BAB 2
LANDASAN TEORI

2.1 Data

Data adalah fakta yang dapat digunakan sebagai input dalam menghasilkan informasi. Data dapat berupa diskusi, pengambilan keputusan, perhitungan, atau pengukuran[10]. Data memiliki dua sifat sebagai berikut:

1. Data Kuantitatif, yaitu data dalam bentuk angka atau bilangan
2. Data Kualitatif, yaitu data bukan dalam bentuk penjumlahan atau angka, melainkan dalam bentuk pernyataan atau kategori

2.2 *Data Mining*

Data mining merupakan proses yang menggunakan berbagai teknik dan alat analisis data untuk menemukan hubungan dan pola yang tersembunyi. Pendekatan dasar dalam data mining adalah untuk meringkas data dan untuk mengekstrak informasi berguna yang masuk akal dan sebelumnya tidak diketahui. Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat interaktif di mana pemakaian terlibat langsung atau dengan perantara *knowledge base* [11]. Ada beberapa tugas yang dapat dilakukan oleh *data mining* dalam proses pemecahan masalah dan pencarian pengetahuan baru, di antaranya adalah sebagai berikut:

1. Klastering *(Clustering)*

Digunakan untuk mengelompokan atau mengidentifikasi data yang memiliki karakteristik tertentu.

2. Klasifikasi *(Classification)*

Digunakan untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.

3. Asosiasi *(Association)*

Digunakan untuk mengatasi masalah bisnis yang khas, yakni dengan menganalisa tabel transaksi penjualan dan mengidentifikasi produk-produk yang seringkali dibeli bersamaan oleh *customer*.

4. Estimasi *(Estimation)*

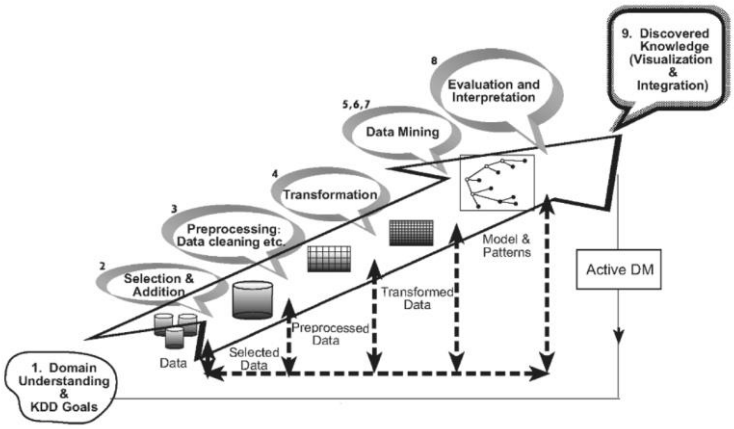
Digunakan untuk memperkirakan atau menilai sesuatu hal yang belum pernah ada sebelumnya yang disajikan dalam bentuk hasil kuantitatif (angka).

5. Prediksi *(Prediction)*

Digunakan untuk memperkirakan atau menilai sesuatu hal yang belum pernah terjadi.

2.3 Metodologi *Knowledge Discovery in Database (KDD)*

Pada penelitian ini metode yang digunakan dalam penyelesaian *data mining* adalah kerangka kerja *Knowledge Discovery in Databases (KDD)*. KDD adalah untuk menemukan pola atau informasi yang menarik atau berguna dari berbagai sumber data seperti *database*, teks, gambar, web, dan lain-lain. Hasil pola atau informasi yang didapatkan harus *valid*, memiliki potensi manfaat, dan mudah dipahami [9].



Gambar 2.1 Metode Knowledge Discovery in Database

Berdasarkan Gambar 2.1 Dapat dijelaskan tahapan-tahapan dalam KDD sebagai berikut:

1. *Domain Understanding & KDD Goals*, yaitu membangun pemahaman tentang domain aplikasi. Hal ini dilakukan sebagai persiapan awal memahami apa yang harus dilakukan dengan banyaknya pertimbangan yang perlu seperti transformasi, algoritma, representasi dan lainnya yang diperlukan. Orang bertanggung jawab atas proyek KDD harus memahani dan menentukan tujuan pengguna akhir dan lingkungan di mana pengetahuan sebelumnya yang relevan.

2. *Selection & Addition*, setelah menentukan tujuan, perlu menentukan data yang akan digunakan untuk menemukan pengetahuan. Langkah ini melibatkan pencarian informasi tentang data yang sudah ada, memperoleh tambahan yang diperlukan, dan menyatukan semua data tersebut dalam satu *dataset*. Atribut-atribut yang akan dipertimbangkan dalam proses ini juga dipilih. Proses ini sangat penting karena *data mining* belajar dan menemukan pola dari data yang digunakan sebagai dasar untuk membangun model. Jika ada beberapa atribut penting yang hilang, maka seluruh studi mungkin gagal. Dari keberhasilan proses, ada baiknya mempertimbangkan sebanyak mungkin atribut pada tahap ini. Di sisi lain, mengumpulkan, mengatur, dan mengelola data dari repositori yang kompleks itu mahal, dan ada pertukaran dengan peluang untuk mehami fenomena dengan baik. *Tradeoff* ini merupakan aspek di mana aspek interaktif dan iteraktif dari KDD berlangsung. Ini dimulai dengan kumpul data terbaik yang tersedia dan kemudian berkembang dan mengamati efeknya dalam hal penemuan dan pemodelan pengetahuan. *Data integration* proses penggabungan data dari beberapa sumber menjadi satu.
3. *Preprocessing*, pada tahap ini, untuk meningkatkan kuakuratan dan efesiensi hasil data mining. Terdapat hal yang harus dilakukan seperti pembersihan data, seperti menghapus data yang tidak perlu[12].
4. *Transformation*, pada tahap ini, data yang akan digunakan dalam penambangan data dipersiapkan dan dikembangkan untuk menjadi lebih baik. Metode ini termasuk reduksi dimensi (seperti pemilihan dan ekstraksi fitur, serta pengambil sampel rekaman) dan transformasi atribut seperti (diskritasi atribut numerik dan transformasi fungsional). Langkah ini penting untuk keberhasilan proyek KDD secara keseluruhan, tetapi bersifat sangat spesifik untuk setiap proyek.
5. *Data Mining*, memilih *data mining* yang sesuai. Memutuskan jenis *data mining* mana yang akan digunakan, misalnya, klasifikasi, regresi atau pengelompokan. Hal ini sangat bergantung pada tujuan KDD, dan juga pada langkah-langkah sebelumnya. Ada dua tujuan utama dalam *data mining*: prediksi dan deskripsi. Prediksi sering disebut sebagai *data mining* yang diawasi, sedangkan *data mining* deskriftif mencakup aspek pengawasan dan visualisasi dari *data mining*.

Sebagian besar *data mining* didasarkan pada pembelajaran induktif, di mana model di bangun secara eksplisit atau implisit dengan menggeneralisasi dari sejumlah contoh pelatihan yang cukup.

6. Memilih algoritma *data mining*, tahap ini mencakup pemilihan metode spesifik yang akan digunakan untuk mencari pola (termasuk beberapa penginduksi). Untuk setiap strategi *meta-learning* ada beberapa kemungkinan bagaimana hal itu dapat dicapai. *Meta-learning* berfokus untuk menjelaskan apa yang menyebabkan suatu algoritma *data mining* berhasil atau tidak dalam suatu pemahaman kondisi di mana algoritma *data mining* paling cepat. Setiap algoritma memiliki parameter dan taktik pembelajaran.
7. Menggunakan algoritma *data mining*, pada tahap ini perlu menggunakan algoritma beberapa kali hingga hasil yang memuaskan diperoleh, misalnya dengan menyetel parameter kontrol algoritma, seperti jumlah minimum *instance* dalam satu daun pohon keputusan.
8. *Evaluation*, pada tahap ini mengevaluasi dan menginterprestasikan pola-pola yang ditambang (aturan, keandalan, dan lain-lain), dengan mengacu pada tujuan yang didefinisikan pada langkah pertama. Di sini mempertimbangkan tahap *preprocessing* dengan mengacu pada hasil algoritma penambangan data (misalnya, menambahan fitur pada langkah 4, dan memulai dari sana). Tahap ini fokus pada kejelasan dan kegunaan model yang dihasilkan. Pada tahap ini, pengetahuan yang ditemukan juga didokumentasikan untuk digunakan kembali.
9. *Integration*, pada tahap ini siap untuk menggabungkan pengetahuan ke dalam sistem lain untuk tindakan lebih lanjut. Pengetahuan menjadi aktif dalam arti dapat membuat perubahan pada sistem dan mengukur efeknya. Keberhasilan dari langkah ini menentukan keefektifan dari keseluruhan proses KDD. Ada banyak tantangan dalam langkah ini, seperti kehilangan “laboratorium” dimana kita telah beroperasi. Sebagai contoh, pengetahuan ditemukan dari cuplikan statis tertentu (biasanya sampel) dari data, tetapi sekarang data menjadi dinamis. Struktur data dapat berubah (atribut tertentu menjadi tidak tersedia), dan domain data dapat dimodifikasi (seperti, sebuah atribut mungkin memiliki nilai yang tidak diasumsikan sebelumnya).

2.4 Normalisasi Data

Normalisasi data adalah proses memberikan bobot yang sama pada semua atribut misalnya [-1, 1], [0, 1], [0.0, 1.0]. Tujuan dari normalisasi data yaitu untuk mencegah atribut yang memiliki rentang besar melebihi atribut yang memiliki rentang kecil[13]. Adapun beberapa metode dalam normalisasi meliputi *min-max normalization, z-score normalization, dan normalization by scalling*. Pada beberapa algoritma *data mining*, perbedaan rentang akan menyebabkan variabel yang memiliki nilai rentang lebih besar memiliki pengaruh lebih besar terhadap hasil algoritma.

2.5 *Min-max Normalization*

Min-max normalization adalah normalisasi data dengan melakukan transformasi linear terhadap data. *Min-max normalization* dapat dilakukan menggunakan persamaan berikut:

$$x = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2.1}$$

Keterangan:

x= nilai yang akan dinormalisasi
min(x) = nilai minimum variabel
max(x) = nilai maksimal variable

2.6 Modus

Modus adalah suatu skor atau nilai yang mempunyai frekuensi paling banyak atau nilai yang mempunyai frekuensi maksmial dalam distribusi data. Dengan kata lain modus merupakan nilai yang paling sering muncul atau memiliki frekuensi tertinggi pada sebuah data. Modus digunakan untuk menyatakan fenomena yang paling banyak terjadi. Modus atau *mode* umumnya dilambangkan dengan *Mo*. Dalam data bisa terdapat satu modus *(unimodus)*, dua modus *(bimodus)*, leboh dari dua modus *(multimodus)*, atau sama sekali tidak memiliki modus. Jika data pengamatan memiliki jumlah frekuensi yang sama berarti data tersebut tidak memiliki modus[14].

2.7 Klasifikasi

Klasifikasi adalah suatu teknik menemukan suatu pola yang mampu memisahkan kelas data yang satu dengan yang lainnya untuk menentukan objek yang masuk dengan kategori tertentu dengan melihat kelakukan dan atribut dari kelompok yang telah didefinisikan. Teknik ini mampu mengklasifikasikan data baru dengan menggunakan hasilnya untuk memberikan sejumlah aturan [15].

2.8 *Naïve Bayes*

Naïve Bayes adalah algoritma yang dapat mengklasifikasi suatu variabel tertentu menggunakan metode probabilitas dan statistik. *Naïve bayes* dapat dilatih untuk melakukan *supervised learning* dengan sangat efektif. *Naïve bayes* tidak memerlukan jumlah data *training* yang banyak [16][17][18]. Adapun langkah-langkah metode *Naïve Bayes* yaitu sebagai berikut:

1. Menghitung nilai *Mean* dan *Standar Deviasi*

   Persamaan untuk menghitung *Mean* dan *Standar Devias* seperti yang disajikan pada persamaan (2.2):

$$\mu = \frac{\sum_i^n xi}{n} \tag{2.2}$$

Keterangan:

μ = nilai rata-rata

xi = nilai data ke-i

n = jumlah data

$$\sigma = \sqrt{\frac{\sum_i^n (xi - \mu)^2}{n - 1}} \tag{2.3}$$

Keterangan:

σ = varian satuan ragam untuk populasi

xi = titik tengah nilai dalam satu atribut

μ = rata-rata atau mean dari populasi

n = jumlah data

2. Menghitung Probabilitas Setiap Kelas

   Persamaan untuk menghitung probabilitas setiap kelas seperti yang disajikan pada persamaan (2.4).

$$P(A) = \frac{XA}{n} \tag{2.4}$$

   Keterangan:

   $P(A)$ = probabilitas untuk kelas A

   XA = jumlah data A

   n = jumlah seluruh data

3. Perhitungan Dengan Gaussian.

   Masing-masing probabilitas yang muncul pada setiap atribut di masing-masing *class* yang didapatkan dengan algoritma Naïve Bayes kemudian dihitung menggunakan gaussian. Persamaan gaussian adalah seperti yang disajikan pada persamaan (2.5).

$$P(Xi = xi|Y = yi)\frac{1}{\sqrt{2\pi.\sigma}}\exp -\frac{(xi-\mu)^2}{2\sigma^2} \tag{2.5}$$

   Keterangan:

   $P(Xi = xi|Y = yi)$ = probabilitas parameter Xi dengan nilai xi dan kelas yi

   σ = standar deviasi

   xi = nilai data pada data uji

   exp = exponent

4. Peritungan Likelihood Dari Masing-masing Kelas

   Menghitung likelihood atau kedekatan nilai probabilitasnya pada setiap *class* dengan mengalikan semua nilai probabilitas pada setiap atribut menggunakan fungsi gaussian.

5. Kelas yang Memiliki Probabilitas Nilai terbesar Adalah Hasil Dari Klasifikasi.

2.9 Keseimbangan Data *(Balance Data)*

Keseimbangan data atau *balance data* adalah keadaan di mana jumlah sampel pada setiap kelas atau kategori dalam suatu dataset memiliki proporsi yang seimbang. Hal ini penting dalam algoritma klasifikasi karena dapat memastikan bahwa model pembelajaran mesin dapat mempelajari setiap kelas secara

proporsional, sehingga tidak bias terhadap kelas yang memiliki jumlah sampel lebih dominan. Ketidakseimbangan data (imbalanced data) dapat menyebabkan model lebih cenderung memprediksi kelas yang lebih sering muncul, sehingga menurunkan akurasi prediksi pada kelas yang kurang terwakili.

2.10 *Confusion Matrix*

Pada penelitian ini *confision matrix* digunakan untuk mengevaluasi kinerja dari metode klasifikasi. *Confusion matrix* adalah metode evaluasi yang digunakan untuk menggambarkan hasil suatu model klasifikasi dengan membandingkan hasil prediksi model dengan nilai yang sebenarnya dari *data testing*. *Confision matrix* dapat membantu memahami kinerja dari model *Naïve Bayes* secara detail dan dapat memperoleh hasil klasifikasi dengan benar [19]. Ada 4 kondisi saat membuat *confusion matrix* dengan ketentuan sebagai berikut:

1. *True Positive (TP)*: data yang diklasifikasi dengan benar sebagai nilai positif
2. *True Negative (TN)*: data yang diklasifikasi dengan benar sebagai nilai negatif
3. *False Positive (FP)*: data yang diklasifikasi salah dengan nilai positif
4. *False Negative (FN)*: data yang diklasifikasi salah dengan nilai negatif

Setelah membuat *confusion matrix* diperlukan akurasi dan *f1-score,* berikut ini adalah rumus perhitungan akurasi dan *f1-score:*

A. Akurasi

Akurasi adalah nilai *matrix* yang dapat mengukur pengujian dalam sebuah sistem klasifikasi yang dapat membuat prediksi secara benar. Akurasi diartikan dengan persentase keakuratan sebuah prediksi. Berikut adalah rumus untuk menghitung nilai akurasi:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.6}$$

Keterangan:

TP = True Positive

TN= True Negative

FP= False Positive

FN= False Negative

B. F1-Score

Nilai *F1-Score* adalah nilai *matrix* yang menggabungkan nilai *presisi* dan *recall* dari sebuah sistem prediksi. *Presisi* adalah rasio antara *true positive* dengan total prediksi *positive*. Sedangkan *recall* adalah rasio antara *true positive* dengan *actual positive*. Berikut adalah rumus dari *F1-Score*:

$$Presisi = \frac{TP}{TP + FP} \tag{2.7}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

$$F1 - Score = \frac{2\ x\ Presisi\ x\ Recall}{Presisi + Recall} \tag{2.9}$$

Keterangan:

TP = True Positive

TN= True Negative

FP= False Positive

FN= False Negative

2.11 Pengujian *Blackbox*

Pengujian *blackbox* adalah tahapan pengujian yang dilakukan berdasarkan fungsional atau spesifikasi dari aplikasi. Pengujian ini tidak melakukan pemeriksaan *source code* program. Hanya melakukan pemeriksaan terhadap fungsionalitas aplikasi untuk memeriksa apakah sudah sesuai dengan kebutuhan user [20].

2.12 Kaggle

Kaggle merupakan sebuah situs daring yang menyediakan berbagai sumber daya dan kompetisi dalam bidang ilmu data dan *coding*. Ada banyak data-data yang bisa kita pelajari di Kaggle dan semua bersifat gratis. Semua data-data tersebut merupakan data yang asli, bukan karangan dan semua data tersebut ada yang bersifat sederhana, bahkan ada yang bersifat kompleks.

2.13 Pupuk NPK

Pada tanah terdapat banyak unsur hara didalamnya contohnya Nitrogen, Posfor dan Kalium sebagai unsur hara makro. Tanah yang kekurangan unsur hara dapat mengakibatkan tanaman menjadi tidak subur, daunnya menguning, kualitas buahnya menurun bahkan bisa menyebabkan gagal panen. Terpenuhinya unsur hara merupakan hal yang wajib untuk dilakukan melalui penambahan pupuk secara berkala karena ketersediaan unsur hara di alam sangat terbatas [21]. Masing - masing unsur tersebut berperan dalam pertumbuhan tanaman, untuk penjelasannya sebagai berikut:

a. Nitrogen

   Nitrogen dibutuhkan dalam jumlah besar karena berperan sebagai pembentukan sel tanaman, jaringan, dan organ tanaman. Fungsi utama nitrogen untuk tanaman adalah sebagai bahan sitesis, klorofil, protein, dan asam amino. Nitrogen merupakan bagian penting dari protein, protoplasma, klorofil, dan asam nukleat.

b. Phosphorus

   Phosphorus merupakan komponen penyusun dari beberapa enzim dan protein untuk proses *transfer* energi. Selain itu, phosphorus juga berperan dalam pertumbuhan benih, akar bunga dan buah.

c. Kalium

   Unsur kalium berperan sebagai pengatur proses fisiologi tanaman seperti fotosistesis, akumulasi, translikasi, transportasi karbohidrat dan mengatur distribusi air dalam jaringan dan makhluk hidup yang ada di dalam dan permukaan tanah, baik makhluk hidup yang paling kecil sampai yang besar.

2.14 pH Tanah

Kemampuan tanaman untuk melakukan proses penyerapan unsur hara dipengaruhi tingkat kesamaan tanah atau pH. pH merupakan kependekan dari *potential of hydrogen*. pH tanah adalah sutu standar pengukuran tingkat keasaman atau kebasaan pada suatu tanah. Sifat kimia tanah dapat dilihat dari nilai pH dan kandungan unsur hara yang terdapat di dalam tanah, dengan nilai pH optimum yaitu 7 [21].

2.15 Kelembaban

Kelembaban udara adalah banyaknya uap air yang terkandung dalam udara atau *atmosfer*. Kandungan uap air dalam udara hangat lebih banyak daripada kandungan air di dalam udara dingin. Banyaknya uap air yang dikandung, tergantung pada suhu udara. Semakin tinggi suhu udara, semakin banyak uap air yang terkandung. Oleh karena itu kelembaban udara memiliki hubungan erat dengan tingkat curah hujan [22].

2.16 Suhu

Suhu merupakan faktor penting dalam tahapan pertumbuhan tanaman. Suhu tanah memiliki peranan dalam proses fotosistesis, penyerapan air, respirasi dan transpirasi. Didapatkan bahwa suhu tanah pada area tertutupi tanaman lebih kecil dari tanah gundul. Tanah yang bagus ditandai dengan pH netral yang berdampak pada kesehatan dari tanaman [23].

2.17 Curah Hujan

Curah hujan merupakan ketinggian air hujan yang terkumpul dalam tempat yang datar, tidak menguap, tidak meresap, dan tidak mengalir. Satuan curah hujan selalu dinyatakan dalam satuan millimeter atau inchi namun untuk di Indonesia satuan curah hujan yang digunakan adalah dalam satuan millimeter (mm). Curah hujan dalam 1 (satu) millimeter memiliki arti dalam luasan satu meter persegi pada tempat yang datar tertampung air setinggi satu millimeter atau tertanpung air sebanyak satu liter [24].

2.18 Internet Of Things (IoT)

Internet Of Things atau biasa disebut IoT merupakan sebuah teknologi canggih yang memiliki konsep yang bertujuan untuk memperluas dan memperkembang manfaat dari konektivitas internet yang tersambung terus menerus, menghubungkan benda-benda sekitar agar aktvivitas sehari-hari menjadi lebih mudah dan efisien yang sangat membantu segala pekerjaan manusia. Istilah Internet Of Things terdiri dari dua bagian kata utama yaitu Internet yang menghubungkan dan mengatur sebuah konektivtas dan *things* yang memiliki arti objek atau sebuah perangkat[25].

2.19 Android

Android adalah sistem perangkat lunak yang digunakan pada *mobile device* yang meliputi sistem operasi, *middleware,* dan aplikasi inti Android SDK *(Standart Development Kit)* menyediakan alat dan *(Application Programming Interface)* API yang diperlakukan untuk memulai pengembangan aplikasi pada *platform* android menggunakan bahasa pemrogaraman java dan kotlin [26].