering (unsupervised)

2.3 Estimasi

Digunakan untuk melakukan estimasi terhadap sebuah data baru yang tidak memiliki keputusan berdasarkan histori data yang telah ada. Contohnya ketika melakukan Estimasi Pembiayaan pada saat pembangunan sebuah Hotel baru pada Kota yang berbeda.

2.4 Prediksi

Algoritma prediksi biasanya digunakan untuk memperkirakan atau forecasting suatu kejadian sebelum kejadian atau peristiwa tertentu terjadi. Contohnya pada bidang Klimatologi dan Geofisika, yaitu bagaimana Badan Meterologi Dan Geofisika (BMKG) memperkirakan tanggal tertentu bagaimana Cuacanya, apakah Hujan, Panas dan lain sebagainya.

2.5 Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan memprediksikan kelas untuk data yang tidak diketahui kelasnya. Dalam klasifikasi, terdapat target variable kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Dalam decision tree tidak menggunakan vector jarak untuk mengklasifikasikan obyek. Seringkali data observasi mempunyai atribut-atribut yang bernilai nominal. Misalkan obyeknya adalah sekumpulan buah-buahan yang bisa dibedakan berdasarkan atribut bentuk, warna, ukuran dan rasa. Bentuk, warna, ukuran dan rasa adalah besaran nominal, yaitu bersifat kategoris dan tiap nilai tidak bisa dijumlahkan atau dikurangkan. Dalam atribut warna ada beberapa nilai yang mungkin yaitu hijau, kuning, merah. Dalam atribut ukuran ada nilai besar, sedang dan kecil. Dengan nilai-nilai atribut ini, kemudian dibuat decision tree untuk menentukan suatu obyek termasuk jenis buah apa jika nilai tiap-tiap atribut diberikan. Contoh lainya adalah pada bidang Akademik yaitu Klasifikasi mahasiswa yang lulus tepat waktu atau yang terlambat lulus di kampus tertentu berdasarkan data kelulusan mahasiswa setiap tahunnya.

2.6 Asosiasi

Digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian. Adapun metode pemecahan masalah yang sering digunakan seperti Algoritma Apriori. Contoh pemanfaatan Algoritma Asosiasi yaitu pada Bidang Marketing ketika sebuah Minimarket melakukan Tata letak produk yang dijual berdasarkan Produkproduk mana yang paling sering dibeli konsumen, selain itu seperti tata letak buku yang dilakukan pustakawan di perpustakaan.

2.7 Klastering

Clustering atau Analisis *Custer* adalah proses pengelompokkan satu set benda-benda fisik atau abstrak kedalam kelas objek yang sama. Tujuan utama dari metode clustering adalah pengelompokan sejumlah data/obyek ke dalam cluster (group) sehingga dalam setiap cluster akan berisi data yang semirip mungkin. Dalam clustering metode ini berusaha untuk menempatkan obyek yang mirip (jaraknya dekat) dalam satu klaster dan membuat jarak antar klaster sejauh mungkin. Ini berarti obyek dalam satu cluster sangat mirip satu sama lain dan berbeda dengan obyek dalam *cluster-cluster* yang lain. Dalam metode ini tidak diketahui sebelumnya berapa jumlah cluster dan bagaimana pengelompokannya.

2.8 Diskusi Data dan Karakterisitik Metode Datamining

Diskusikas Dengan Kelompok Kalian :

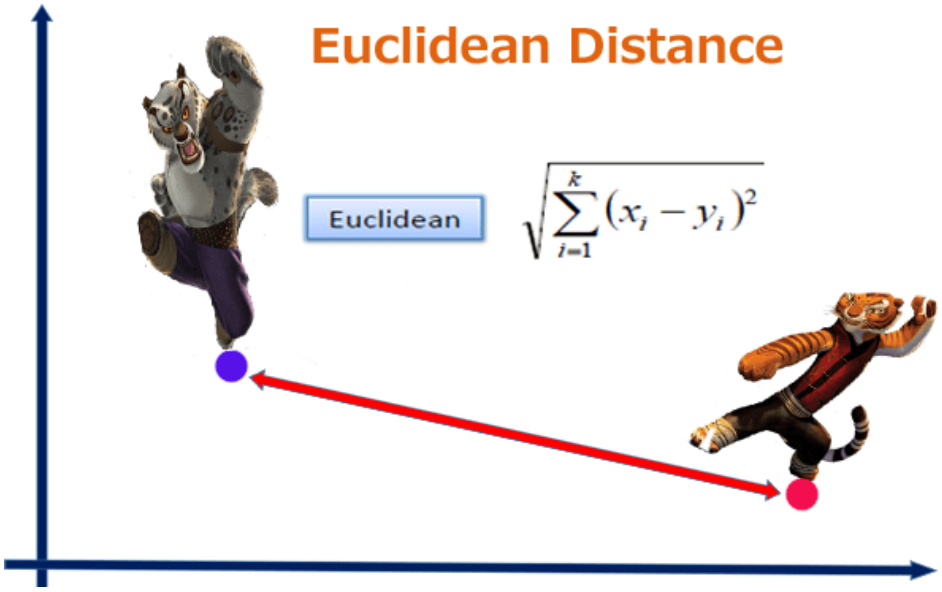
1. Sebutkan 5 peran utama data mining!
2. Jelaskan perbedaan estimasi dan prediksi!
3. Jelaskan perbedaan prediksi dan klasifikasi!
4. Jelaskan perbedaan klasifikasi dan klastering!
5. Jelaskan perbedaan klastering dan association!
6. Jelaskan perbedaan estimasi dan klasifikasi!
7. Jelaskan perbedaan estimasi dan klastering!
8. Jelaskan perbedaan supervised dan unsupervised learning!
9. Sebutkan tahapan utama proses data mining!

BAB III
Teori Pengukuran Jarak dan Similaritas

Jarak antara dua buah objek data bisa dihitung menggunakan dissimilarity (ketidakmiripan) atau similarity (kemiripan). Penggunaan ukuran jarak, dissimilarity atau similarity, bergantung pada teknik dan metode machine learning yang anda gunakan. Jarak antar dua objek data yang memiliki atribut numerik dapat dihitung menggunakan sejumlah formula di antanya adalah: Euclidean distance dan Manhattan distance, sedangkan untuk mengukur similaritas menggunakan Cosine Similarity.

3.1 Euclidean Distance

Euclidean distance adalah perhitungan jarak dari 2 buah titik dalam Euclidean space. Euclidean space diperkenalkan oleh Euclid, seorang matematikawan dari Yunani sekitar tahun 300 B.C.E. untuk mempelajari hubungan antara sudut dan jarak. Euclidean ini berkaitan dengan Teorema Phytagoras dan biasanya diterapkan pada 1, 2 dan 3 dimensi. Tapi juga sederhana jika diterapkan pada dimensi yang lebih tinggi. Eunclidean distance atau biasa disebut sebagai jarak garis lurus merupakan ukuran jarak untuk atribut numerik yang paling populer.



Gambar 3.1 Pengukuran jarak dengan Euclidean distance (sumber:dataaspirant.com)

Formula yang ini menggunakan rumus

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_i - y_i)^2}$$

atau

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Dimana ***x*** dan ***y*** adalah dua objek data yang memiliki ***n*** atribut bernilai numerik.

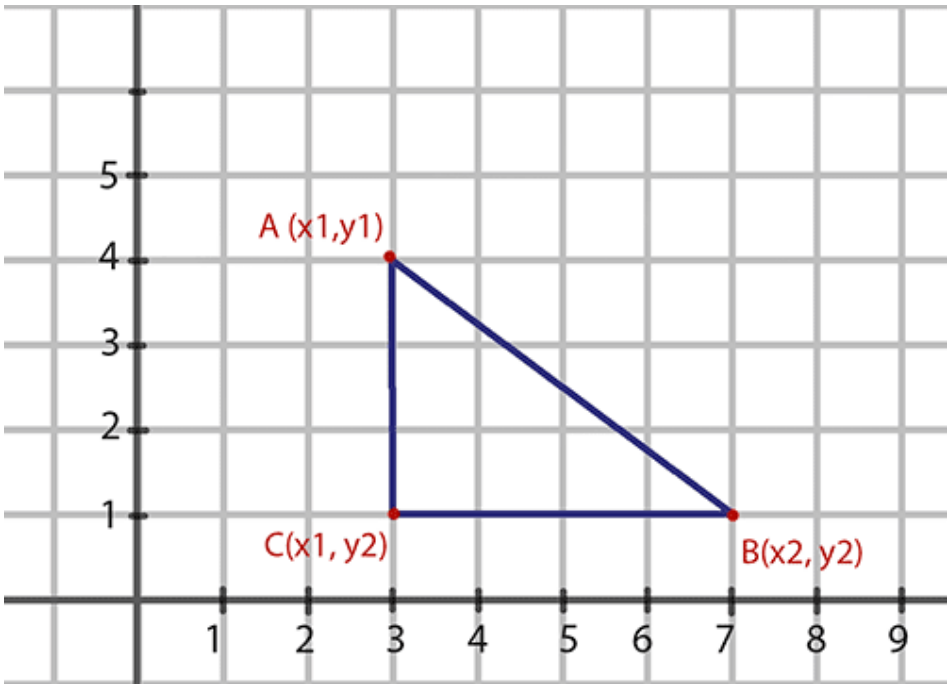
Contoh :

1. Diketahui dua buah titik (2,1) dan (3,2), hitunglah jarak menggunakan persamaan *Euclidean.*

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$= \sqrt[2]{(2-3)^2 + (1-2)^2} = \sqrt{2}$$

2. Diketahui tiga buah titik seperti gambar 3.2, hitunglah jarak titik A ke B menggunakan persamaan *Euclidean distance.*



Gambar 3.2 Tiga Buah Titik Berbentuk Segitiga Siku-Siku

Jawab : Titik A (3,4) dan Titik B (7, 1)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$
$$= \sqrt[2]{(3-7)^2 + (4-1)^2} = \sqrt{16+9} = \sqrt{25} = 5$$

3.2 *Square Euclidean Distance*

Sebagaimana namanya, jarak square Euclidean merupakan jarak dengan cara mengkuadratkan jarak antar dua titik yang akan diukur. Cara ini jarang sekali digunakan dalam pengukuran jarak. Formulasi dari jarak square Euclidean sebagai berikut:

$$d(x,y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_i - y_i)^2$$

atau

$$d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$

Dimana ***x*** dan ***y*** adalah dua objek data yang memiliki ***n*** atribut bernilai numerik.
Contoh :

1. Diketahui dua buah titik (2,1) dan (3,2), hitunglah jarak menggunakan persamaan *Square Euclidean.*

$$d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$
$$= (2-3)^2 + (1-2)^2 = 2$$

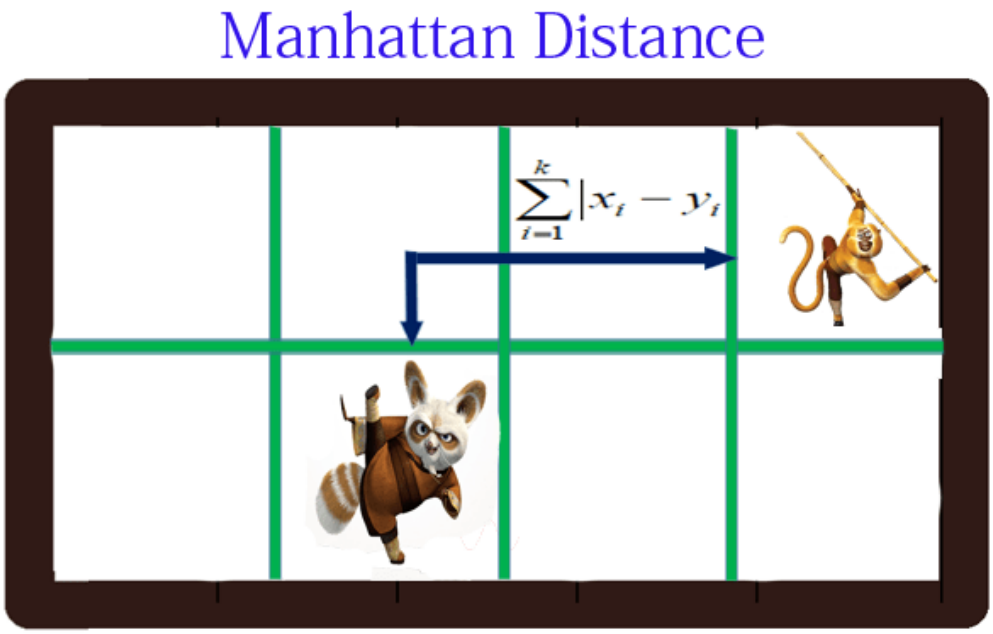
2. Diketahui tiga buah titik seperti gambar 3.2, hitunglah jarak titik A ke B menggunakan persamaan *Square Euclidean distance.*

Jawab : Titik A (3,4) dan Titik B (7, 1)

$$d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2$$
$$= (3-7)^2 + (4-1)^2 = 16 + 9 = 25$$

3.3 Manhattan distance

Jarak Rectilinear atau jarak manhattan adalah jarak yang diukur tegak lurus dari pusat fasilitas ke fasilitas yang lain. Disebut Manhattan ini berdasar pada kota Manhattan yang tersusun menjadi blok-blok. Sehingga sering juga disebut city block distance, juga sering disebut sebagai ablosute value distance atau boxcar distance. Sebagai ilustrasi, semisal kita berjalan dari lokasi A menuju utara 2 meter, kemudian belok ke timur 3 meter. Berapakah jarak kita yang sekarang dengan posisi titik A tadi. City Block distance adalah panjang jalan yang sudah kita tempuh dari B ke A sehingga menghasilkan total jarak = 2 + 3 = 5 blok.



Gambar 3.3 Pengukuran jarak dengan Manhattan distance (sumber:dataaspirant.com)

Manhattan distance dirumuskan sebagai

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_i - y_i|$$

atau

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

Dimana $|\boldsymbol{x_1} - \boldsymbol{y_1}|$ menyatakan selisih absolut antara nilai atribut ke-1 pada objek $\boldsymbol{x}$ dan nilai atribut ke-1 untuk objek $\boldsymbol{y}$. Misalnya, jika $\boldsymbol{x_1} = \boldsymbol{2}$ dan $\boldsymbol{y_1} = \boldsymbol{7}$, maka :

$$|x_1 - y_1| = |2 - 7| = 5.$$

Contoh :

1. Diketahui dua buah titik (2,1) dan (3,2), hitunglah jarak menggunakan persamaan *Square Euclidean.*

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

$$= |2 - 3| + |1 - 2| = 2$$

2. Diketahui tiga buah titik seperti gambar 3.2, hitunglah jarak titik A ke B menggunakan persamaan *Square Euclidean distance.*

Jawab : Titik A (3,4) dan Titik B (7, 1)

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

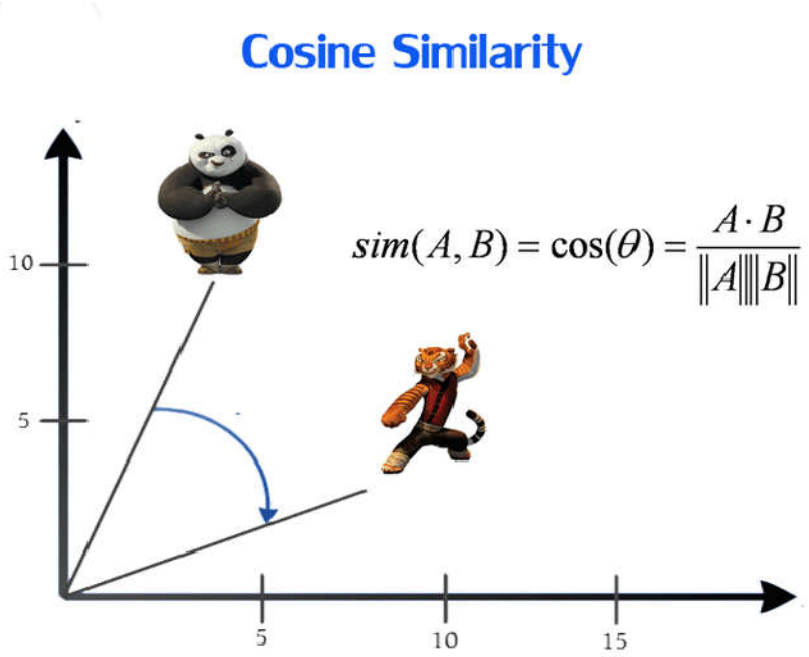
$$= |3 - 7| + |4 - 1| = 4 + 3 = 7$$

Eunclidean distance maupun Manhattan distance memiliki empat karakteritik berikut:

- Jarak antar objek tidak pernah bernilai negative, **d(I,j) ≥ 0**;
- Jarak suatu objek dengan dirinya 0, **d(i.i) = 0**;
- Simetris, **d(i,j) = d(j,i)**; dan
- Berlaku hukum pertidaksamaan segitiga; jarak langsung dari suatu titik ke titik kedua selalu lebih kecil dibanding jarak tidak langsung dengan melewati titik ketiga, **d(i,j) < d(i,k) + d(k,j)**.

3.4 Cosine Similarity

Ukuran jarak ini umumnya digunakan untuk data yang berupa vector dokumen, yang memiliki ribuan atribut (kata) yang frekuensinya kebanyakan bernilai 0. Anda bisa saja menggunakan dissimilarity untuk atribut numerik, misalnya Euclidean distance, tetapi anda akan memerlukan komputasi yang sangat kompleks hanya untuk menghitung ribuan atribut yang bernilai 0. Oleh karena itu, anda bisa menggunakan formula yang khusus di desain untuk menghitung jaraj antar vector dokumen yang disebut cosine similarity atau kemiripan kosinus.



Gambar 3.4 Pengukuran kemiripan dengan Cosine Similarity (sumber:dataaspirant.com)

Secara sistematis, cosine similarity diformulasikan sebagai
Rumus :

$$Cos(x.y) = \frac{x.y}{\|x\|\|y\|}$$

Atau

$$Cos(x.y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n y_i^2}}$$

adalah Euclidean norm (atau panjang) dari vector

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$$

$$\|y\| = \sqrt{y_1^2 + y_2^2 + \cdots + y_p^2}$$

Secara konsep, formula ini menyatakan kosinus sudut antara vector *x* dan vector ***y.*** Dengan demikian, formula ini menghasilkan nilai dalam rentang [0,1]. Jika vector ***x*** dan vector ***y*** tidak memiliki kemiripan sama sekali, maka kedua vector tersebut akan membentuk sudut 90 derajat (orthogonal atau tegak lurus) sehingga nilai kosinus sudutnya adalah 0, artinya ***cos(x.y)=0***. Sebaliknya, jika kedua vector

sama persis, maka keduanya akan membentuk sudut 0 derajat (berimpit) sehinggan ***cos****(x,y)=1*.

1. Contoh kasus yang kami ambil adalah tentang kemiripan 2 buah kalimat yaitu kalimat A dan B:
   A: Juli lebih suka aku dari pada Linda yang suka aku
   B: Jane lebih suka aku dari pada Juli yang suka aku
   Dalam perhitungan berikut digunakan metode consine similarity, pada aplikasi juga digunakan contoh kasus yang sama yaitu perhitungan kemiripan antara 2 buah kalimat.

   Dari kedua kalimat itu, maka didapat data sebagai berikut:
   1. Juli: pada kalimat A ada 1, kalimat B ada 1
   2. Lebih: pada kalimat A ada 1, kalimat B ada 1
   3. Suka: pada kalimat A ada 2, kalimat B ada 2
   4. Aku: pada kalimat A ada 2, kalimat B ada 2
   5. Dari: pada kalimat A ada 1, kalimat B ada 1
   6. Pada: pada kalimat A ada 1, kalimat B ada 1
   7. Linda: pada kalimat A ada 1, kalimat B ada 0
   8. Yang: pada kalimat A ada 1, kalimat B ada 1
   9. Jane: pada kalimat A ada 0, kalimat B ada 1

   Kemudian, untuk mencari tingkat kemiripannya yaitu:

$$\frac{(1*1)+(1*1)+(2*2)+(2*2)+(1*1)+(1*1)+(1*0)+(1*1)+(0*1)}{\sqrt{1^2+1^2+2^2+2^2+1^2+1^2+1^2+1^2+0^2}*\sqrt{1^2+1^2+2^2+2^2+1^2+1^2+0^2+1^2+1^2}} = 0.92857142857$$

   Dengan begitu, dapat disimpulkan bahwa tingkat kemiripan kalimat A dengan kalimat B adalah 0.92857142857.

2. Akan dihitung nilai similaritas dengan metode *cosine coefficient* antara kasus baru (P) dan pusat klaster (C) yang tunjukan pada Tabel 3.1.

Tabel 3.1 Contoh kasus baru dan nilai pusat klaster

	Usia	G01	G02	G03	G04	G05
Kasus Baru (P)	0.33	1	1	0	0	1
Pusat Klaster (C)	0.59	0.65	0.17	0.4	0.9	0.6

$$\langle P, C\rangle = (P_1 * C_1) + (P_2 * C_2) + (P_3 * C_3) + (P_4 * C_4) + (P_5 * C_5) + (P_6 * C_6)$$

$$\langle P, C\rangle = (0.33 * 0.59) + (1 * 0.65) + (1 * 0.17) + (0 * 0.4) + (0 * 0.9) + (1 * 0.6) = 1.6$$

$$||\mathrm{P}|| = \sqrt{P_1^2 + P_2^2 + P_3^2 + P_4^2 + P_5^2 + P_6^2}$$

$$||\mathrm{P}|| = \sqrt{0.33^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2} = 1.8$$

$$||\mathrm{C}|| = \sqrt{C_1^2 + C_2^2 + C_3^2 + C_4^2 + C_5^2 + C_6^2}$$

$$||\mathrm{C}|| = \sqrt{0.59^2 + 0.65^2 + 0.17^2 + 0.4^2 + 0.9^2 + 0.6^2} = 1.46$$

$$Sim = COS(P, C) = \frac{\langle P, C \rangle}{\|P\|\|C\|} = \frac{1.6}{1.8 * 1.46} = 0.63$$

$$Dis\ Sim = 1 - Sim = 1 - 0.63 = 0.37$$

3.5 Diskusi Teori Pengukuran Jarak dan Similaritas

Diskusikas Dengan Kelompok Kalian :

1. Bagaimana algoritma, contoh perhitungan dan implementasi dari algoritma :
   a. Jaccard similarity
   b. Mahalanobis Distance
   c. Chebyshev distance
   d. Minkowski distance
   e. Euclidean distance
   f. Manhattan distance
2. Hitung dengan metode Euclidean, Minkowski, Manhattan :

Atribut/Feature	X	Y	X
Objek A	2	3	4
Objek B	7	6	3

BAB IV
Regresi Linier Untuk Estimasi dan Prediksi

Dalam kondisi sehari-hari kita sering menjumpai adanya hubungan antara satu variabel dengan variabel lainnya. Sebagai contoh tingkat pendidikan seseorang berhubungan dengan gaji yang diperolehnya; dalam bidang pemasaran kita ketahui adanya hubungan antara volume penjualan dengan biaya advertensi dan lain-lain. Hubungan variabel diatas digambarkan adanya variabel bebas (X) dan tak bebas (Y). Hubungan antara dua atau lebih variabel tersebut ada dua macam, yaitu bentuk hubungan dan keeratan hubungan. Bila ingin diketahui bentuk hubungan, maka digunakan analisis regresi. Sedangkan bila yang ingin diketahui adalah keeratan hubungan, maka digunakan analisis korelasi. Analisis regresi adalah suatu proses melakukan estimasi untuk memperoleh suatu hubungan fungsional antara variabel acak Y dengan variabel X. Persamaan regresi digunakan untuk memprediksi nilai Y untuk nilai X tertentu. Analisis regresi sederhana adalah analisis regresi antara satu variabel Y dan satu variabel X. Pada materi ini kita hanya membahas persamaan regresi sederhana linier. Alat lain untuk mempelajari hubungan antara dua variabel adalah analisis Korelasi. Analisis ini meliputi pengukuran arah dan kekuatan suatu hubungan linier antara dua variabel. Arah dan kekuatan hubungan ini dinyatakan dalam koefisien korelasi.

4.1 Scatter Diagram

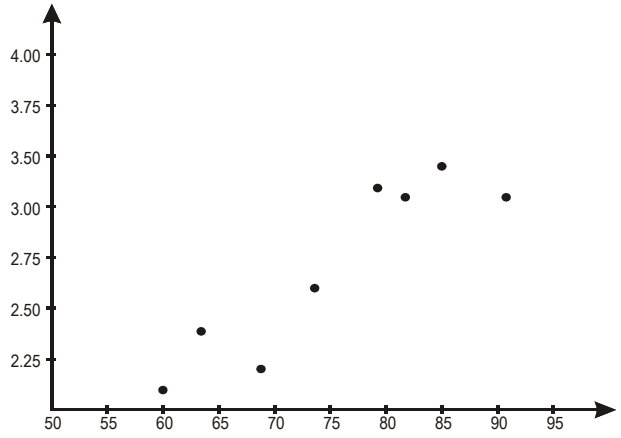
Bila dua variabel X dan Y berhubungan sebab akibat, dengan variabel X sebagai variabel independent (variabel bebas, variabel yang nilainya mempengaruhi nilai variabel tak bebas) dan variabel Y sebagai variabel dependent (variabel tak bebas, variabel yang nilainya dipengaruhi oleh variabel bebas), maka bila nilai variabel X diketahui, nilai tersebut dapat dipergunakan untuk memperkirakan nilai variabel Y jika bentuk hubungan kedua variabel tersebut diketahui. Untuk mengethaui pola hubungan yang mungkin terbentuk dari dua variabel X dan Y dapat dipergunakan Scatter diagram (diagram pencar).

Scatter diagram adalah grafik yang menunjukkan titik-titik perpaduan nilai observasi dari 2 variabel (X & Y). Pada umumnya dalam grafik, variabel independent (X) diletakkan pada garis horisontal, sedangkan variabel dependent (Y) pada garis vertikal. Dari scatter diagram dapat diperoleh informasi tentang bentuk hubungan antara dua variabel X dan Y dengan melihat macam pola yang terbentuk. Selain memberikan informasi tentang bentuk hubungan dari kedua variabel, polayang terbentuk juga dapat menggambarkan keeratan hubungan dari kedua variabel tersebut.

Tabel 4.1 Nilai Test Masuk dan IP Mahasiswa

Mahasiswa	A	B	C	D	E	F	G	H
Nilai test masuk (X)	74	69	85	63	82	60	79	91
IP (Y)	2.6	2.2	3.4	2.3	3.1	2.1	3.2	3.1

Dari informasi tersebut jika nilai test masuk digunakan untuk memprediksikan keberhasilan studi mahasiswa, maka test masuk merupakan variabel independent, sedangkan IP sebagai variabel dependent. Bila dibuat plot atas pasangan nilai diatas, akan diperoleh scatter diagram berikut :



Gambar 4.1 Scatter diagram Nilai Test Masuk dan IP Mahasiswa

Dari scatter diagram yang terbentuk dapat diberikan beberapa penjelasan sebagai berikut :

1. Hubungan kedua variabel tersebut adalah positif karena peningkatan nilai X juga diikuti peningkatan nilai Y (searah)
2. Derajat atau tingkat hubungan kedua variabel X dan Y sangat erat (titik-titik yang menunjukkan pertemuan nilai X dan Y mendekati garis lurus)
3. Hubungan kedua variabel adalah linier, karena titik-titik yang menunjukkan pertemuan nilai X dan Y tersebut dapat menggambarkan garis lurus.

Berdasarkan pola hubungan natara X dan Y yang diperoleh dari scatter diagram maka secara garis besar sifat hubungan antara variabel independent (X) dan variabel dependent (Y) dapat diklasifikasikan sebagai hubungan linier dan hubungan nonlinier. Sifat hubungan yang nonlinier (curvalinier) justru banyaj terjadi dalam masalah ekonomi, meskipun demikian pembahasan pada bab ini dibatasi hanya untuk hubungan yang linier.

Persamaan Regresi Linier

Y' = a + bX
Y' = nilai Y prediksi
Y = Variabel terikat
a = nilai rata-rata Y prediksi jika X = 0
b = rata-rata perubahan pada Y jika X berubah 1 satuan
X = Variabel bebas

Untuk menghitung koefisien a dan b pada persamaan diatas digunakan rumus :

$$a = \frac{\sum Y - b\sum X}{n} \qquad b = \frac{n\sum XY - \sum X \sum Y}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

Contoh :
Berikut data hasil test karyawan dengan unit penjualan perminggu :

Tabel 4.2 Hasil Test Karyawan dengan Unit Penjualan Perminggu

Salesman	Hasil Test (X)	Penjualan (Y)
A	4	5
B	7	12
C	3	4
D	6	8
E	10	11

Tentukan persamaan regresi linier sederhana data diatas
a. Hitunglah nilai penjualan, apabila salesman memiliki hasil test sebesar 8
Jawab

X	Y	X^2	XY	Y^2
4	5	16	20	25
7	12	49	84	144
3	4	9	12	16
6	8	36	48	64
10	11	100	110	121
30	40	210	274	370

a. Y' = a + b X

$$b = \frac{n\sum XY - \sum X \sum Y}{n\left(\sum X^2\right) - \left(\sum X\right)^2} = \frac{5(274) - (30)(40)}{5(210) - (30)^2} = \frac{1370 - 1200}{1050 - 900} = 1.133$$

$$a = \frac{\sum Y - b\sum X}{n} = \frac{40 - (1.133)(30)}{5} = 1.202$$

$\therefore$ Y = 1.202 + 1.133

b. Jika X = 8 Y = 1.202 + 1.133 (8)
= 1.202 + 9.1
= 10.302 ≈ 10

Contoh :

Seorang pengusaha usaha transportasi ingin mengetahui hubungan antara umur kendaraan dengan biaya perawatannya. Setelah dilakukan pengamatan, diketahui hubungan antara umur kendaraan dengan biaya perawatan sebagai berikut :

Tabel 4.3 Hasil Biaya Perawatan Kendaran

Nomor Kendaraan	Umur Kendaraan (tahun) (X)	Biaya Reparasi (Juta) (Y)
H 101 CC	5	3.1
H 104 CC	11	4
H 207 CC	4	3
H 532 CC	5	3.4
H 227 CC	3	2.5
H 438 CC	2	2

Berdasarkan informasi tersebut diatas dapat dilakukan, estimasi garis regresi berdasarkan metode kuadrat terkecil sebagai berikut :

No.	**Umur (X)**	**Biaya Perawatan (Y)**	**XY**	**X^2**	**Y^2**
1	5	3.1	15.5	25	9.61
2	11	4	44.0	121	16
3	4	3	12.0	16	9
4	5	3.4	17.0	25	11.56
5	3	2.5	7.5	9	6.25
6	2	2	4	4	4
	ΣX = 30	ΣY = 18	ΣXY = 100	$\Sigma X^2 = 200$	$\Sigma Y^2 =$ 56.42

Y = a + bX

$$b = \frac{n\sum XY - \sum X \sum Y}{n\left(\sum X^2\right) - \left(\sum X\right)^2} = \frac{6(100) - (30)(18)}{6(200) - (30)^2} = \frac{600 - 540}{1200 - 900} = \frac{60}{300} = 0.2$$

$$a = \frac{\sum Y - b\sum X}{n} = \frac{18 - (0.2)(30)}{6} = \frac{18 - 6}{6} = 2$$

$\therefore$ Y = 2 + 0.2X

Jika X = 8

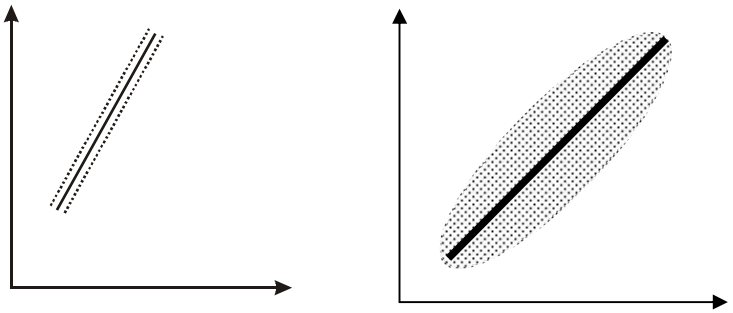
Y = 2 + 0.2 (8) = 3.6

4.2 Standard Error Estimasi

Proses selanjutnya dalam mempelajari analisis regresi adalah mengukur ketepatan persamaan estimasi. Ukuran keteparan persamaan-persamaan estimasi tersebut disebut standard error estimasi yang dilambangkan dengan Se.

Standard Error Estimasi adalah standard deviasi yang digunakan untuk mengukur penyebaran nilai observasi di sekitar garis regresi.

Standard error estimasi, mendekati sama dengan standard deviasi, keduanya merupakan ukuran penyebaran. Standard deviasi digunakan untuk mengukur penyebaran dari kumpulan nilai observasi dengan bertitik tolak pada mean, sedangkan standard error estimasi bertitik tolak pada garis pregresi.



(a)
Garis Regresi ini lebih tepat sebagai estimatr dari hubungan X dan Y

(b)
Garis Regresi ini kurang sebagai estimatr dari hubungan X dan Y

Formula dari Standard Error Estimasi (Se) adalah :

$$Se = \sqrt{\frac{\sum (Y - Y')^2}{n - 2}} \text{ atau}$$

$$Se = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}$$

Contoh :

Kembali pada pemilik usaha angkutan yang berupaya mengadakan prediksi terhadap biaya perawatan tiap mobil dengan melihat masa pakainya, telah ditemukan persamaan estimasi :

Y' = 2 + 0.2X

$$Se = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n-2}}$$

$$Se = \sqrt{\frac{56.42 - 2(18) - 0.2(100)}{6-2}} = \sqrt{\frac{56.42 - 36 - 20}{4}} = \sqrt{\frac{0.42}{4}} = 0.324$$

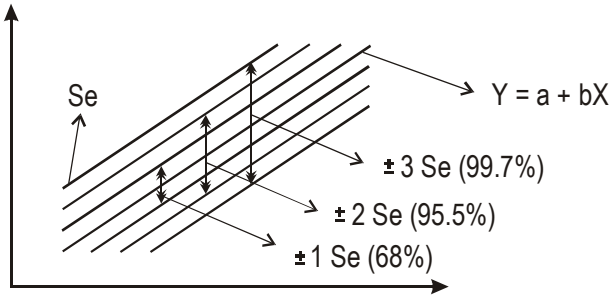
Sebenarnya standard error estimasi dapat diinterpretasikan seperti halnya standard deviasi terhadap nilai mean. Semakin besar nilai Se, semakin tersebar nilai observasi yang berada di sekitar garis regresi atau sebaliknya semakin kecil nilai Se, maka penyebaran nilai observasi akan mendekati garis regresi. Apabila Se = 0 berarti tidak ada penyebaran atau semua nilai observasi terletak pada garis regresi sehingga garis regresi yang terbentuk dapat digunakan secara sempurna untuk mengadakan prediksi nilai variabel dependent.

Dengan asumsi bahwa semua nilai observasi yang berada di sekitar garis regresi mengikuti distribusi normal maka :

68% Nilai observasi berada dalam jarak ± 1 Se

95% Nilai observasi berada dalam jarak ± 2 Se

99.7% Nilai observasi berada dalam jarak ± 3 Se

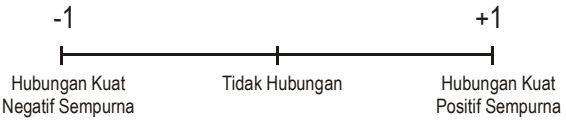


± 1 Se, ± 2 Se, dan ± 3 Se terletak disekitar garis regresi

4.3 Koefisien Korelasi Linier Sederhana

Bila analisis regresi berusaha memprediksi bentuk hubungan antara variabel Y dan X agar dapat memprediksi variabel Y untuk variabel X tertentu, analisis korelasi berusaha meghitung arah dan kekuatan hubungan antara variabel Y dan variabel X. Perbedaan utama regresi dengan korelasi adalah jika pada analisis regresi terdapat hubungan sebab akibat, pada analisis korelasi hubungan semacam ini tidak ada. Artinya korelasi antara Y dengan X akan sama dengan korelasi antara X dengan Y.

Kekuatan dan arah hubungan antara 2 variabel diukur dengan koefisien korelasi. Koefisien korelasi bertanda + (positif) atau – (negatif), dengan angka yang berkisar dari –1 hingga +1.



Semakin mendekati +1, koefisien korelasi menunjukkan adanya hubungan yang poasitif dan kuat. Koefisien korelasi yang mendekati –1 menunjukkan hubungan yang negatif dan kuat. Jika koefisien korelasi mendekati 0, memberikan indikasi bahwa ke 2 variabel tidak memiliki hubungan.

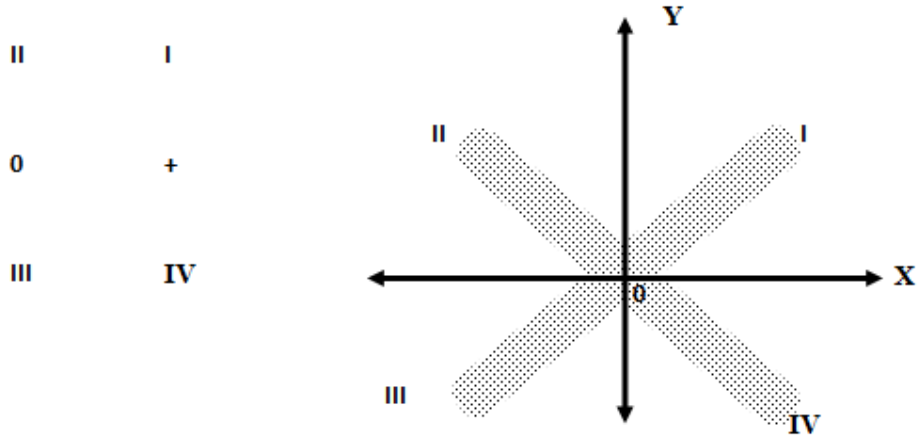
Untuk mencari koefisien korelasi linier sederhana digunakan rumus sebagai berikut :

$$r = \frac{n.\sum XY - \sum X.\sum Y}{\sqrt{n.\sum X^2 - (\sum X)^2}\sqrt{n.\sum Y^2 - (\sum Y)^2}}$$

dimana :

n = jumlah pasangan data
X = variabel bebas
Y = variabel terikat

Apa yang dimaksud dengan korelasi positif dan negative ? Jika 2 variabel berkorelasi positif, kenaikan variabel satu akan diikuti kenaikan variabel lain dan penurunan variabel satu diikuti dengan penurunan variabel lain. Sedangkan korelasi negatif menunjukkan jika satu variabel naik, variabel lain akan turun. Perhatikan gambar berikut :



Gambar 4.2 Hubungan Koefisien Korelasi Kuadaran

Kuandran I : Jika X naik, Y naik (korelasi positif)
Kuandran III : Jika X turun, Y turun (korelasi positif)
Kuandran II : Jika X turun, Y naik (korelasi negatif)
Kuandran IV : Jika X naik, Y turun (korelasi negatif)

Contoh :

Mencari koefisien korelasi antara variabel penjualan dengan variabel hasil test.

Salinan	Hasil Test (X)	Penjualan (Y)	X^2	XY	Y^2
A	4	5	16	20	25
B	7	12	49	84	144
C	3	4	9	12	16
D	6	8	36	48	64
E	10	11	100	110	120
Σ	30	40	210	274	370

$$r = \frac{n.\sum XY - \sum X.\sum Y}{\sqrt{n.\sum X^2 - (\sum X)^2} - \sqrt{n.\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{5(274) - (30)(40)}{\sqrt{5(210) - (30)^2} - \sqrt{5.(370) - (40)^2}} = 0.87$$

artinya antara hasil test dengan penjualan memiliki hubungan yang positif dan cukup kuat.

4.4 Diskusi Data Regresi Linier untuk Estimasi dan Prediksi

Ada keyakinan bahwa antara penjualan dan keuntungan ada hubungannya. Berikut adalah nilai penjualan dan keuntungan dari 11 pedagang yang ada di Jl. Mawar Kota XYZ pada tahun 20xx.

Pedagang	Penjualan (juta)	Keuntungan (juta)
Hermanto	89,2	4,9
Usman	18,6	4,4
Marwan	18,2	1,3
Syamsul	71,7	8,0
Kadam	58,6	6,6
Herman	46,8	4,1
Iskandar	17,5	2,6
Hamid	11,9	1,7
Agus	19,6	3,5
Bahrul Alam	51,2	8,2
Suyanto	28,6	6,1

Diminta :

a. Tentukan persamaan regresi linier pengaruh penjualan terhadap keuntungan! Dan tentukan nilai prediksi keuntungan jika penjualan sebesar 45
b. Tentukan nilai koefisien korelasinya. Dan berikan interpretasinya?

BAB V
Klasifikasi Dengan Algoritma K-Nearest Neighbor

5.1 Pendahuluan

Tujuan dari algoritma klasifikasi adalah untuk memprediksi kelas baru dari data set yang mempunyai kelas (J. a. Sáez, Galar, Leungo, & Herrera, 2013). Algoritma k-Nearest Naighbor (k-NN) masuk dalam algoritma klasifikasi (Haixiang, Yijing, Yanan, Xiao, & Jinling, 2015). Berikut adalah ilustrasi untuk memahami konsep algoritma k-NN.

Sebagai contoh, terdapat objek dengan kelas singa dan kambing. Singa mempunyai ciri-ciri, yaitu tidak bertanduk, gigi bertaring, pemakan daging, suara mengaum, kaki berjumlah empat, dan mempunyai ekor. Sedangkan kambing memiliki ciri-ciri, yaitu bertanduk, gigi tidak bertaring, pemakan tumbuhan, suara mengembik, kaki berjumlah empat, dan mempunyai ekor. Kemudian terdapat objek "X" yang mempunyai ciri-ciri, yaitu tidak bertanduk, gigi bertaring, pemakan tumbuhan, suara mengaum, jumlah kaki tiga, dan memiliki ekor. Berikut adalah tabel kemiripan antar objek "X", singa dan kambing.

Tabel 5.1 Kemiripan Objek "X" dengan Singa dan Kambing

Ciri-Ciri Ojek "X"	Singa	Kambing
Tidak Bertanduk	Mirip	Tidak Mirip
Gigi Bertaring	Mirip	Tidak Mirip
Pemakan Tumbuhan	Tidak Mirip	Mirip
Suara Mengaum	Mirip	Tidak Mirip
Kaki Berjumlah Tiga	Tidak Mirip	Mirip
Mempunyai Ekor	Mirip	Tidak Mirip
Jumlah Kemiripan	**4**	**2**

Dengan melihat *Similarity* (kemiripan) pada tabel 1, antara objek "X" dengan objek yang sudah diketahui kelasnya (singa dan kambing), dapat disimpulkan bahwa objek "X" tersebut masuk ke dalam kelas singa karena nilai *similarity* pada kelas singa (4 kemiripan) lebih besar daripada nilai *similarity* pada kelas kambing (2 kemiripan).

Nilai *similarity* pada algoritma k-NN dihitung berdasarkan jarak (distance) antara data training dan data testing. Metode *Euclidean distance* merupakan perhitungan jarak pada algoritma k-NN yang paling banyak digunakan oleh para

peneliti (Liu & Zhang, 2012). Menurut Harrington (Harrington, 2012), algoritma k-NN banyak digunakan peneliti karena mempunyai kelebihan, antara lain nilai akurasi tinggi, *insentive* terhadap *outlier*, dan tidak ada asumsi terhadap data. Namun algoritma k-NN juga mempunyai kelemahan, antara lain perlu untuk menentukan nilai k optimal , komputasi yang mahal, dan membutuhkan banyak memori.

Rumus Euclidean Distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Dimana,

- d(x,y) adalah jarak antara data x ke data y
- x_i adalah data testing ke-i
- y_i adalah data training ke-i

5.2 Perhitungan Manual Algortima k-Nearest Neighbor

Dataset dibagi menjadi dua bagian, yaitu data training dan data testing. Data training adalah data yang sudah mempunyai kelas, sedangkan data testing adalah data yang akan dicari kelasnya. Data training akan membentuk model/pola/pengetahuan, sedangkan data testing digunakan untuk mengukur evaluasi algoritma. Dataset yang digunakan pada kasus ini data dataset palsu/dummy yang dapat dilihat pada tabel dibawah. Kita akan menerapkan algoritma k-NN untuk mengetahui nilai prediksi (klasifikasi) kelas pada data testing, berdasarkan model pada data training.

Tabel 5.2 Data Training untuk K-NN

X	Y	Kelas
1,2	2,3	A
2,6	4,6	A
4	1	B
5,6	1,2	B
6	3,5	B

Tabel 5.3 Data Testing untuk K-NN

X	Y	Kelas
3	2	?

Berikut adalah langkah-langkah perhitungan manual algoritma k-NN:

1. Tentukan nilai parameter k

   Nilai parameter pada kasus ini adalah 3 (tiga). Jadi, kita akan menyimpulkan data testing yang akan kita cari kelasnya tersebut berdasarkan dengan 3 (tiga) data training terdekat dengan data testing (lihat gambar 1).

   Gambar 1. Algoritma k-NN dengan nilai k = 3

2. Hitung jarak antara data training dan data testing

   Perhitungan jarak yang digunakan untuk mengatur jarak antara data training dan data testing adalah *Euclidean distance* dengan rumus yang sudah ada diatas.

   - $Ed_1 = \sqrt[2]{(1{,}2-3)^2+(2{,}3-2)^2} = 1{,}825$
   - $Ed_2 = \sqrt[2]{(2{,}5-3)^2+(4{,}6-2)^2} = 2{,}648$
   - $Ed_3 = \sqrt[2]{(4-3)^2+(1-2)^2} = 1{,}414$
   - $Ed_4 = \sqrt[2]{(5{,}6-3)^2+(1{,}2-2)^2} = 2{,}721$
   - $Ed_5 = \sqrt[2]{(6-3)^2+(3{,}5-2)^2} = 3{,}354$

3. Urutkan data berdasarkan jarak terkecil

   Data diurutkan secara *ascending* (naik) berdasarkan jarak terkecil data training ke data testing.

Tabel 5.4 Jarak Data Training ke Data Testing untuk K-NN

Data Training		Kelas	Data Testing		Jarak Data Training ke Data Testing
X	Y		X	Y	
1,2	2,3	A	3	2	1,825
2,5	4,6	A	3	2	2,648
4	1	B	3	2	1,414
5,6	1,2	B	3	2	2,721
6	3,5	B	3	2	3,354

Tabel 5.5 Urutan Data berdasarkan jarak Terkecil untuk K-NN

Data Training		Kelas	Data Testing		Jarak Data Training ke Data Testing
X	Y		X	Y	
4	1	B	3	2	1,414
1,2	2,3	A	3	2	1,825
2,5	4,6	A	3	2	2,648
5,6	1,2	B	3	2	2,721
6	3,5	B	3	2	3,354

4. Menetapkan kelas

Setelah data training diurutkan berdasarkan jarak terkecil, langkah selanjutnya adalah menetapkan kelas. Dari tabel 5, dipilih tiga data terdekat (karena nilai k=3) antara data training dan data testing.

Tabel 5.6 Tiga Data dengan Jarak Terdekat antara Data Training dan Data Testing Untuk KNN

Data Training		Kelas	Data Testing		Jarak Data Training ke Data Testing
X	Y		X	Y	
4	1	B	3	2	1,414
1,2	2,3	A	3	2	1,825
2,5	4,6	A	3	2	2,648

Pada tabel diatas dapat dilihat data dengan kelas A lebih banyak dari pada kelas B dengan proporsi kelas A sebesar 67%, sedangkan kelas B sebesar 33%. Dapat disimpulkan bahwa data testing dengan nilai X = 3 dan Y = 2 masuk kedalam kelas A.

5.3 Implementasi Algoritma k-Nearest Neighor

Setelah memahami konsep perhitungan manual algoritma k-NN, kita implementasikan algoritma k-NN menggunakan Bahasa pemrograman PHP. Dataset yang kita gunakan dalam implementasi ini adalah data training dan data testing yang sudah kita gunakan pada perhitungan manual algoritma k-NN. Berikut adalah langkah-langkah implementasi algoritma k-NN menggunakan pemrograman PHP.

1. Buat file dengan nama knn.php.
2. Simpan file tersebut ke dalam folder …\xamm\htdocs\belajar-phpdm\public.
3. Ketikkan source code knn.php, seperti terlihat pada source code dibawah.

BAB VI
Klasifikasi Dengan Naïve Bayes

6.1 Pendahuluan

Naïve Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas(Patil and Sherekar 2013).
Definisi lain mengatakan Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami 2013).

Naïve Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu (Ridwan 2013). Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan (Pattekari and Parveen 2012)

Naive Bayes Classifier dinilai bekerja sangat baik dibanding dengan model classifier lainnya, yaitu Naïve Bayes Classifier memiliki tingkat akurasi yg lebih baik dibanding model classifier lainnya(Xhemali 2009).

Pre-requisites

- Pemahaman terhadap dasar-dasar Statistika terutama mengenai Probabilitas/Peluang
- Pemahaman terhadap dasar-dasar teknologi Web,HTML dan CSS
- Pemahaman terhadap dasar-dasar basis data/database, terutama query SQL pada MySQL/mariaDB
- Pemahaman terhadap dasar-dasar pemrograman PHP, terutama fungsi-fungsi koneksi database dan pengelolaan tipe data array

6.2 Teorema Naïve Bayes

Sebelum menjelaskan **Naïve Bayes Classifier** ini, akan dijelaskan terlebih dahulu **Teorema Bayes** yang menjadi dasar dari metoda tersebut. Pada **Teorema Bayes**, bila terdapat dua kejadian yang terpisah (misalkan X dan H), maka **Teorema Bayes** dirumuskan sebagai berikut (Bustami 2013).:

$$P(H|X) = \frac{P(X|H)}{P(X)}.P(H)$$ .. [NBC-01]

Keterangan

- **X** : Data dengan *class* yang belum diketahui
- **H** : Hipotesis data merupakan suatu *class* spesifik
- **P(H|X)** : Probabilitas hipotesis H berdasar kondisi X (*posteriori probabilitas*)
- **P(H)** : Probabilitas hipotesis H (*prior probabilitas*)
- **P(X|H)** : Probabilitas X berdasarkan kondisi pada hipotesis H
- **P(X)** : Probabilitas X

Teorema Bayes sering pula dikembangkan mengingat berlakunya hukum probabilitas total, menjadi seperti berikut:

$$P(H|X) = \frac{P(X|H)}{\sum_{i=1}^{n} P(H_i|X)}.P(H)$$ .. [NBC-02]

Keterangan

- **i** : 1,2,3, ... , n jumlah data Hipotesis (*prior probabilitas*)
- dimana : H_1 U H_2 U H_3 ... U H_n = S
- **S** : Probabilitas total H

Untuk menjelaskan **Teorema Naïve 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:

$$P(C|F_1, \dots, F_n) = \frac{P(F_1, \dots, F_n|C)}{P(F_1, \dots, F_n)}.P(C)$$ .. [NBC-03]

Di mana Variabel C merepresentasikan kelas, sementara variabel F_1 ... F_n merepresentasikan 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 di atas dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior\ x\ likehood}{evidence}$$ .. [NBC-04]

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 $(C, F_1, \ldots, F_n)$ menggunakan aturan perkalian sebagai berikut:

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

.. [NBC-05]

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 ($F_1, F_2, \ldots, F_n$) saling bebas (*independen*) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(F_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i).P(F_j)}{P(F_j)} = P(F_i)$$ .. [NBC-06]

Untuk $i \neq j$, sehingga

$$P(F_i|C,F_j) = P(F_i|C)$$ .. [NBC-07]

Persamaan di atas merupakan model dari teorema Naïve Bayes yang selanjutnya akan digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinyu digunakan rumus **Densitas Gauss** :

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

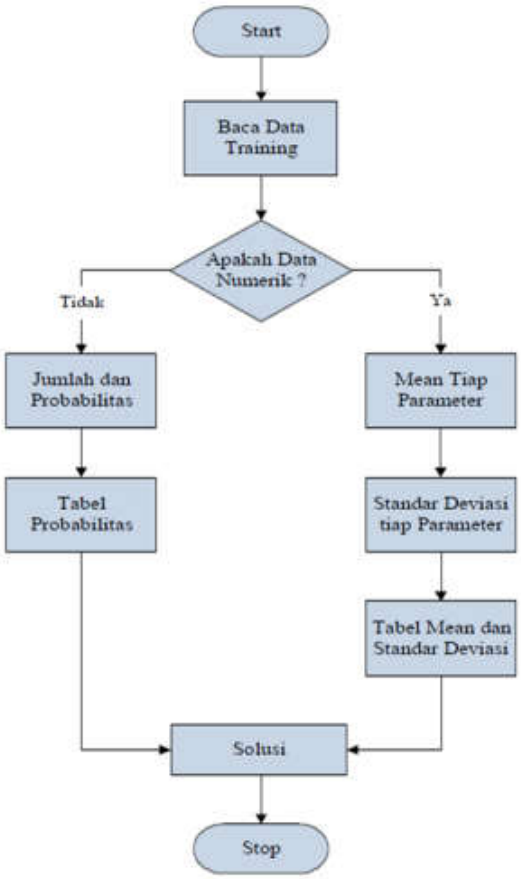
.. [NBC-08]

Keterangan

- **P** : Peluang
- **X_i** : Atribut ke-i
- **x_i** : Nilai Atribut ke-i
- **Y** : Kelas yang dicari
- **y_i** : Sub-kelas yang dicari
- **μ** : *mean*, menyatakan rata-rata dari seluruh atribut
- **σ** : Deviasi Standar, menyatakan varian dari seluruh atribut
- **e** : 2,7183

6.3 Alur Metode Naive Bayes

Alur dari metode Naive Bayes dapat dilihat pada Gambar 1 (Saleh 2015)sebagai berikut:



Gambar 6.1 Alur Metode Naïve Bayes

Adapun keterangan dari gambar di atas adalah sebagai berikut:

1. Membaca Data Training
2. Menghitung Jumlah dan Probabilitas, namun jika data numerik maka
   a. Menghitung nilai *mean* dan Dtandar Deviasi dari masing-masing parameter yang merupakan numerik. Adapun persamaan untuk mencari nilai rata-rata hitung (*mean*) adalah seperti dalam persamaan [NBC-09] berikut ini:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$ .. [NBC-09]

atau

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$ .. [NBC-10]

Keterangan

- **μ** : nilai rata-rata hitung (mean)
- **x_i** : nilai x ke-i
- **n** : jumlah sampel

Sedangkan persamaan untuk menghitung nilai Nilai Simpangan Baku (Standar Deviasi) dirumuskan sebagai berikut :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$ .. [NBC-11]

Keterangan

- **σ** :standar deviasi
- **x_i** : nilai x ke-i
- **μ** : nilai rata-rata hitung (mean)
- **n** : jumlah sampel

   b. Menghitung nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.
3. Mendapatkan nilai dalam tabel mean, Standar Deviasi dan Probabilitas
4. Menghasilkan Solusi

6.4 Kelebihan dan Kekurangan

Teorema Naïve Bayes memiliki beberapa kelebihan dan kekurangan yaitu sebagai berikut

Kelebihan
Teori Bayesian, mempunyai beberapa kelebihan (Grainner 1998), yaitu:

- Mudah untuk dipahami
- Hanya memerlukan pengkodean yang sederhana
- Lebih cepat dalam penghitungan
- Menangani kuantitatif dan data diskrit
- Kokoh untuk titik noise yang diisolasi, misalkan titik yang dirata – ratakan ketika mengestimasi peluang bersyarat data.
- Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata – rata dan variansi dari variabel) yang dibutuhkan untuk klasifikasi.
- Menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang
- Cepat dan efisiensi ruang
- Kokoh terhadap atribut yang tidak relevan

Kekurangan
Sedangkan kekurangan dari Teorema ini adalah :

- Kekurangan dari Teori probabilitas Bayesian yang banyak dikritisi oleh para ilmuwan adalah karena pada teori ini, satu probabilitas saja tidak bisa mengukur seberapa dalam tingkat keakuratannya. Dengan kata lain, kurang bukti untuk membuktikan kebenaran jawaban yang dihasilkan dari teori ini.
- Tidak berlaku jika probabilitas kondisionalnya adalah 0 (nol), apabila nol maka probabilitas prediksi akan bernilai nol juga
- Mengasumsikan variabel bebas

6.5 Laplace Correction

Laplace Correction (Laplacian Estimator) atau *additive smoothing* adalah suatu cara untuk menangani nilai probabilitas 0 (nol). Dari sekian banyak data di training set, pada setiap perhitungan datanya ditambah 1 (satu) dan tidak akan membuat perbedaan yang berarti pada estimasi probabilitas sehingga bisa menghindari kasus nilai probabilitas 0 (nol). *Laplace Correction* ini dapat dinyatakan dalam persamaan seperti pada persamaan [NBC-12] berikut ini

$$\rho_i = \frac{m_i + 1}{n + k}$$ .. [NBC-12]

Keterangan

- **ρ_i** : probabilitas dari atribut m_i
- **m_i** : jumlah sampel dalam kelas dari atribut m_i
- **k** : jumlah kelas dari atribut m_i
- **n** : jumlah sampel

Sebagai contoh, asumsikan ada *class* ***Perlengkapan*** = *sedang* di suatu training set, memiliki 17 sampel, ada 0 sampel dengan ***Penggunaan Listrik*** = *rendah*, 4 sampel dengan ***Penggunaan Listrik*** = *sedang*, dan 13 sampel dengan ***Penggunaan Listrik*** = *tinggi*.

Probabilitas dari kejadian ini tanpa *Laplacian Correction* adalah

P(Penggunaan Listrik = *rendah* | Perlengkapan = *sedang***)** = 0

P(Penggunaan Listrik = *sedang* | Perlengkapan = *sedang***)** = 0.235 (dari 4/17), dan

P(Penggunaan Listrik = *tinggi* | Perlengkapan = *sedang***)** = 0.764 (dari 13/17).

Menggunakan *Laplacian Correction* dari tiga kejadian diatas, diasumsikan ada 1 sampel lagi untuk masing – masing nilai ***Perlengkapan*** = *sedang*. Dengan cara ini, didapatkanlah probabilitas sebagai berikut (dibulatkan menjadi 3 angka dibelakang koma):

P(Penggunaan Listrik = rendah | Perlengkapan = sedang**)** = 0.050 (dari 1/20),

P(Penggunaan Listrik = sedang | Perlengkapan = sedang**)** = 0.250 (dari 5/20), dan

P(Penggunaan Listrik = tinggi | Perlengkapan = sedang**)** = 0.700 (dari 14/20)

Probabilitas yang "*dibenarkan*" hasilnya tidak berbeda jauh dengan hasil probabilitas sebelumnya sehingga nilai probabilitas 0 (nol) dapat dihindari.

Contoh Kasus :

Tabel 6.1 Data Training untuk Berolahraga

#	Cuaca	Temperatur	Kecepatan Angin	Berolah-raga
1	Cerah	Normal	Pelan	Ya
2	Cerah	Normal	Pelan	Ya
3	Hujan	Tinggi	Pelan	Tidak
4	Cerah	Normal	Kencang	Ya
5	Hujan	Tinggi	Kencang	Tidak
6	Cerah	Normal	Pelan	Ya

Apakah bila cuaca cerah, temperatur normal, kecepatan angin kencang, orang akan berolah raga?

Asumsi:

Y = berolahraga,

X_1 *= cuaca,*

X_2 *= temperatur,*

X_3 *= kecepatan angin.*

P(Y=ya)	4/6	P(Y=tidak)	2/6
P(X1=Cerah | Y=ya)	4/4	P(X1=Cerah | Y=tidak)	0/2
P(X2=Normal | Y=ya)	4/4	P(X2=Normal | Y=tidak)	0/2
P(X3=Kencang | Y=ya)	1/4	P(X3=Kencang | Y=tidak)	1/2

HMAP dari keadaan ini dapat dihitung dengan:

P(X1=cerah, X2=normal, X3=kencang | Y=ya)

= { P(X1=cerah|Y=ya) . P(X2=normal|Y=ya) . P(X3=kencang|Y=ya) } .

P(Y=ya)

= { (1) . (1/4) . 1} . (4/6) = 1/6

P(X1=cerah, X2=normal, X3=kencang | Y=tidak)

= { P(X1=cerah|Y=tidak).P(X2=normal|Y=tidak) .

P(X3=kencang|Y=tidak) }.P(Y=tidak)

= { (0) . (1/2) .0 } . (2/6) = 0

Studi Kasus dan Perhitungan

- Data yang digunakan BUKAN merupakan data *real*, tapi data yang di*generate* secara otomatis/random/acak dari sistem
- Data dan Nilai Perhitungan yang ditampilkan akan SELALU BERBEDA jika halaman di *refresh/reload*
- Jumlah Data Training ditampilkan secara acak/*random* antara 40 s.d 60
- Jika ditemukan probabilitas kondisional bernilai 0 (nol) maka otomatis diberlakukan *Laplace Coreection*

Penerapan **Metode Naïve Bayes** diharapkan mampu untuk memprediksi besarnya penggunaan listrik tiap rumah tangga agar lebih mudah mengatur penggunaan listrik.
Peranan listrik sangat penting bagi setiap lapisan masyarakat bahkan listrik juga sangat dibutuhkan sebagai sarana produksi dan untuk kehidupan sehari-hari, begitu pentingnya peranan listrik tentu saja berdampak pada permintaan listrik yang semakin besar tapi hal ini kiranya tidak linier dengan persediaan listrik yang belum mampu memenuhi permintaan listrik yang begitu besar tersebut. Untuk mengatasi hal ini perlu adanya campur tangan pemerintah dan masyarakat dalam menggunakan listrik dengan bijak sehingga kebutuhan listrik tidak menjadi lebih besar dari persediaan listrik. Oleh karena itu setiap rumah tangga haruslah paham penggunaan listrik yang efektif.

Pembacaan Data Training

Untuk menentukan data yang nantinya akan dianalisis dengan **Metode Naïve Bayes** maka langkah pertama yang dilakukan adalah membaca data training. Pada contoh kasus ini ada sejumlah **56** data training/sampel. Adapun data training yang digunakan dapat dilihat pada tabel 1 berikut(Nasari 2014):

Tabel 6.2 Data Training untuk Penggunaan Listrik

No	Jmlh. Tanggungan Keluarga	Luas Rumah	Pendapatan/bulan	Daya Listrik	Perlengkapan Yang Dimiliki	Penggunaan Listrik
1	banyak	besar	sedang	tinggi	banyak	tinggi
2	sedang	besar	besar	sedang	sedikit	tinggi
3	sedang	standar	besar	rendah	sedang	tinggi
4	sedikit	standar	besar	tinggi	banyak	tinggi
5	banyak	besar	sedang	sedang	banyak	tinggi
6	banyak	besar	besar	tinggi	banyak	tinggi
7	sedang	besar	besar	rendah	banyak	tinggi

8	banyak	standar	besar	tinggi	banyak	rendah
9	banyak	besar	besar	sedang	banyak	rendah
10	banyak	besar	besar	rendah	banyak	tinggi
11	banyak	standar	sedang	sedang	banyak	sedang
12	banyak	besar	besar	sedang	banyak	tinggi
13	sedang	besar	besar	tinggi	banyak	tinggi
14	sedang	standar	besar	tinggi	sedang	tinggi
15	banyak	standar	besar	sedang	sedang	sedang
16	sedang	besar	sedang	rendah	banyak	tinggi
17	sedang	besar	besar	tinggi	banyak	tinggi
18	sedikit	besar	sedang	tinggi	banyak	sedang
19	banyak	besar	besar	sedang	banyak	sedang
20	banyak	besar	kecil	sedang	sedang	tinggi
21	banyak	besar	besar	rendah	sedikit	sedang
22	banyak	besar	kecil	rendah	sedikit	tinggi
23	banyak	standar	sedang	tinggi	sedang	tinggi
24	banyak	besar	sedang	sedang	sedang	tinggi
25	banyak	standar	besar	tinggi	banyak	tinggi
26	sedikit	standar	besar	tinggi	banyak	tinggi
27	banyak	besar	sedang	tinggi	banyak	rendah
28	sedang	standar	besar	sedang	banyak	tinggi
29	sedikit	kecil	sedang	tinggi	sedang	tinggi
30	banyak	besar	kecil	sedang	banyak	tinggi
31	banyak	kecil	besar	tinggi	banyak	sedang
32	sedang	standar	besar	tinggi	sedang	tinggi
33	sedang	besar	sedang	sedang	banyak	tinggi
34	banyak	besar	sedang	tinggi	banyak	rendah
35	banyak	besar	besar	tinggi	banyak	tinggi
36	sedang	standar	besar	tinggi	sedang	tinggi
37	banyak	standar	besar	tinggi	banyak	tinggi
38	sedang	standar	besar	tinggi	banyak	sedang
39	banyak	besar	besar	tinggi	sedang	tinggi
40	banyak	besar	sedang	tinggi	sedang	tinggi
41	banyak	standar	besar	tinggi	sedang	tinggi
42	banyak	besar	besar	rendah	sedikit	tinggi
43	banyak	besar	sedang	tinggi	sedang	tinggi
44	banyak	besar	kecil	tinggi	sedang	sedang
45	banyak	besar	besar	tinggi	sedang	tinggi
46	sedang	besar	besar	sedang	banyak	tinggi

47	banyak	kecil	besar	sedang	banyak	tinggi
48	sedang	besar	sedang	sedang	banyak	sedang
49	sedang	standar	sedang	rendah	sedang	sedang
50	sedang	besar	besar	tinggi	sedikit	sedang
51	sedang	besar	sedang	rendah	banyak	tinggi
52	banyak	besar	sedang	tinggi	sedikit	sedang
53	banyak	besar	besar	tinggi	sedang	sedang
54	banyak	besar	besar	sedang	sedikit	sedang
55	sedang	kecil	sedang	tinggi	banyak	sedang
56	banyak	besar	sedang	tinggi	banyak	tinggi

Kriteria dan Probabilitas

Adapun nilai probabilitas setiap kriteria didapatkan dari data training pada tabel 1. Adapun nilai probabilitas setiap kriteria sebagai berikut:

Probabilitas Kriteria Jumlah Tanggungan

Pada kriteria jumlah tanggungan dapat diketahui dari **56** data terdapat : **tidak ada** data rumah tangga dengan jumlah tanggungan *sedikit*dan penggunaan listrik *rendah* , **1** data rumah tangga dengan jumlah tanggungan *sedikit* dan penggunaan listrik *sedang* , **3** data rumah tangga dengan jumlah tanggungan *sedikit* dan penggunaan listrik *tinggi* , **tidak ada** data rumah tangga dengan jumlah tanggungan *sedang* dan penggunaan listrik *rendah* , **5** data rumah tangga dengan jumlah tanggungan *sedang* dan penggunaan listrik *sedang* , **13** data rumah tangga dengan jumlah tanggungan *sedang* dan penggunaan listrik *tinggi* , **4** data rumah tangga dengan jumlah tanggungan *banyak* dan penggunaan listrik *rendah* , **9** data rumah tangga dengan jumlah tanggungan *banyak* dan penggunaan listrik *sedang* , **21** data rumah tangga dengan jumlah tanggungan *banyak* dan penggunaan listrik *tinggi*. Probabilitas kriteria jumlah tanggungan dapat dilihat pada tabel 2

Tabel 6.3 Probabilitas Kriteria Jumlah Tanggungan

	Jumlah Kejadian 'Penggunaan Listrik'			**Probabilitas**		
Jumlah Tanggungan	**Rendah**	**Sedang**	**Tinggi**	**Rendah**	**Sedang**	**Tinggi**
sedikit	0	1	3	0.142	0.111	0.1
sedang	0	5	13	0.142	0.333	0.35
banyak	4	9	21	1	0.6	0.567

Probabilitas Kriteria Luas Tanah

Pada kriteria luas tanah dapat diketahui dari **56** data terdapat : **tidak ada** data rumah tangga dengan luas tanah *kecil* dan penggunaan listrik *rendah* , **2** data rumah tangga dengan luas tanah *kecil* dan penggunaan listrik *sedang* , **2** data rumah tangga dengan luas tanah *kecil* dan penggunaan listrik *tinggi* , **1** data rumah tangga dengan luas tanah *standar* dan penggunaan listrik *rendah* , **4** data rumah tangga dengan luas tanah *standar* dan penggunaan listrik *sedang* , **11** data rumah tangga dengan luas tanah *standar* dan penggunaan listrik *tinggi* , **3** data rumah tangga dengan luas tanah *besar* dan penggunaan listrik *rendah* , **9** data rumah tangga dengan luas tanah *besar* dan penggunaan listrik *sedang* , **24** data rumah tangga dengan luas tanah *besar* dan penggunaan listrik *tinggi*. Probabilitas kriteria luas tanah dapat dilihat pada tabel 3

Tabel 6.4 Probabilitas Kriteria Luas Tanah

	Jumlah Kejadian 'Penggunaan Listrik'			**Probabilitas**		
Luas Tanah	**Rendah**	**Sedang**	**Tinggi**	**Rendah**	**Sedang**	**Tinggi**
kecil	0	2	2	0.142	0.166	0.075
standar	1	4	11	0.25	0.266	0.297
besar	3	9	24	0.75	0.6	0.648

Probabilitas Kriteria Pendapatan

Pada kriteria pendapatan dapat diketahui dari **56** data terdapat : **tidak ada** data rumah tangga dengan pendapatan *kecil* dan penggunaan listrik *rendah* , **1** data rumah tangga dengan pendapatan *kecil* dan penggunaan listrik *sedang* , **3** data rumah tangga dengan pendapatan *kecil* dan penggunaan listrik *tinggi* , **2** data rumah tangga dengan pendapatan *sedang* dan penggunaan listrik *rendah* , **6** data rumah tangga dengan pendapatan *sedang* dan penggunaan listrik *sedang* , **11** data rumah tangga dengan pendapatan *sedang* dan penggunaan listrik *tinggi* , **2** data rumah tangga dengan pendapatan *besar* dan penggunaan listrik *rendah* , **8** data rumah tangga dengan pendapatan *besar* dan penggunaan listrik *sedang* , **23** data rumah tangga dengan pendapatan *besar* dan penggunaan listrik *tinggi*. Probabilitas kriteria pendapatan dapat dilihat pada tabel 4.

Tabel 6.5 Probabilitas Kriteria Pendapatan

	Jumlah Kejadian 'Penggunaan Listrik'			Probabilitas		
Pendapatan	**Rendah**	**Sedang**	**Tinggi**	**Rendah**	**Sedang**	**Tinggi**
Kecil	0	1	3	0.142	0.111	0.1
sedang	2	6	11	0.5	0.4	0.297
Besar	2	8	23	0.5	0.533	0.621

Probabilitas Kriteria Daya Listrik

Pada kriteria daya listrik dapat diketahui dari **56** data terdapat : **tidak ada** data rumah tangga dengan daya listrik *rendah* dan penggunaan listrik *rendah* , **2** data rumah tangga dengan daya listrik *rendah* dan penggunaan listrik *sedang* , **7** data rumah tangga dengan daya listrik *rendah* dan penggunaan listrik *tinggi* , **1** data rumah tangga dengan daya listrik *sedang* dan penggunaan listrik *rendah* , **5** data rumah tangga dengan daya listrik *sedang* dan penggunaan listrik *sedang* , **10** data rumah tangga dengan daya listrik *sedang* dan penggunaan listrik *tinggi* , **3** data rumah tangga dengan daya listrik *tinggi* dan penggunaan listrik *rendah* , **8** data rumah tangga dengan daya listrik *tinggi*dan penggunaan listrik *sedang* , **20** data rumah tangga dengan daya listrik *tinggi* dan penggunaan listrik *tinggi*. Probabilitas kriteria daya listrik dapat dilihat pada tabel 5.

Tabel 6.6 Probabilitas Kriteria Daya Listrik

	Jumlah Kejadian 'Penggunaan Listrik'			Probabilitas		
Daya Listrik	**Rendah**	**Sedang**	**Tinggi**	**Rendah**	**Sedang**	**Tinggi**
Rendah	0	2	7	0.142	0.166	0.2
Sedang	1	5	10	0.25	0.333	0.27
Tinggi	3	8	20	0.75	0.533	0.54

Probabilitas Kriteria Perlengkapan

Pada kriteria perlengkapan dapat diketahui dari **56** data terdapat : **tidak ada** data rumah tangga dengan perlengkapan *sedikit* dan penggunaan listrik *rendah* , **4** data rumah tangga dengan perlengkapan *sedikit* dan penggunaan listrik *sedang* , **3** data rumah tangga dengan perlengkapan *sedikit* dan penggunaan listrik *tinggi* , **tidak ada** data rumah tangga dengan perlengkapan *sedang* dan penggunaan listrik *rendah* , **4** data rumah tangga dengan perlengkapan *sedang* dan penggunaan listrik *sedang* , **13** data rumah tangga dengan perlengkapan *sedang* dan penggunaan listrik *tinggi* , **4** data rumah tangga dengan

perlengkapan *banyak* dan penggunaan listrik *rendah* , **7**data rumah tangga dengan perlengkapan *banyak* dan penggunaan listrik *sedang* , **21** data rumah tangga dengan perlengkapan *banyak* dan penggunaan listrik *tinggi*. Probabilitas kriteria perlengkapan dapat dilihat pada tabel 6

Tabel 6.7 Probabilitas Kriteria Perlengkapan

	Jumlah Kejadian 'Penggunaan Listrik'			Probabilitas		
Perlengkapan	**Rendah**	**Sedang**	**Tinggi**	**Rendah**	**Sedang**	**Tinggi**
Sedikit	0	4	3	0.142	0.277	0.1
Sedang	0	4	13	0.142	0.277	0.35
Banyak	4	7	21	1	0.466	0.567

Probabilitas Kriteria Penggunaan Listrik

Berdasarkan tabel 1 dapat diketahui dari **56** data terdapat : **37** data rumah tangga dengan penggunaan listrik *tinggi* , **4** data rumah tangga dengan penggunaan listrik *rendah* , **15** data rumah tangga dengan penggunaan listrik *sedang*. Probabilitas kriteria penggunaan listrik dapat dilihat pada tabel 7

Tabel 6.8 Probabilitas Kriteria Penggunaan Listrik

Jumlah Kejadian 'Penggunaan Listrik'			Probabilitas		
Rendah	**Sedang**	**Tinggi**	**Rendah**	**Sedang**	**Tinggi**
37	4	15	0.66	0.071	0.267

Pengujian Metode Naive Bayes

Dari nilai probabilitas di atas akan diuji data sebanyak **56** data dan dihasilkan hasil klasifikasi penggunaan listrik seperti terlihat pada TABEL 8 berikut ini:

Berdasarkan TABEL 8 di atas dapat dilihat persentase untuk *Correctly Classified Instance* adalah sebesar **64.285 %** sementara persentase untuk *Incorrectly Classified Instance* adalah sebesar **35.714 %**. Di mana dari **56** data penggunaan listrik rumah tangga, ada sebanyak **36** data penggunaan listrik rumah tangga berhasil diklasifikasikan dengan benar dan sebanyak **20** data penggunaan listrik rumah tangga tidak berhasil diklasifikasikan dengan benar.

Tabel 6.9 Hasil Klasifikasi Penggunaan Listrik

No	Input Kategorikal					Class	Probabilitas			Prediction
	X_1	X_2	X_3	X_4	X_5		rendah	sedang	tinggi	
1	Banyak	besar	sedang	tinggi	banyak	tinggi	0.02009	0.00960	0.02219	tinggi
2	Sedang	besar	besar	sedang	sedikit	tinggi	0.00014	0.00265	0.00252	sedang
3	Sedang	standar	besar	rendah	sedang	tinggi	0.00003	0.00059	0.00299	tinggi
4	Sedikit	standar	besar	tinggi	banyak	tinggi	0.00096	0.00105	0.00375	tinggi
5	Banyak	besar	sedang	sedang	banyak	tinggi	0.00670	0.00600	0.01109	tinggi
6	Banyak	besar	besar	tinggi	banyak	tinggi	0.02009	0.01280	0.04639	tinggi
7	Sedang	besar	besar	rendah	banyak	tinggi	0.00055	0.00222	0.01058	tinggi
8	Banyak	standar	besar	tinggi	banyak	rendah	0.00670	0.00569	0.02126	tinggi
9	Banyak	besar	besar	sedang	banyak	rendah	0.00670	0.00800	0.02319	tinggi
10	Banyak	besar	besar	rendah	banyak	tinggi	0.00383	0.00400	0.01716	tinggi
11	Banyak	standar	sedang	sedang	banyak	sedang	0.00223	0.00267	0.00508	tinggi
12	Banyak	besar	besar	sedang	banyak	tinggi	0.00670	0.00800	0.02319	tinggi
13	Sedang	besar	besar	tinggi	banyak	tinggi	0.00287	0.00711	0.02861	tinggi
14	Sedang	standar	besar	tinggi	sedang	tinggi	0.00014	0.00188	0.00809	tinggi
15	Banyak	standar	besar	sedang	sedang	sedang	0.00032	0.00212	0.00656	tinggi
16	Sedang	besar	sedang	rendah	banyak	tinggi	0.00055	0.00167	0.00506	tinggi
17	Sedang	besar	besar	tinggi	banyak	tinggi	0.00287	0.00711	0.02861	tinggi
18	Sedikit	besar	sedang	tinggi	banyak	sedang	0.00287	0.00178	0.00391	tinggi
19	Banyak	besar	besar	sedang	banyak	sedang	0.00670	0.00800	0.02319	tinggi
20	Banyak	besar	kecil	sedang	sedang	tinggi	0.00027	0.00099	0.00230	tinggi
21	Banyak	besar	besar	rendah	sedikit	sedang	0.00055	0.00238	0.00302	tinggi
22	Banyak	besar	kecil	rendah	sedikit	tinggi	0.00016	0.00050	0.00049	sedang
23	Banyak	standar	sedang	tinggi	sedang	tinggi	0.00096	0.00254	0.00627	tinggi
24	Banyak	besar	sedang	sedang	sedang	tinggi	0.00096	0.00357	0.00684	tinggi
25	Banyak	standar	besar	tinggi	banyak	tinggi	0.00670	0.00569	0.02126	tinggi
26	Sedikit	standar	besar	tinggi	banyak	tinggi	0.00096	0.00105	0.00375	tinggi
27	Banyak	besar	sedang	tinggi	banyak	rendah	0.02009	0.00960	0.02219	tinggi
28	Sedang	standar	besar	sedang	banyak	tinggi	0.00032	0.00198	0.00656	tinggi
29	Sedikit	kecil	sedang	tinggi	sedang	tinggi	0.00008	0.00029	0.00028	sedang
30	Banyak	besar	kecil	sedang	banyak	tinggi	0.00191	0.00167	0.00373	tinggi
31	Banyak	kecil	besar	tinggi	banyak	sedang	0.00383	0.00356	0.00536	tinggi
32	Sedang	standar	besar	tinggi	sedang	tinggi	0.00014	0.00188	0.00809	tinggi
33	Sedang	besar	sedang	sedang	banyak	tinggi	0.00096	0.00333	0.00684	tinggi
34	Banyak	besar	sedang	tinggi	banyak	rendah	0.02009	0.00960	0.02219	tinggi
35	Banyak	besar	besar	tinggi	banyak	tinggi	0.02009	0.01280	0.04639	tinggi

36	Sedang	standar	besar	tinggi	sedang	tinggi	0.00014	0.00188	0.00809	tinggi
37	Banyak	standar	besar	tinggi	banyak	tinggi	0.00670	0.00569	0.02126	tinggi
38	Sedang	standar	besar	tinggi	banyak	sedang	0.00096	0.00316	0.01311	tinggi
39	Banyak	besar	besar	tinggi	sedang	tinggi	0.00287	0.00762	0.02861	tinggi
40	Banyak	besar	sedang	tinggi	sedang	tinggi	0.00287	0.00571	0.01368	tinggi
41	Banyak	standar	besar	tinggi	sedang	tinggi	0.00096	0.00339	0.01311	tinggi
42	Banyak	besar	besar	rendah	sedikit	tinggi	0.00055	0.00238	0.00302	tinggi
43	Banyak	besar	sedang	tinggi	sedang	tinggi	0.00287	0.00571	0.01368	tinggi
44	Banyak	besar	kecil	tinggi	sedang	sedang	0.00082	0.00159	0.00460	tinggi
45	Banyak	besar	besar	tinggi	sedang	tinggi	0.00287	0.00762	0.02861	tinggi
46	Sedang	besar	besar	sedang	banyak	tinggi	0.00096	0.00444	0.01430	tinggi
47	Banyak	kecil	besar	sedang	banyak	tinggi	0.00128	0.00222	0.00268	tinggi
48	Sedang	besar	sedang	sedang	banyak	sedang	0.00096	0.00333	0.00684	tinggi
49	Sedang	standar	sedang	rendah	sedang	sedang	0.00003	0.00044	0.00143	tinggi
50	Sedang	besar	besar	tinggi	sedikit	sedang	0.00041	0.00423	0.00504	tinggi
51	Sedang	besar	sedang	rendah	banyak	tinggi	0.00055	0.00167	0.00506	tinggi
52	Banyak	besar	sedang	tinggi	sedikit	sedang	0.00287	0.00571	0.00391	sedang
53	Banyak	besar	besar	tinggi	sedang	sedang	0.00287	0.00762	0.02861	tinggi
54	Banyak	besar	besar	sedang	sedikit	sedang	0.00096	0.00476	0.00409	sedang
55	Sedang	kecil	sedang	tinggi	banyak	sedang	0.00055	0.00148	0.00158	tinggi
56	Banyak	besar	sedang	tinggi	banyak	tinggi	0.02009	0.00960	0.02219	tinggi

Aplikasi PHP

Sebagai bahan pembelajaran Metode Naïve Bayes Classifier ini; dibuat database (dalam hal ini menggunakan MySQL/MariaDB Database server) sebagai berikut:

```
CREATE DATABASE IF NOT EXISTS db_dm;

USE db_dm;
```

Kesimpulan

- Metode **Naive Bayes** memanfaatkan data training untuk menghasilkan probabilitas setiap kriteria untuk *class* yang berbeda, sehingga nilai-nilai probabilitas dari kriteria tersebut dapat dioptimalkan untuk memprediksi suatu kondisi berdasarkan proses klasifikasi yang dilakukan oleh metode **Naive Bayes** itu sendiri.

6.6 Diskusi Teori Klasifikasi dengan Naïve Bayes

Misalnya terdapat ingin diketahui apakah suatu objek masuk dalam ketegori dipilih untuk perumahan atau tidak dengan algoritma Naive Bayes Classifier. Untuk menetapkan suatu daerah akan dipilih sebagai lokasi untuk mendirikan perumahan, telah dihimpun 10 aturan.

Ada 4 atribut yang digunakan, yaitu:

- harga tanah per meter persegi (C1),
- jarak daerah tersebut dari pusat kota (C2),
- ada atau tidaknya angkutan umum di daerah tersebut (C3), dan
- keputusan untuk memilih daerah tersebut sebagai lokasi perumahan (C4).

Aturan ke-	Harga tanah (C1)	Jarak dari pusat kota (C2)	Ada angkutan umum (C3)	Dipilih untuk perumahan (C4)
1	Murah	Dekat	Tidak	Ya
2	Sedang	Dekat	Tidak	Ya
3	Mahal	Dekat	Tidak	Ya
4	Mahal	Jauh	Tidak	Tidak
5	Mahal	Sedang	Tidak	Tidak
6	Sedang	Jauh	Ada	Tidak
7	Murah	Jauh	Ada	Tidak
8	Murah	Sedang	Tidak	Ya
9	Mahal	Jauh	Ada	Tidak
10	Sedang	Sedang	Ada	Ya

Untuk jenis data harga tanah dan jarak pusat kota yang kontinue, misalnya :

Aturan ke-	Harga tanah (C1)	Jarak dari pusat kota (C2)	Ada angkutan umum (C3)	Dipilih untuk perumahan (C4)
1	100	2	Tidak	Ya
2	200	1	Tidak	Ya
3	500	3	Tidak	Ya
4	600	20	Tidak	Tidak
5	550	8	Tidak	Tidak
6	250	25	Ada	Tidak
7	75	15	Ada	Tidak
8	80	10	Tidak	Ya
9	700	18	Ada	Tidak
10	180	8	Ada	Ya

BAB VII
Klastering dengan K-Means

7.1 Clustering

Pada dasarnya clustering terhadap data adalah suatu proses untuk mengelompokkan sekumpulan data tanpa suatu atribut kelas yang telah didefinisikan sebelumnya, berdasarkan pada prinsip konseptual clustering yaitu memaksimalkan dan juga meminimalkan kemiripan intra kelas. Misalnya, sekumpulan obyek-obyek komoditi pertama-tama dapat di clustering menjadi sebuah himpunan kelas-kelas dan lalu menjadi sebuah himpunan aturan-aturan yang dapat diturunkan berdasarkan suatu klasifikasi tertentu.

Proses untuk mengelompokkan secara fisik atau abstrak obyek-obyek ke dalam bentuk kelas-kelas atau obyek-obyek yang serupa, disebut dengan clustering atau unsupervised classification. Melakukan analisa dengan clustering, akan sangat membantu untuk membentuk partisi-partisi yang berguna terhadap sejumlah besar himpunan obyek dengan didasarkan pada prinsip "divide and conquer" yang mendekomposisikan suatu sistem skala besar, menjadi komponen-komponen yang lebih kecil, untuk menyederhanakan proses desain dan implementasi. Perbedaan utama antara Clustering Analysis dan klasifikasi adalah bahwa Clustering Analysis digunakan untuk memprediksi kelas dalam format bilangan real dan pada format katagorikal atau Boolean.

a. Data Clustering

Data Clustering merupakan salah satu metode data mining yang bersifat tanpa arahan (unsupervised). Ada dua jenis data clustering yang sering dipergunakan dalam proses pengelompokan data yaitu hierarchical dataclustering dan non-hierarchical dataclustering. K-Means merupakan salah satu metode data clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster/kelompok. Metode ini mempartisi data ke dalam cluster/kelompok sehingga data yang memiliki karakteristik yang sama dikelompokkan ke dalam satu cluster yang sama dan data yang mempunyai

karakteristik yang berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan dari data clustering ini adalah untuk meminimalisasikan objective function yang diset dalam proses clustering, yang pada umumnya berusaha meminimalisasikan variasi di dalam suatu cluster dan memaksimalisasikan variasi antar cluster.

Data clustering menggunakan metode K-Means ini secara umum dilakukan dengan algoritma dasar sebagai berikut:

1. Tentukan jumlah cluster
2. Alokasikan data ke dalam cluster secara random
3. Hitung centroid (rata-rata) dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid (rata-rata) terdekat

Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan

7.2 Algoritma *K-Means*

K-means merupakan salah satu algoritma clustering.Tujuan algoritma ini yaitu untuk membagi data menjadi beberapa kelompok. Algoritma ini menerima masukan berupa data tanpa label kelas. Hal ini berbeda dengan supervised learning yang menerima masukan berupa vektor (-x-1, y1), (-x-2 , y2) , …, (-x-i, yi), di mana xi merupakan data dari suatu data pelatihan dan yi merupakan label kelas untuk xi.

Pada algoritma pembelajaran ini, komputer mengelompokkan sendiri data-data yang menjadi masukannya tanpa mengetahui terlebih dulu target kelasnya. Pembelajaran ini termasuk dalam unsupervised learning. Masukan yang diterima adalah data atau objek dan k buah kelompok (cluster) yang diinginkan. Algoritma ini akan mengelompokkan data atau objek ke dalam k buah kelompok tersebut. Pada setiap cluster terdapat titik pusat (centroid) yang merepresentasikan cluster tersebut.

Data clustering menggunakan metode K-Means ini secara umum dilakukan dengan algoritma dasar sebagai berikut:

1. Tentukan jumlah cluster
2. Alokasikan data ke dalam cluster secara random
3. Hitung centroid (rata-rata) dari data yang ada di masing-masing cluster
4. Alokasikan masing-masing data ke centroid (rata-rata) terdekat
5. Kembali ke Step 3, apabila masih ada data yang berpindah cluster atau apabila perubahan nilai centroid, ada yang di atas nilai threshold yang ditentukan atau apabila perubahan nilai pada objective function yang digunakan di atas nilai threshold yang ditentukan.

Beberapa distance space telah diimplementasikan dalam menghitung jarak (distance) antara data dan centroid termasuk di antaranya

1. L1 (Manhattan/City Block) distance space,
2. L2 (Euclidean) distance space,
3. dan Lp (Minkowski) distance space.

Jarak antara dua titik x1 dan x2 pada Manhattan/City Block distance space dihitung dengan menggunakan rumus sebagai berikut:

$$D_{l1}(x_1, x_2) = \| x_2 \updownarrow x_1 \|^1 = \sum_{j-1}^{p} \left| x_{2j} \updownarrow x_{1j} \right|$$

Dimana:

P : dimensi data

|.| : nilai absolut

Sedangkan untuk L_2 (Euclidean) distance space, jarak antara dua titik dihitung menggunakan rumus sebagai berikut:

$$D_{l2}(x_2, x_1) = \| x_2 \updownarrow x_1 \| = \sqrt{\sum_{j-1}^{p} (x_{2j} \updownarrow x_{1j})^2}$$

Dimana:

P : dimensi data

L_p (Minkowski) distance space yang merupakan generalisasi dari beberapa distance space yang ada seperti L_1 (Manhattan/City Block) dan L_2 (Euclidean), juga telah diimplementasikan. Tetapi secara umum distance space yang sering digunakan adalah Manhattan dan Euclidean. Euclidean sering digunakan karena penghitungan jarak dalam distance space ini merupakan jarak terpendek yang bisa didapatkan antara dua titik yang diperhitungkan, sedangkan Manhattan sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus seperti keberadaaan outliers dengan lebih baik.

Pembaharuan suatu titik centroid dapat dilakukan dengan rumus berikut:

$$\mu k = \frac{1}{N_k} \sum_{q=1}^{N_k} X_q$$

Dimana:

μk = titik centroid dari cluster ke-K

Nk = banyaknya data pada cluster ke-K

xq = data ke-q pada cluster ke-K

7.3 Kelemahan dan Kelebihan *K-Means*

Ada beberapa kelebihan pada algoritma k-means, yaitu:

a. Mudah untuk diimplementasikan dan dijalankan.
b. Waktu yang dibutuhkan untuk menjalankan pembelajaran ini relatif cepat.
c. Mudah untuk diadaptasi.
d. Umum digunakan. Algoritma k-means memiliki beberapa kelebihan,

Adapun kelemahan dari K-Means, yaitu:

a. Sebelum algoritma di jalankan, titik K diinisialisasikan secara random sehingga pengelompokan data yang di dapatkan bisa berbeda-beda. Namun apabila nilai yang diperoleh acak untuk penginisialisasi kurang baik maka pengelompokan yang didapatkn menjadi tidak optimal.
b. Apabila terjebak dalam kasus yang biasanya di sebut dengan curse of dimensionality. Hal ini pun akan terjadi apabila salah satu data untuk

melakukan pelatihan mempunyai dimensi yang sangat banyak, sebagai contoh; jika ada data pelatihan yang terdiri dari 2 buah atribut saja maka dimensinya ada 2 dimensi pula, namun akan berbeda jika ada 20 atribut maka akan ada 20 dimensi yang di miliki. Adapun salah satu dari cara kerja algoritma cluster ini ialah untuk mencari jarak terdekat dari antara k titik dangan titik lainnya. Apabila ingin mencari jarak untuk antar titik dari 2 dimensi hal itu masih mudah untuk di lakukan, namun bagaimana dengan 20 buah dimensi hal tersebut akan menjadi lebih sulit untuk di lakukan pencarian jarak.

c. Apabila hanya ada terdapat beberapa buah titik sampel data yang ada, maka hal yang mudah untuk melakukan penghitungan dan mencari jarak titik terdekat dengan k titik yang telah di lakukan inisialisasi yang secara acak. Namun jika ada banyak titik data, misalkan satu juta data, maka perhitungan dan pencarian titik terdekat akan sangat membutuhkan waktu yang lama. Proses tersebut dapat dipercepat namun dibutuhkan sebuah struktur data yang lebih rumit seperti kD-tree atau hashing untuk melakukan proses tersebut.

d. Adanya penggunaan k buah random, tidak ada jaminan untuk menemukan kumpulan cluster yang optimal.

7.4 Perhitungan manual metode *clustering* K-Means

Berikut adalah contoh perhitungan manual mengenai algoritma k-means antara lain:

1. Data set

   Tabel 6. 1 merupakan tabel dataset dari 15 mahasiswa yang memprogramkan mata kuliah Data mining. Dari 15 mahasiswa tersebut akan dikelompokkan menjadi 3 bagian yaitu kelompok pintar, sedang dan kurang.

Tabel 7.1 Dataset Mahasiswa

NO	NAMA MAHASISWA	UTS	TUGAS	UAS
1	Roy	89	90	75
2	Sintia	90	71	95
3	Iqbal	70	75	80
4	Dilan	45	65	59
5	Ratna	65	75	53
6	Merry	80	70	75
7	Rudi	90	85	81
8	Hafiz	70	70	73
9	Gede	96	93	85
10	Christian	60	55	48
11	Justin	45	60	58
12	Jesika	60	70	72
13	Ayu	85	90	88
14	Siska	52	68	55
15	Reitama	40	60	70

2. Setelah menentukan dataset, maka perlu menentukan jumlah cluster yang akan dibentuk. Adapun cluster yang akan dibentuk adalah:
   a. Cluster 1 (C1) = Pintar
   b. Cluster 2 (C1) = Sedang
   c. Cluster 3 (C1) = Kurang
3. Tetapkan C pusat cluster awal secara random

   Dari dataset diatas terpilih 3 cluster pusat diantaranya :

Tabel 7.2 Nilai Pusat Cluster ditentukan secara rabdom

Kluster 1	96	93	85
Kluster 2	70	75	80
Kluster 3	60	55	48

4. Alokasikan semua data/obyek ke dalam cluster terdekat. Berikut hasil dari alokasi data ke jarak cluster.

   Adapun hasil dari jarak ke cluster diperoleh dari perhitungan dengan rumus :

$$d_{ij} = \sqrt{\sum_{k-1}^{p} \{x_{jk} - x_{jk}\}^2}$$

$$d(1,1) = \sqrt{(89-96)^2 + (90-93)^2 + (75-85)^2} = 12{,}56981$$
$$d(1,2) = \sqrt{(89-96)^2 + (90-75)^2 + (75-80)^2} = 24{,}71841$$
$$d(1,3) = \sqrt{(89-60)^2 + (90-55)^2 + (75-48)^2} = 52{,}86776$$

$$d(2,1) = \sqrt{(90-96)^2 + (71-93)^2 + (95-85)^2} = 24{,}8998$$
$$d(2,2) = \sqrt{(90-70)^2 + (71-75)^2 + (95-80)^2} = 25{,}3179$$
$$d(2,3) = \sqrt{(90-60)^2 + (71-55)^2 + (95-80)^2} = 58{,}0086$$

$$d(3,1) = \sqrt{(70-96)^2 + (75-93)^2 + (80-85)^2} = 32{,}01562$$
$$d(3,2) = \sqrt{(70-70)^2 + (75-75)^2 + (80-80)^2} = 0$$
$$d(3,3) = \sqrt{(70-60)^2 + (75-55)^2 + (80-48)^2} = 39{,}03844$$
$$d(4,1) = \sqrt{(45-96)^2 + (65-93)^2 + (59-85)^2} = 63{,}72598$$
$$d(4,2) = \sqrt{(45-70)^2 + (65-75)^2 + (59-80)^2} = 34{,}14674$$
$$d(4,3) = \sqrt{(45-60)^2 + (65-55)^2 + (59-48)^2} = 21{,}11871$$

$$d(5,1) = \sqrt{(65-96)^2 + (75-93)^2 + (53-85)^2} = 48{,}05206$$
$$d(5,2) = \sqrt{(45-70)^2 + (65-75)^2 + (59-80)^2} = 27{,}45906$$
$$d(5,3) = \sqrt{(45-60)^2 + (65-55)^2 + (59-48)^2} = 21{,}2132$$

(lakukan perhitungan tersebut sampai data ke 15)

Setelah melakukan perhitungan maka didapat hasil seperti berikut ini :

Tabel 7.3 Hasil *Clustering* Data Mahasiswa yang Mengikuti Kuliah Data Mining Iterasi Ke-1

No	Nama Mahasiswa	Jarak Ke Cluster			Hasil
		C1	C2	C3	
1.	Roy	12,56980509	24,71841419	52,86775955	1
2.	Sintia	24,8997992	25,3179778	58,00862005	1
3.	Iqbal	32,01562119	0	39,03844259	2
4.	Dilan	63,72597587	34,14674216	21,11871208	3
5.	Ratna	48,05205511	27,45906044	21,21320344	3
6.	Merry	29,74894956	12,24744871	36,79673899	2
7.	Rudi	10,77032961	22,38302929	53,7494186	1
8.	Hafiz	36,72873534	8,602325267	30,82207001	2
9.	Gede	0	32,01562119	64,10148204	1
10.	Christian	64,10148204	39,03844259	0	3
11.	Justin	66,47555942	36,52396474	18,70828693	3
12.	Jesika	44,65422712	13,74772708	28,3019434	2
13.	Ayu	11,78982612	22,6715681	58,73670062	1
14.	Siska	58,83026432	31,591138	16,79285562	3
15.	Reitama	66,70832032	35	30,14962686	3

5. Tentukan kembali titik pusat cluster yang baru berdasarkan rata-rata

   Cluster baru tersebut didapat dari rumus = nilai hasil /banyak hasil

   Kluster 1 (UTS) = (89+90+90+90+85)/5=90

   Kluster 1 (Tugas) = (90+71+85+93+90)/5=85,8

   Kluster 1 (UAS) = (75+95+81+85+88)/5=84,8

   Lakukan, perhitungan tersebutuntuk kluster 2 dan 3, sehingga didapat nilai cluster baru antara lain :

Tabel 7.4 Nilai Pusat *Cluster* Data Mahasiswa yang Mengikuti Kuliah Data Mining Iterasi Ke-1

Kluster 1	90	85,8	84,8
Kluster 2	70	71,25	75
Kluster 3	51,16666667	63,83333333	57,16666667

6. Lakukan kembali langkah 4 hingga titik pusat dari setiap cluster tidak berubah. Berikut hasil yang didapat sesuai dengan langkah ke 4.

Tabel 7.5 Hasil *Clustering* Data Mahasiswa yang Mengikuti Kuliah Data Mining Iterasi Ke-2

No	Nama Mahasiswa	Jarak Ke Cluster			Hasil
		C1	C2	C3	
1.	Roy	10,70887482	26,69386634	49,33643008	1
2.	Sintia	17,97442628	28,28537608	54,68775	1
3.	Iqbal	23,23101375	6,25	31,63463292	2
4.	Dilan	55,88631317	30,33253204	6,538348415	3
5.	Ratna	41,86740976	22,87055968	18,25970062	3
6.	Merry	21,1111345	10,07782219	34,45891273	2
7.	Rudi	3,883297568	25,00124997	50,2402561	1
8.	Hafiz	28,08700767	2,358495283	25,3656592	2
9.	Gede	9,374433316	35,34207832	60,29441655	1
10.	Christian	56,59399261	33,06149573	15,49462272	3
11.	Justin	58,38561467	32,25775101	7,30867065	3
12.	Jesika	36,24196463	10,51487042	18,33257574	2
13.	Ayu	7,271863585	27,30499039	52,72649555	1
14.	Siska	51,46727115	27,10281351	4,769696007	3
15.	Reitama	58,17800272	32,42780443	17,43798536	3

7. Hasil dari tahapan yang pertama dan kedua tidak berubah, maka hasil sudah sesuai dengan pengelompokkan kluster. Berikut adalah hasil dari pengelompokkan tersebut.

Tabel 7.6 Hasil *Clustering* Data Mahasiswa yang Mengikuti Kuliah Data Mining Iterasi Ke-3

No	Nama Mahasiswa	UTS	Tugas	UAS	Kelompok
1.	Roy	89	90	75	Pintar
2.	Sintia	90	71	95	Pintar
3.	Iqbal	70	75	80	Sedang
4.	Dilan	45	65	59	Kurang
5.	Ratna	65	75	53	Kurang
6.	Merry	80	70	75	Sedang
7.	Rudi	90	85	81	Pintar
8.	Hafiz	70	70	73	Sedang
9.	Gede	96	93	85	Pintar
10.	Christian	60	55	48	Kurang
11.	Justin	45	60	58	Kurang
12.	Jesika	60	70	72	Sedang
13.	Ayu	85	90	88	Pintar
14.	Siska	52	68	55	Kurang
15.	Reitama	40	60	70	Kurang

DAFTAR PUSTAKA

Abrori, M., & Setiyani, N. (2015). Implementasi Algoritma Best-First Search(BeFS) Pada Penyelesaian Traveling Salesman Problem(TSP) (Studi Kasus: Perjalanan Wisata di Kota Yogyakarta). *JURNAL FOURIER*, *4*(2), 93–111.

Balaji, S., & Murugaiyan, M. S. (2012). Wateerfall Vs V-model Vs Agile : A Comparative Study On SDLC. *International Journal of Information Technology and Business Management*, *2*(1), 26–30.

Dincer, A., & Uraz, B. (2013). *Google Maps JavaScript API Cookbook*. Birmingham: Packt Publishing.

Hutami, D. W., & Mahmudy, W. F. (2017). Implementasi Algoritma Nearest Insertion Heuristic dan Modified Nearest Insertion Heuristic Pada Optimasi Rute Kendaraan Pengangkut Sampah (Studi Kasus : Dinas Kebersihan dan Pertamanan Kota Malang). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(2), 95–99.

Jacobson, L., & Kanber, B. (2015). *Genetic Algorithms in Java Basics*. Berkeley, California: Apress Media.

Kramer, O. (2017). *Genetic Algorithm Essentials*. Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-52156-5

Kumari, J., & Dubey, A. K. (2016). A Review Paper on Genetic Algorithm. *International Journal of Advance Research in Computer Science and Management Studies*, *4*(7), 122–125.

Mahmudy, W. F. (2014). Improved Simulated Annealing for Optimization of Vehicle Routing Problem with Time Windows (VRPTW). *Kursor Journal*, *7*(3), 109–116.

N. Sivanandam, S., & Deepa, S. N. (2008). *Introduction to Genetic Algorithms* (1st ed.). Berlin: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-73190-0

Petroutsos, E. (2014). *Google Maps Power Tools for Maximizing the API*. New York: McGraw-Hill Education.

Priandani, N. D., & Mahmudy, W. F. (2015). Optimasi Travelling Salesman Problem With Time Windows (TSP-TW) pada Penjadwalan Paket Rute Wisata di Pulau Bali Menggunakan Algoritma Genetika. *Seminar Nasional Sistem Informasi Indonesia*, 259–266.

Purnia, D. S., & Riana, D. (2016). Pencarian Rute Terpendek Perjalan Promosi Marketing Menggunakan Algoritma Genetika dan Algoritma Greedy. *INFORMATIKA*, *3*(2), 299–313.

Risdwiyanto, A., & Kurniyati, Y. (2015). Strategi Pemasaran Perguruan Tinggi Swasta di Kabupaten Sleman Yogyakarta Berbasis Rangsangan Pemasaran. *Jurnal MAKSIPRENEUR*, *5*(1), 1–23.

Samana, E., Prihandono, B., & Noviani, E. (2015). Aplikasi Simulated Annealing Untuk Menyelesaikan Travelling Salesman Problem. *Bimaster*, *03*(1), 25–32.

Shita, R. T., & Subandi. (2017). Implementasi Algoritma Genetika Pada Aplikasi Pemetaan Distribusi Barang Berbasis Web. *Jurnal Telematika Mkom*, *9*(3), 114–118.

Sholeh, M., Widyastuti, N., & Pratama, M. (2017). Google Map for Implementation of Geographic Information System Development Search Location SMEs. *International Journal of Engineering Research & Technology (IJERT)*, *6*(2), 501–504.

Whitten, J. L., & Bentley, L. D. (2007). *System Analysis and Design Methods 7th*. (McGrawHill, Ed.) (7th ed.). New York: McGrawHill.