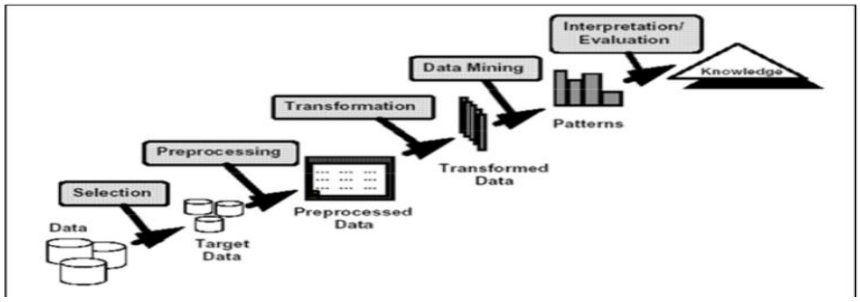
BAB II

LANDASAN TEORI

2.1 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) adalah suatu metode dalam data mining yang digunakan untuk melakukan proses pengolahan data yang telah ada dalam database. Tujuan utamanya adalah untuk mengolah data tersebut guna mendapatkan informasi baru yang lebih mudah dipahami.[1]

Knowledge Discovery in Databases yang sangatlah besar.KDD adalah proses terorganisir identifikasi yang pola data baru valid data berguna dan dapat dimengerti dari kumpulan data sangat besar dan kompleks.(DM) Data Mining adalah inti dari merupakan tahap KDD (Knowledge Discovery in Database), yang melibatkan menyimpulkan algoritma yang mengeksplorasi data, mengembangkan model dan menemukan pola-pola sebelumnya yang tidak diketahui. Model ini digunakan untuk memahami fenomena dari analisis, data dan prediksi[2]



Gambar 2.1 Tahapan Proses KDD

Berikut adalah penjelasan mendetail mengenai tahapan-tahapan dalam Knowledge Discovery In Database (KDD):

1. Seleksi atau Pemilihan Data

   Pada fase ini, data yang relavan untuk analisis didentifikasi dan dipilih. Pemilihan data yang tepat sangatlah penting karena kualitas dan relavansi data mempengaruhi hasil akhir dari proses KDD.

2. Pemrosesan dan Pembersihan Data

   Pemrosesan dan pembersihan data adalah dua tahapan krusial dalam siklus *Knowledge Discovery In Database (KDD).* Proses ini bertujuan untuk memastikan bahwa data yang digunakan dalam analisis atau pemodelan tidak hanya lengkap dan konsisten, tetapi juga sesuai dan siap digunakan dalam model analisis. Tanpa pemrosesan dan pembersihan data yang tepat, hasil analisis atau model yang dibangun mungkin menyesatkan atau tidak valid.

3. Transformasi

   Setelah data dibersihkan, Langkah selanjutnya adalah mentransformasikan data. Transformasi data bertujuan untuk mengubah data yang telah diseleksi dan dibersihkan kedalam format atau bentuk yang lebih sesuai untuk dianalisis lebih lanjut. Pada tahap ini, menggunakan Teknik tertentu, seperti data mining.

4. Data Mining

   Data mining adalah langkah utama KDD. Pada fase ini diterapkan teknik analisis algoritma untuk menemukan pola atau informasi tersembunyi dalam data. Data mining melibatkan penggunaan berbagai teknik statistik dan pembelajaran mesin untuk menganalisis data dan menemukan hubungan atau pola yang sebelumnya tidak diketahui.

5. Evaluasi

   Setelah pola atau model sudah ditemukan maka langkah selanjutnya adalah evaluasi. Pada fase ini model yang ditemukan diuji untuk mengevaluasi efektivitasnya dalam mendeskripsikan data hasil yang dinginkan. Evaluasi bertujuan untuk mengetahui apakah pola atau model yang ditemukan cukup valid dan layak digunakan dalam pengambilan keputusan.

2.2 Data Mining

Data mining merupakan suatu proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting yang tekandung dalam data. Informasi yang diperoleh dapat berupa angka atau informasi yang digunakan untuk berbagai tujuan. Data mining juga dikenal dengan sebagai "penambangan data" atau "penemuan pengetahuan basis data" atau *Knowledge Discovery In Database (KDD),* yang mencakup serangkaian proses yang bertujuan untuk menemukan informasi yang relevan dan dapat digunakan dari Big Data.

Menekankan bahwa data mining adalah serangkaian proses untuk mengekstrak pola data menggunakan algoritma seperti Naive Bayes dan metode pembobotan seperti SAW. Mereka menunjukkan bahwa data mining juga dapat membantu dalam pengambilan keputusan berbasis data dalam organisasi[3]

Data mining adalah proses yang menggunakan teknik statistic, matematika, kecerdasan buatan dan pembelajaran mesin (machine learning) mengekstrasi danmengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database yang terkait.[4]

mendefinisikan data mining sebagai serangkaian proses analitik yang menggunakan teknik statistik dan algoritma machine learning untuk mengidentifikasi pola tersembunyi dalam kumpulan data besar. Penelitian ini menggarisbawahi bahwa data mining sangat relevan untuk segmentasi pasien dalam layanan kesehatan, di mana proses ini memungkinkan pengelompokan pasien berdasarkan pola tertentu untuk meningkatkan efisiensi layanan kesehatan[4]

menjelaskan bahwa data mining, khususnya menggunakan algoritma seperti *Naive Bayes* dan *Chi-Square*, adalah alat yang sangat berguna untuk analisis sentimen berbasis teks. Penerapannya mencakup klasifikasi opini dan prediksi kepuasan pelanggan di berbagai sektor[5]

2.3 Algoritma Naïve Bayes

Naive Bayes adalah salah satu metode klasifikasi statistik yang konsep dasarnya adalah teorema Bayes yang digunakan untuk menghitung peluang dari satu kelas dari masing-masing kelompok kriteria/fitur yang ada, serta dapat menentukan mana kelas yang paling optimal. Dalam studi pembandingan algoritma klasifikasi, didapatkan bahwa metode naïve bayes nenunjukkan kecepatan dan menghasilkan akurasi yang tinggi jika diterapkan pada dataset yang besar. Naive bayes sering dipakai untuk mengatasi permasalahan klasifikasi dikarenakan naïve bayes menurut literature yang telah ada memiliki tingkat akurasi yang tinggi serta cara perhitungannya yang sederhana [3]

Naïve bayes mengidentifikasi nilai atau nilai kunci dengan menghitung variabel atau kemunculan nilai dalam data yang menghasilkan probabilitas elemen tertentu. Untuk menghitung probabilitas (probabilitas posterior) dalam naïve

bayes, algoritma harus menentukan probabilitas sebelumnya, kemungkinan, prediktor probabilitas sebelumnya. Persamaan 3 menjelaskan bahwa jika datanya multi-fitur,[6]

Naïve bayes adalah salah satu algoritma dalam teknik data mining yang menerapkan teori bayes dalam klasifikasi . Model algoritma naïve bayes memiliki tingkat kesalahan yang sangat minimum dan di kenal dengan perhitungannya yang sederhana, cepat, dan sangat akurat. Metode naïve bayes juga dianggap bekerja lebih baik dari pada model pengklasifikasi lainnya karena memiliki tingkat akurasi yang lebih baik . Namun, algoritma naïve bayes juga memiliki kelemahan, yaitu prediksi probalitas berjalan tidak optimal, dan kurangnya pemilihan fitur yang relavan dengan klasifikasi menyebabkan akurasi yang rendah.[7]

Dalam aplikasinya, Naive Bayes sering digunakan dalam berbagai bidang, termasuk analisis sentimen, klasifikasi email spam, dan pengelompokan dokumen. Misalnya, dalam analisis sentimen, Naive Bayes dapat digunakan untuk mengklasifikasikan opini pengguna terhadap produk atau layanan berdasarkan data teks yang diambil dari platform media sosial[8]. Selain itu, algoritma ini juga dapat dioptimalkan dengan teknik lain, seperti pemilihan fitur menggunakan Information Gain, untuk meningkatkan akurasi klasifikasi [9]. Penelitian juga menunjukkan bahwa Naive Bayes dapat bersaing dengan algoritma lain seperti K-Nearest Neighbor (K-NN) dan Decision Tree dalam hal akurasi klasifikasi [10].

Untuk perhitungan pada metode Naive Bayes menggunakan rumus sebagai berikut.

$$P\,(A|B) = \frac{P\,(B|A)\,.P\,(A)}{P\,(B)}$$

P (A|B) = Probabilitas A bersyarat yang diberikan oleh B
P (B|A) = Probabilitas B bersyarat yang diberikan oleh A
P (A) = Probabilitas kejadian A
P (B) = Probabilitas kejadian B

2.3.1 Kepuasan Pelayanan di RSUD

Kepuasan pelayanan di RSUD (Rumah Sakit Umum Daerah) menjadi salah satu indikator penting dalam menentukan kualitas pelayanan kesehatan yang diberikan. Tingkat kepuasan pasien dipengaruhi oleh berbagai dimensi, seperti kualitas interaksi dengan tenaga kesehatan, fasilitas yang tersedia, dan efisiensi sistem administrasi.

kepuasan pasien di RSUD sangat bergantung pada kualitas pelayanan kesehatan yang mencakup aspek-aspek seperti responsivitas tenaga medis, keandalan sistem, dan empati terhadap pasien. Penelitian ini menemukan bahwa rumah sakit dengan tenaga medis yang responsif dan ramah memiliki tingkat kepuasan pasien yang lebih tinggi dibandingkan rumah sakit dengan tenaga medis yang kurang komunikatif.[11]

menyoroti peran fasilitas fisik dan kebersihan sebagai faktor penting dalam meningkatkan kenyamanan pasien. Temuan ini menunjukkan bahwa RSUD yang memperhatikan kebersihan ruang rawat inap dan ruang tunggu memiliki tingkat kepuasan yang lebih baik. Selain itu, pelayanan administrasi yang cepat dan mudah juga memberikan kontribusi signifikan terhadap pengalaman pasien.[12]

Penelitian oleh [13].menunjukkan bahwa responsivitas penyedia layanan berkontribusi secara signifikan terhadap tingkat kepuasan pasien. Faktor-faktor ini sering dievaluasi melalui survei yang mencakup dimensi-dimensi seperti empati, kecepatan layanan, dan keandalan penyedia layanan kesehatan.

Indikator utama kepuasan pengguna dalam layanan kesehatan sangat penting untuk mengevaluasi kualitas pelayanan yang diberikan. Berbagai penelitian menunjukkan bahwa ada beberapa dimensi yang berkontribusi terhadap kepuasan pengguna, termasuk kualitas layanan, kualitas informasi, dan kualitas sistem. Kualitas layanan, yang mencakup aspek seperti kecepatan, ketepatan, dan empati dalam pelayanan, merupakan faktor dominan yang mempengaruhi kepuasan pengguna.[14]

Pengukuran kepuasan pengguna juga dapat dilakukan dengan menggunakan berbagai metode, seperti servqual dana eucs, yang masing-masing mengukur dimensi kualitas layanan dari perspektif pengguna. Metode servqual, misalnya, mencakup lima dimensi yaitu tangibles, reliability, responsiveness, assurance, dan empathy, yang semuanya berkontribusi terhadap persepsi pengguna mengenai kualitas layanan yang diterima[15]. Sementara itu, EUCS lebih fokus pada aspek kepuasan pengguna terhadap sistem informasi yang digunakan[16].

Dalam konteks aplikasi teknologi informasi dalam layanan kesehatan, survei kepuasan pengguna juga dilakukan untuk mengevaluasi aplikasi kesehatan digital. Penelitian menunjukkan bahwa aplikasi m-health yang digunakan selama pandemi COVID-19 mengalami berbagai kendala, dan evaluasi kepuasan pengguna dilakukan untuk mengidentifikasi masalah dan meningkatkan kualitas layanan[17].

Salah satu pendekatan yang umum digunakan dalam survei kepuasan adalah Survei Kepuasan Masyarakat (SKM), yang bertujuan untuk mengukur tingkat kepuasan pasien sebagai pengguna layanan kesehatan. Penelitian menunjukkan bahwa variabel yang digunakan dalam survei harus relevan, valid, dan reliabel agar dapat mencerminkan faktor-faktor yang benar-benar mempengaruhi kepuasan pasien[18].

2.3.2 Fasilitas BPJS Kesehatan

BPJS Kesehatan adalah program jaminan sosial yang bertujuan memberikan akses pelayanan kesehatan yang terjangkau dan berkualitas bagi seluruh masyarakat Indonesia. Di rumah sakit, BPJS Kesehatan bekerja sama dengan fasilitas kesehatan rujukan tingkat lanjutan (FKRTL) untuk menyediakan berbagai jenis layanan kesehatan sesuai kebutuhan peserta. Penyelenggaraan fasilitas BPJS Kesehatan di rumah sakit ini meliputi berbagai jenis layanan, dari rawat jalan hingga rawat inap, dengan prosedur yang diatur berdasarkan sistem rujukan berjenjang.

Adapun Jenis Fasilitas BPJS Kesehatan di Rumah Sakit sebagai berikut:

1. Pelayanan Rawat Jalan

   Pelayanan ini mencakup konsultasi dengan dokter spesialis, pemeriksaan diagnostik, dan pengobatan non-rawat inap. Berdasarkan penelitian [11]kualitas pelayanan rawat jalan yang cepat dan akurat menjadi salah satu indikator utama kepuasan peserta BPJS di rumah sakit

2. Pelayanan Rawat Inap

   Peserta BPJS Kesehatan berhak mendapatkan pelayanan rawat inap di rumah sakit sesuai dengan kelas yang didaftarkan (kelas I, II, atau III). fasilitas rawat inap untuk kelas III sering kali mengalami kekurangan kapasitas, menyebabkan waktu tunggu yang lama bagi pasien[18]

3. Pelayanan Diagnostik dan Tindakan Medis

   Rumah sakit yang bekerja sama dengan BPJS Kesehatan menyediakan fasilitas untuk tindakan medis dan diagnostik seperti operasi, radiologi, dan laboratorium. Penelitian oleh[17] menunjukkan bahwa integrasi fasilitas diagnostik dengan teknologi digital, seperti rekam medis elektronik, meningkatkan efisiensi pelayanan bagi peserta

4. Penyediaan Obat-Obatan

   Peserta BPJS Kesehatan berhak mendapatkan obat-obatan yang tercantum dalam Formularium Nasional (FORNAS).[14] mencatat bahwa ketersediaan obat di rumah sakit menjadi salah satu faktor utama yang memengaruhi kepuasan pasien.

2.3.3 Analisis Sentimen

Analisis sentimen adalah teknik yang digunakan untuk mengevaluasi opini atau emosi yang terkandung dalam data berbasis teks. Dalam konteks ini, algoritma Naive Bayes digunakan untuk mengklasifikasikan sentimen ke dalam kategori seperti "positif," "negatif," atau "netral." Penelitian oleh[5] Hamzah menunjukkan bahwa algoritma ini memiliki tingkat akurasi yang baik untuk analisis sentimen ulasan pelanggan dan dapat diterapkan untuk mengevaluasi tingkat kepuasan pengguna layanan kesehatan.[19]

Secara keseluruhan, analisis sentimen merupakan alat yang sangat berharga bagi perusahaan untuk memahami dan meningkatkan kepuasan pelanggan. Dengan memanfaatkan teknologi analisis yang canggih, perusahaan dapat mengumpulkan dan menganalisis data pelanggan secara efektif, yang pada akhirnya dapat meningkatkan loyalitas pelanggan dan kinerja bisnis secara keselur uhan[20].

2.4. Tahapan Penelitian

Dalam implementasinya, algoritma Naive Bayes memerlukan data pelatihan untuk menghitung probabilitas bersyarat dari fitur-fitur yang ada. Proses ini melibatkan penghitungan frekuensi nilai fitur dalam dataset yang diberikan. Berdasarkan perhitungan tersebut, algoritma akan mengklasifikasikan data baru berdasarkan model probabilistik yang dihasilkan.

Metodologi dalam penerapan algoritma ini meliputi langkah-langkah berikut:

1. **Pengumpulan Data**:

   Dataset yang relevan dikumpulkan dari survei, ulasan, atau sumber lainnya.

2. **Preprocessing Data**:

   Data yang tidak terstruktur, seperti teks, dibersihkan dan diubah menjadi format numerik.

3. **Pelatihan Model**:

   Model probabilistik dibangun berdasarkan data pelatihan.

4. **Evaluasi Model**:

   Akurasi algoritma diuji menggunakan data validasi untuk menilai performa model.

5. **Prediksi Data Baru**:

Data baru diklasifikasikan menggunakan model yang telah dilatih[21]

2.5 Penelitian Terdahulu

Penelitian sebelumnya menunjukkan bahwa algoritma Naive Bayes memiliki performa yang baik dalam berbagai aplikasi, termasuk analisis sentimen dan klasifikasi teks. Sebagai contoh:

Tabel 2.1. Penelitian Terdahulu

Judul	Penulis	Hasil	Hubungan	Kekurangan
Penerapan Algoritma Naive Bayes untuk Analisis Kepuasan Penggunaan Aplikasi Bank [22]	D. Sepri	Naive Bayes menunjukkan akurasi tinggi untuk memprediksi tingkat kepuasan pengguna aplikasi perbankan berbasis fitur layanan.	Memberikan kerangka yang relevan dalam menganalisis data kepuasan pengguna aplikasi berdasarkan fitur spesifik.	Tidak mengintegrasikan dimensi psikologis pengguna ke dalam model analisis.
Analisis Tingkat Kepuasan Pengguna Aplikasi JMO (Jamsostek Mobile) Menggunakan Algoritma Naive Bayes Classifier [23]	A. Novirian dini, H. Hermanto, DA Ambarsari	Algoritma Naive Bayes menghasilkan prediksi tingkat kepuasan pengguna aplikasi dengan akurasi tinggi.	Metode ini relevan karena menunjukkan efektivitas Naive Bayes dalam aplikasi layanan publik seperti BPJS Kesehatan.	Tidak mencakup analisis mendalam terhadap dimensi layanan spesifik.

Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes	Esti Mulyani, Fachrul Pralienka Bani Muhamad, Kurnia Adi Cahyanto	Akurasi tertinggi menggunakan unigram dengan 74.4%. Menggunakan TF-IDF dan Information Gain untuk seleksi fitur.	Memberikan wawasan tentang peningkatan akurasi Naive Bayes melalui teknik seleksi fitur dan pengolahan data teks.	Fokus pada klasifikasi judul buku, tidak mencakup sentimen pengguna media sosial.
Analisis Data Kepuasan Pengguna Layanan E-Wallet Gopay Menggunakan Metode Algoritma Naive Bayes Classifier [23]	IGI Sudipa, dkk.	Naive Bayes menunjukkan kinerja efisien dalam memprediksi kepuasan pengguna layanan berbasis data kuantitatif.	Memperlihatkan kesamaan Naive Bayes dalam klasifikasi tingkat kepuasan layanan berbasis teknologi.	Terbatas pada data numerik tanpa pengolahan data opini berbasis teks.
Perbandingan Akurasi Metode C4.5 dan Naive Bayes untuk Klasifikasi Tingkat Kepuasan Mahasiswa [24]	F. Fatmawati, W. Widiantoro, N. Narti	Naive Bayes memiliki kinerja kompetitif dibandingkan metode lain, seperti C4.5, dalam menilai kepuasan berdasarkan data survei multidimensi.	Menguatkan relevansi Naive Bayes untuk menilai kepuasan pengguna berdasarkan survei multidimensi seperti SKM BPJS.	Tidak mencakup evaluasi kualitas sistem yang digunakan dalam layanan.

2.6 Kelebihan dan Kekurangan Penelitian

Kelebihan

1. Naive Bayes mampu menangani dataset besar dengan waktu komputasi yang relatif singkat.
2. Algoritma ini bekerja baik dengan data berdimensi tinggi karena proses perhitungannya yang sederhana.
3. Dapat diterapkan pada berbagai jenis data, baik kategorikal maupun numerik.

Kekurangan

1. Asumsi ini sering tidak terpenuhi dalam data dunia nyata, sehingga dapat menurunkan akurasi prediksi.
2. Ketidakseimbangan dalam distribusi kelas dapat menyebabkan bias dalam hasil klasifikasi.
3. Tidak mempertimbangkan hubungan antar fitur secara mendalam [25]

2.7 Evaluasi Metode Naïve Bayes

Evaluasi metode Naive Bayes menggunakan confusion matrix adalah salah satu cara yang efektif untuk menilai kinerja model klasifikasi. Confusion matrix memberikan gambaran yang jelas tentang bagaimana model memprediksi kelas-kelas yang ada dalam dataset, dengan membandingkan prediksi model dengan label sebenarnya. Terdapat empat elemen utama dalam confusion matrix, yaitu True Positives (TP), False Positives (FP), True Negatives (TN), dan False Negatives (FN). Dari elemen-elemen ini, kita dapat menghitung berbagai metrik penting seperti akurasi, precision, recall, dan F1-score, yang masing-masing memberikan informasi berbeda tentang kinerja model dalam mengklasifikasikan

data. Penggunaan confusion matrix memudahkan untuk mengidentifikasi masalah pada model, seperti ketidakseimbangan kelas yang dapat memengaruhi akurasi dan interpretasi hasil.

Tabel . 2.2. Evaluasi

		Kelas Prediksi	
	Kelas	**Benar**	**Salah**
Kelas Atribut	Benar	True Positive (TP)	False Positive (FP)
	Salah	False Negative (FN)	True Negative (TN)

Dimana tabel 1 berisi:

1) TP (True Positive), yaitu jumlah data positif yang memiliki nilai benar.
2) TN (True Negative), yaitu jumlah data negatif yang memiliki nilai benar.
3) FN (False Negative), yaitu jumlah data negatif tetapi yang memiliki nilai salah.
4) FP (False Positive), yaitu jumlah data yang positif tetapi yang memiliki nilai salah.

$$\boldsymbol{Acuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FN+FP}} \times \mathbf{100\%} \quad [1]$$

$$\boldsymbol{Presisi} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \times \mathbf{100\%} \quad [2]$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \times \mathbf{100\%} \quad [3]$$

Evaluasi dengan menggunakan confusion matrix sangat relevan untuk menilai kinerja metode Naive Bayes dalam mengklasifikasikan tingkat kepuasan pengguna BPJS kesehatan. Dengan memanfaatkan confusion matrix, dapat dilihat seberapa baik model dalam memprediksi kategori kepuasan masyarakat, baik yang puas maupun yang tidak puas, berdasarkan data survei yang telah dikumpulkan. Hasil dari confusion matrix ini akan menunjukkan seberapa tepat

Naive Bayes dalam mengklasifikasikan data yang ada, serta mengungkapkan apakah model cenderung menghasilkan kesalahan prediksi pada salah satu kategori, seperti mengklasifikasikan orang yang puas sebagai tidak puas atau sebaliknya.