BAB II
LANDASAN TEORI

2.1. Konsep Data Mining

Data mining merupakan salah satu bidang dalam ilmu komputer yang berfokus pada proses menemukan pola, hubungan, dan pengetahuan baru dari kumpulan data dalam jumlah besar [1]. Konsep ini lahir dari kebutuhan untuk mengolah data yang terus bertambah seiring berkembangnya teknologi informasi, di mana data tidak lagi hanya disimpan, tetapi juga harus dimanfaatkan sebagai sumber informasi yang bernilai [2]. Melalui data mining, data mentah yang semula tidak terstruktur dapat diubah menjadi pengetahuan yang berguna untuk mendukung pengambilan keputusan. Proses ini sangat relevan di berbagai bidang, mulai dari bisnis, kesehatan, pendidikan, hingga pemerintahan, karena mampu mengungkap informasi tersembunyi yang tidak dapat dilihat hanya dengan analisis manual sederhana [3].

Secara umum, data mining sering disebut juga sebagai proses "Knowledge Discovery in Database" (KDD) [4]. Proses ini mencakup beberapa tahapan, mulai dari pengumpulan data, pembersihan data (data cleaning), transformasi, hingga tahap inti berupa penerapan metode atau algoritma tertentu untuk menemukan pola [5]. Setelah itu, hasil analisis perlu diinterpretasikan agar dapat dipahami dan dimanfaatkan secara praktis. Dalam praktiknya, data mining tidak hanya berhubungan dengan pengolahan data numerik, tetapi juga mampu mengolah data teks, gambar, hingga data multimedia lainnya. Dengan demikian, data mining

bukan hanya sekadar proses teknis, tetapi juga bagian dari strategi organisasi untuk memahami kondisi yang ada serta memprediksi tren di masa depan.

Metode dalam data mining beragam dan dapat diklasifikasikan menjadi beberapa kategori utama, yaitu klasifikasi, regresi, clustering, asosiasi, serta anomaly detection [6]. Klasifikasi digunakan untuk mengelompokkan data ke dalam kategori tertentu berdasarkan atribut yang ada, regresi bertujuan untuk memprediksi nilai numerik, clustering digunakan untuk mengelompokkan data berdasarkan kesamaan karakteristik, sementara asosiasi mencari hubungan antar item dalam data, seperti yang sering diterapkan pada analisis keranjang belanja [7]. Selain itu, anomaly detection digunakan untuk menemukan data yang menyimpang dari pola umum, misalnya dalam mendeteksi penipuan atau aktivitas tidak wajar. Setiap metode ini memiliki algoritma berbeda, seperti Naïve Bayes, Decision Tree, Apriori, K-Means, dan Neural Network, yang disesuaikan dengan tujuan penelitian atau kebutuhan analisis [8].

Manfaat utama dari penerapan data mining adalah memberikan wawasan yang lebih mendalam bagi pengguna dalam mengambil keputusan strategis [9]. Misalnya, dalam dunia bisnis, data mining dapat digunakan untuk memahami perilaku konsumen, meningkatkan layanan, hingga merancang strategi pemasaran yang lebih efektif [10]. Di bidang kesehatan, data mining dapat membantu dalam menganalisis rekam medis pasien untuk menemukan pola penyakit tertentu. Sementara itu, dalam pemerintahan, data mining dapat mendukung penyaluran bantuan sosial agar lebih tepat sasaran, seperti pada kasus klasifikasi penerima bantuan. Dengan keunggulan tersebut, data mining dipandang sebagai teknologi

penting dalam era big data, karena mampu mengubah data yang melimpah menjadi pengetahuan berharga yang bermanfaat bagi kemajuan organisasi maupun masyarakat.

2.2. Langkah-Langkah Data Mining

1. Pengumpulan Data

Langkah awal dalam proses data mining adalah pengumpulan data. Data yang dikumpulkan bisa berasal dari berbagai sumber, seperti basis data organisasi, arsip transaksi, hasil survei, data sensor, maupun data publik yang tersedia secara online. Tahap ini sangat penting karena kualitas data yang dikumpulkan akan menentukan hasil dari proses analisis. Jika data yang digunakan tidak lengkap, bias, atau tidak relevan, maka pola yang dihasilkan dari data mining juga berpotensi menyesatkan. Oleh sebab itu, pengumpulan data harus dilakukan dengan hati-hati, memperhatikan relevansi, kelengkapan, serta representasi data agar sesuai dengan tujuan penelitian.

Selain itu, pengumpulan data juga perlu memperhatikan aspek teknis dan non-teknis. Secara teknis, data harus dalam format yang dapat diolah, misalnya tabel, spreadsheet, atau database. Sedangkan dari sisi non-teknis, pengumpulan data harus memperhatikan etika, terutama jika data menyangkut informasi pribadi masyarakat. Contoh penerapan pada penelitian penerima bantuan BPJS adalah mengumpulkan data calon penerima berupa usia, pekerjaan, pendapatan, jumlah tanggungan, dan status rumah. Data inilah yang nantinya akan menjadi bahan utama untuk diolah pada tahap-tahap berikutnya sehingga menghasilkan sistem klasifikasi yang lebih objektif dan adil.

2. Pembersihan Data (Data Cleaning)

Setelah data terkumpul, tahap berikutnya adalah pembersihan data. Proses ini dilakukan untuk memastikan data yang akan dianalisis dalam kondisi bersih, bebas dari kesalahan, serta konsisten. Dalam praktiknya, data mentah sering kali mengandung berbagai masalah, seperti data yang hilang (missing value), duplikasi, kesalahan pengetikan, atau data yang tidak relevan. Jika masalah ini tidak diperbaiki, algoritma data mining bisa salah membaca pola sehingga menghasilkan klasifikasi yang tidak akurat. Oleh karena itu, tahap cleaning menjadi fondasi penting sebelum melangkah lebih jauh.

Pada penelitian penerima bantuan BPJS, misalnya, data tentang pendapatan mungkin ditulis dengan format yang berbeda (Rp 500,000.00, 500000, atau "500 ribu"). Perbedaan format ini harus distandarkan agar sistem bisa membaca dengan baik. Begitu juga dengan data status pekerjaan yang kosong atau tanggungan keluarga yang tidak diisi, harus diisi ulang atau ditangani sesuai metode tertentu, seperti imputasi. Dengan demikian, pembersihan data memastikan bahwa input yang masuk ke algoritma benar-benar berkualitas dan dapat mendukung hasil klasifikasi yang akurat.

3. Transformasi Data

Transformasi data merupakan tahap di mana data mentah yang sudah dibersihkan diubah ke dalam bentuk yang sesuai dengan kebutuhan algoritma [11]. Tidak semua algoritma dapat langsung memproses data dalam bentuk asli, sehingga data perlu dinormalisasi, dikategorikan, atau dikodekan ulang [12]. Misalnya, data

pendapatan yang awalnya berbentuk angka nominal dapat dikategorikan ke dalam kelompok tertentu, seperti “rendah”, “sedang”, dan “tinggi”. Begitu pula data pekerjaan yang awalnya berupa teks (seperti "buruh", "pedagang", "PNS") bisa diubah menjadi simbol atau angka agar mudah diproses oleh algoritma klasifikasi.

Transformasi juga mencakup proses reduksi dimensi jika data memiliki terlalu banyak variabel yang tidak relevan [13]. Dalam penelitian penerima bantuan BPJS, variabel yang dianggap penting misalnya usia, pendapatan, jumlah tanggungan, pekerjaan, dan status rumah [14]. Data ini kemudian ditransformasikan ke dalam bentuk tabel probabilitas agar sesuai dengan metode Naïve Bayes. Dengan adanya transformasi data, algoritma dapat bekerja lebih efisien, cepat, dan mampu menghasilkan pola yang lebih jelas [15].

4. Penerapan Algoritma Data Mining

Tahap inti dari proses data mining adalah penerapan algoritma. Pada tahap ini, data yang sudah siap diolah akan diproses menggunakan metode tertentu sesuai dengan tujuan analisis. Algoritma yang digunakan bisa berupa klasifikasi, clustering, asosiasi, regresi, atau anomaly detection. Misalnya, pada penelitian ini digunakan algoritma Naïve Bayes yang berfungsi untuk melakukan klasifikasi penerima bantuan BPJS berdasarkan variabel-variabel yang sudah ditentukan. Naïve Bayes dipilih karena sederhana, efisien, dan mampu memberikan hasil akurat meskipun data cukup bervariasi.

Dalam tahap ini, data training digunakan untuk membangun model, sedangkan data testing dipakai untuk mengevaluasi kinerja model tersebut. Misalnya, sistem akan mempelajari pola dari data training seperti hubungan antara

pendapatan rendah dengan status rumah menumpang dan label “dapat bantuan”. Setelah itu, pola tersebut diuji pada data testing untuk melihat seberapa akurat model dalam memprediksi kategori baru. Hasilnya bisa diukur melalui metrik evaluasi seperti akurasi, presisi, recall, dan F1-score.

5. Evaluasi dan Interpretasi

Setelah algoritma diterapkan, langkah berikutnya adalah evaluasi hasil. Evaluasi dilakukan untuk mengetahui seberapa baik model yang dibangun dapat mengklasifikasikan atau memprediksi data baru. Jika model memiliki akurasi yang tinggi dan seimbang antara presisi serta recall, maka model dianggap valid dan bisa digunakan. Namun, jika performanya rendah, maka perlu dilakukan perbaikan, baik dari sisi pembersihan data, transformasi, atau pemilihan algoritma yang lebih sesuai. Evaluasi juga memastikan bahwa hasil klasifikasi tidak hanya akurat secara matematis, tetapi juga relevan secara praktis sesuai konteks masalah.

Interpretasi hasil sangat penting agar temuan dari data mining dapat dipahami oleh pengambil keputusan. Misalnya, hasil menunjukkan bahwa mayoritas calon penerima dengan pendapatan di bawah Rp 850.000 dan jumlah tanggungan lebih dari tiga orang cenderung masuk kategori "Dapat Bantuan". Informasi ini bisa digunakan untuk merumuskan kebijakan lebih tepat sasaran. Dengan demikian, data mining tidak berhenti pada angka akurasi, tetapi juga menghasilkan wawasan yang dapat diimplementasikan dalam kehidupan nyata.

2.3. Metode Naïve Bayes

Metode Naïve Bayes adalah salah satu algoritma klasifikasi dalam data mining yang didasarkan pada Teorema Bayes dengan asumsi bahwa setiap atribut

atau fitur dalam data bersifat independen satu sama lain [16]. Artinya, keberadaan suatu atribut dalam satu kelas tidak dipengaruhi oleh atribut lainnya. Meskipun asumsi ini sering kali tidak sepenuhnya sesuai dengan kenyataan, algoritma Naïve Bayes tetap terbukti efektif dalam berbagai kasus klasifikasi karena kesederhanaan dan efisiensi yang dimilikinya [17]. Prinsip kerja dasarnya adalah menghitung probabilitas suatu data termasuk ke dalam kelas tertentu berdasarkan distribusi nilai fitur yang dimiliki, kemudian memilih kelas dengan probabilitas terbesar sebagai hasil prediksi [18].

Kelebihan utama dari metode ini adalah kecepatan dan kemudahannya dalam implementasi. Naïve Bayes dapat bekerja dengan baik pada data yang jumlahnya besar serta variabel yang bervariasi tanpa memerlukan komputasi yang kompleks [19]. Algoritma ini juga mampu menghasilkan hasil klasifikasi yang akurat meskipun data yang digunakan memiliki keterbatasan. Karena berbasis probabilitas, Naïve Bayes dapat menangani data yang bersifat kategorikal maupun numerik yang telah ditransformasikan [19]. Dengan sifatnya yang ringan, metode ini sering digunakan dalam aplikasi dunia nyata seperti klasifikasi teks, deteksi spam email, analisis sentimen, hingga prediksi dalam bidang kesehatan dan sosial.

Dalam penerapannya, Naïve Bayes bekerja dengan membangun model dari data training [20]. Model ini terbentuk dari perhitungan frekuensi atau probabilitas setiap atribut terhadap kelas yang ada. Misalnya, jika pada data penerima bantuan BPJS diketahui bahwa sebagian besar responden dengan penghasilan rendah dan jumlah tanggungan tinggi masuk ke dalam kategori “Dapat Bantuan”, maka probabilitas untuk kombinasi atribut tersebut terhadap kelas “Dapat Bantuan” akan

lebih besar. Ketika data testing dimasukkan, sistem akan menghitung probabilitas tiap kelas berdasarkan atribut yang ada, lalu menentukan kelas dengan nilai probabilitas tertinggi sebagai hasil klasifikasi.

Metode ini sangat cocok digunakan dalam penelitian penerima bantuan BPJS di Rantauprapat karena karakteristik datanya sederhana dan variabel yang digunakan relatif jelas, seperti usia, pekerjaan, pendapatan, jumlah tanggungan, dan status rumah. Dengan mengaplikasikan Naïve Bayes, proses klasifikasi penerima bantuan menjadi lebih objektif, terstruktur, dan transparan dibandingkan metode manual yang cenderung subjektif. Hasilnya tidak hanya memberikan akurasi tinggi, tetapi juga mudah dipahami dan ditafsirkan, sehingga dapat menjadi dasar dalam pengambilan keputusan yang lebih adil dan tepat sasaran.

2.3.1. Algoritma

Algoritma adalah serangkaian langkah logis dan sistematis yang dirancang untuk menyelesaikan suatu permasalahan atau mencapai tujuan tertentu [21]. Dalam konteks ilmu komputer, algoritma menjadi dasar dari semua proses komputasi karena ia menentukan bagaimana data diproses, dianalisis, dan diubah menjadi informasi yang bermanfaat [22]. Setiap algoritma harus memiliki karakteristik tertentu, seperti jelas, terstruktur, memiliki awal dan akhir, serta dapat dijalankan dalam jumlah langkah terbatas. Dengan demikian, algoritma tidak hanya digunakan dalam pemrograman komputer, tetapi juga dalam kehidupan sehari-hari, misalnya langkah-langkah membuat resep makanan atau prosedur kerja yang mengikuti aturan tertentu.

Peran algoritma dalam dunia teknologi sangatlah penting karena menjadi inti dari berbagai sistem aplikasi, perangkat lunak, dan kecerdasan buatan. Algoritma yang baik dapat meningkatkan efisiensi, mempercepat proses, dan meminimalisir kesalahan dalam pengolahan data. Dalam bidang data mining misalnya, algoritma digunakan untuk mengidentifikasi pola, membuat klasifikasi, hingga melakukan prediksi berdasarkan data yang tersedia. Setiap algoritma memiliki pendekatan yang berbeda, seperti algoritma Decision Tree yang berbasis pada struktur pohon keputusan, algoritma K-Means yang berfokus pada pengelompokan, atau algoritma Naïve Bayes yang mengandalkan probabilitas untuk klasifikasi.

Selain itu, algoritma juga menjadi fondasi dalam membangun sistem pendukung keputusan, terutama dalam penelitian seperti klasifikasi penerima bantuan BPJS. Dengan algoritma yang tepat, proses seleksi penerima bantuan bisa dilakukan lebih objektif dan transparan. Misalnya, algoritma Naïve Bayes digunakan karena mampu memberikan hasil klasifikasi dengan cepat dan akurat meskipun dengan data sederhana. Hal ini menunjukkan bahwa pemilihan algoritma yang sesuai sangat memengaruhi kualitas hasil yang diperoleh. Oleh karena itu, pemahaman mendalam tentang konsep, cara kerja, dan karakteristik algoritma menjadi kunci untuk menghasilkan solusi yang efektif dalam berbagai bidang ilmu pengetahuan maupun praktik nyata.

2.3.2. Alur atau Rumus Metode Naive Bayes

Naïve Bayes berlandaskan pada teorema Bayes, yaitu teori dasar dalam probabilitas yang menjelaskan bagaimana menghitung peluang suatu hipotesis berdasarkan bukti yang ada. Rumus utamanya adalah:

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

Di mana:

P(H|X) adalah probabilitas posterior, yaitu peluang hipotesis H benar jika diberikan bukti X.

P(X|H) adalah likelihood, yakni peluang bukti X muncul jika hipotesis H benar.

P(H) adalah prior, yaitu peluang awal dari hipotesis H sebelum ada bukti.

P(X) adalah evidence, yaitu total peluang terjadinya bukti X.

Dengan dasar ini, Naïve Bayes digunakan untuk menghitung seberapa besar kemungkinan suatu data termasuk ke dalam kategori tertentu berdasarkan bukti atau atribut yang dimiliki.

Dalam praktik klasifikasi, nilai P(X) biasanya sama untuk semua kelas, sehingga tidak terlalu berpengaruh dalam perbandingan antar kelas. Oleh karena itu, rumus Naïve Bayes sering disederhanakan menjadi:

$$P(H|X) \propto P(X|H) \cdot P(H)$$

Tanda "∝" artinya berbanding lurus. Jadi kita hanya perlu menghitung perkalian antara prior kelas P(H) dengan likelihood P(X|H). Proses ini membuat perhitungan menjadi lebih sederhana namun tetap akurat, karena yang kita cari hanyalah kelas dengan probabilitas paling tinggi. Dengan kata lain, data baru akan diklasifikasikan ke dalam kelas yang memberikan hasil perkalian terbesar antara prior dan likelihood.

Naïve Bayes juga mengasumsikan bahwa setiap atribut dalam data bersifat independen, artinya tidak saling memengaruhi. Inilah alasan metode ini disebut "Naïve" (naif). Dengan asumsi independensi, probabilitas bersyarat dapat dihitung dengan mengalikan probabilitas dari masing-masing atribut. Rumusnya adalah:

$$P(H|X_1, X_2, ..., X_n) \propto P(H) \cdot P(X_1|H) \cdot P(X_2|H) \cdot ... \cdot P(X_n|H)$$

Keterangan:

X1,X2,...,Xn adalah atribut data (misalnya pekerjaan, pendapatan, jumlah tanggungan, status rumah).

Masing-masing probabilitas dihitung berdasarkan data latih, kemudian dikalikan untuk menentukan kecenderungan suatu data masuk ke kelas tertentu.

Alur pertama dalam metode Naïve Bayes adalah pengolahan data latih (training data). Pada tahap ini, dataset yang sudah dipilih akan dianalisis untuk menghitung probabilitas prior dari setiap kelas. Misalnya, dari sekian banyak data penerima BPJS, dihitung berapa persentase yang termasuk ke dalam kategori "Dapat" dan berapa yang "Tidak Dapat". Nilai probabilitas awal ini penting karena akan menjadi dasar untuk mengukur kecenderungan suatu data baru masuk ke salah satu kelas. Tanpa adanya prior, proses klasifikasi akan kehilangan arah karena tidak ada titik awal untuk membandingkan antar kelas.

Tahap berikutnya adalah menghitung probabilitas bersyarat (likelihood) dari setiap atribut terhadap masing-masing kelas. Contohnya, dalam penelitian ini terdapat atribut seperti pekerjaan, pendapatan, jumlah tanggungan, dan status rumah. Untuk setiap atribut tersebut dihitung peluang kemunculannya di dalam kelas tertentu. Misalnya, berapa probabilitas seseorang dengan status pekerjaan

“Buruh” berada dalam kelas “Dapat” dan berapa yang masuk kelas “Tidak Dapat”. Dengan cara ini, setiap nilai atribut akan memiliki bobot probabilitas yang akan dikalikan pada tahap berikutnya.

Setelah semua probabilitas prior dan likelihood diperoleh, langkah selanjutnya adalah mengalikan probabilitas berdasarkan atribut yang dimiliki data uji. Proses perkalian ini sesuai dengan asumsi independensi antar fitur pada Naïve Bayes, sehingga seluruh atribut dapat dihitung secara terpisah lalu dikombinasikan. Hasil perkalian untuk setiap kelas menunjukkan seberapa besar kecenderungan data uji masuk ke kelas tersebut. Kelas dengan nilai probabilitas terbesar kemudian dipilih sebagai hasil akhir. Misalnya, jika data uji dengan pekerjaan “Petani” dan pendapatan Rp 800.000 menghasilkan probabilitas lebih tinggi pada kelas “Dapat”, maka data tersebut dikategorikan sebagai penerima bantuan.

Tahap terakhir dalam alur Naïve Bayes adalah evaluasi hasil klasifikasi. Setelah semua data uji diproses dan masing-masing diberi label sesuai hasil perhitungan, langkah berikutnya adalah membandingkan hasil prediksi dengan label sebenarnya. Dari sini dapat dihitung metrik evaluasi seperti akurasi, presisi, recall, atau F1-score. Evaluasi ini penting untuk memastikan sejauh mana metode Naïve Bayes mampu bekerja dengan baik dalam konteks penelitian. Semakin tinggi nilai akurasi yang diperoleh, maka semakin efektif algoritma ini dalam memprediksi kelayakan penerima bantuan BPJS.

2.3.3. Example Metode Naïve Bayes

Sebagai contoh penerapan metode Naïve Bayes dalam penelitian ini, misalnya terdapat data masyarakat dengan atribut: pekerjaan sebagai buruh,

pendapatan Rp 600.000 per bulan, jumlah tanggungan 3 orang, dan status rumah sewa. Berdasarkan data latih, masing-masing atribut memiliki probabilitas tertentu yang mengarah pada kategori "Dapat" atau "Tidak Dapat". Naïve Bayes akan menghitung peluang dari setiap kombinasi atribut tersebut terhadap masing-masing kelas. Proses ini diawali dengan menghitung prior probability untuk kelas "Dapat" dan "Tidak Dapat" dari distribusi data training yang ada.

Setelah prior probability dihitung, langkah berikutnya adalah mencari likelihood atau probabilitas kemunculan atribut tertentu dalam setiap kelas. Misalnya, kemungkinan seseorang yang berstatus buruh masuk ke kategori "Dapat" lebih tinggi dibandingkan dengan kategori "Tidak Dapat", berdasarkan data yang tersedia. Begitu juga dengan atribut pendapatan Rp 600.000 yang lebih sering muncul pada kelas "Dapat". Setiap probabilitas ini dihitung secara terpisah, lalu dikalikan untuk menghasilkan total probabilitas gabungan sesuai dengan prinsip independensi antar atribut pada Naïve Bayes.

Dari hasil perkalian probabilitas gabungan tersebut, Naïve Bayes kemudian membandingkan nilai likelihood antara kategori "Dapat" dan "Tidak Dapat". Kategori yang memiliki nilai probabilitas lebih besar akan menjadi hasil klasifikasi akhir. Misalnya, jika probabilitas total untuk kelas "Dapat" lebih tinggi dibandingkan dengan "Tidak Dapat", maka sistem akan mengklasifikasikan individu tersebut sebagai penerima bantuan BPJS. Proses ini menunjukkan bagaimana Naïve Bayes menggabungkan informasi dari berbagai atribut untuk menghasilkan prediksi yang terstruktur dan objektif.

Dengan menggunakan pendekatan ini, setiap data baru dapat diproses secara cepat dan efisien tanpa harus membangun model yang rumit. Contoh sederhana ini memperlihatkan bagaimana Naïve Bayes mampu bekerja dengan data sosial-ekonomi masyarakat dan tetap memberikan hasil klasifikasi yang akurat. Kelebihan lain dari metode ini adalah kemudahannya dalam interpretasi, karena setiap langkah perhitungan dapat ditelusuri dan dijelaskan secara jelas. Hal ini sangat bermanfaat dalam konteks penelitian maupun pengambilan keputusan, karena hasil yang diperoleh tidak hanya akurat tetapi juga transparan.

2.4. Alur Data Mining

1. Data Cleaning (Pembersihan Data)

Tahap pertama dalam alur Data Mining adalah data cleaning atau pembersihan data. Proses ini bertujuan untuk menghilangkan noise, data yang tidak konsisten, atau data yang hilang sehingga kualitas data lebih terjamin. Misalnya, pada dataset calon penerima BPJS, mungkin ada data yang tidak lengkap seperti pendapatan yang kosong atau status rumah yang salah input. Jika dibiarkan, data ini bisa menurunkan akurasi hasil klasifikasi. Oleh karena itu, pembersihan dilakukan dengan cara melengkapi data yang kosong, memperbaiki kesalahan format, atau menghapus data yang benar-benar tidak valid.

Selain memperbaiki kualitas data, pembersihan juga membantu mengurangi bias yang mungkin muncul saat proses klasifikasi. Misalnya, jika ada entri ganda (duplicate data), maka hasil probabilitas bisa menjadi tidak seimbang karena atribut tertentu akan muncul lebih banyak dari yang seharusnya. Hal ini bisa menimbulkan hasil prediksi yang salah. Dengan pembersihan, setiap entri akan mewakili kondisi

riil calon penerima bantuan sehingga hasil klasifikasi menjadi lebih akurat dan dapat dipertanggungjawabkan.

Dalam penelitian yang menggunakan Naïve Bayes, tahap cleaning sangat penting karena algoritma ini berbasis pada perhitungan probabilitas. Jika data mengandung error atau nilai ekstrim yang tidak wajar, maka hasil perhitungan peluang juga akan menyimpang. Oleh karena itu, tahap pembersihan merupakan pondasi utama agar alur data mining berjalan lancar dan menghasilkan sistem klasifikasi yang benar-benar sesuai dengan kondisi lapangan.

2. Data Integration dan Selection

Setelah data dibersihkan, alur selanjutnya adalah integrasi dan seleksi data. Integrasi berarti menggabungkan berbagai sumber data menjadi satu dataset yang utuh, sedangkan seleksi adalah memilih atribut atau variabel yang relevan untuk analisis. Dalam konteks penerima BPJS, data mungkin diperoleh dari hasil wawancara masyarakat, catatan administrasi desa, atau survei lapangan. Semua data ini kemudian digabungkan agar bisa dianalisis secara lebih komprehensif.

Namun, tidak semua atribut yang ada perlu digunakan. Proses seleksi data dilakukan untuk memilih atribut yang paling berpengaruh terhadap hasil klasifikasi. Misalnya, atribut seperti usia, pekerjaan, pendapatan, jumlah tanggungan, dan status rumah jelas relevan, sedangkan atribut lain yang tidak berkaitan langsung bisa diabaikan. Dengan demikian, dataset menjadi lebih ringkas dan fokus pada variabel yang benar-benar memiliki kontribusi besar dalam penentuan kelayakan bantuan.

Seleksi data juga mempermudah proses perhitungan Naïve Bayes karena semakin sedikit atribut yang digunakan, semakin ringan komputasi yang dibutuhkan. Selain itu, pemilihan atribut yang tepat dapat meningkatkan akurasi model karena hanya variabel penting saja yang dipertimbangkan. Proses ini juga membantu menghindari overfitting, yaitu kondisi di mana model terlalu mengikuti data latih sehingga tidak dapat bekerja dengan baik pada data uji.

3. Data Transformation

Tahap ketiga adalah transformasi data, yaitu mengubah data mentah ke dalam bentuk yang sesuai untuk dianalisis oleh algoritma. Transformasi dapat berupa normalisasi, pengelompokan kategori, atau pengkodean data. Misalnya, pendapatan masyarakat bisa dikelompokkan ke dalam kategori "rendah", "sedang", dan "tinggi" agar perhitungannya lebih sederhana. Begitu juga dengan pekerjaan, status rumah, dan jumlah tanggungan, semuanya perlu dikodekan dalam bentuk angka agar bisa diolah oleh sistem klasifikasi.

Transformasi ini penting karena tidak semua algoritma bisa bekerja langsung dengan data mentah, terutama jika data masih dalam bentuk teks atau memiliki format yang berbeda-beda. Dengan melakukan transformasi, semua atribut menjadi seragam sehingga dapat diproses oleh algoritma dengan lebih cepat dan efisien. Misalnya, status rumah yang memiliki nilai "Sewa", "Milik Sendiri", dan "Menumpang" bisa dikonversi menjadi angka 1, 2, dan 3. Proses ini tidak hanya mempermudah perhitungan, tetapi juga memastikan tidak ada kesalahan interpretasi dalam analisis data.

Selain itu, transformasi data juga memungkinkan dilakukannya reduksi dimensi, yaitu penyederhanaan data tanpa mengurangi informasi penting. Dengan cara ini, algoritma dapat bekerja lebih optimal karena hanya memproses informasi yang benar-benar dibutuhkan. Tahap ini memastikan data dalam kondisi terbaik sebelum masuk ke tahap inti data mining, yaitu proses mining atau analisis dengan algoritma tertentu.

4. Data Mining dan Evaluation

Tahap inti adalah data mining, yaitu penerapan metode atau algoritma untuk menemukan pola dalam data. Pada penelitian ini, algoritma yang digunakan adalah Naïve Bayes, yang bekerja berdasarkan prinsip probabilitas. Data latih digunakan untuk menghitung peluang setiap atribut terhadap kelas, lalu data uji diprediksi berdasarkan perhitungan probabilitas tersebut. Hasil dari tahap ini berupa klasifikasi apakah seseorang termasuk kategori “Dapat” atau “Tidak Dapat” menerima bantuan BPJS.

Namun, hasil klasifikasi tidak bisa langsung digunakan tanpa evaluasi. Oleh karena itu, tahap selanjutnya adalah evaluasi pola. Evaluasi dilakukan dengan cara membandingkan hasil prediksi dengan label aktual pada data uji. Dari perbandingan ini bisa dihitung akurasi, presisi, recall, atau F1-score. Evaluasi sangat penting karena menentukan sejauh mana metode yang digunakan benar-benar efektif dalam memprediksi penerima bantuan secara tepat sasaran.

Selain itu, evaluasi juga membantu dalam mengidentifikasi kelemahan model. Jika akurasi masih rendah, maka bisa dilakukan perbaikan pada tahap sebelumnya seperti menambah data latih, melakukan transformasi data yang lebih

baik, atau bahkan mencoba algoritma lain untuk perbandingan. Dengan evaluasi yang tepat, proses data mining tidak hanya menghasilkan pola semu, tetapi juga pengetahuan yang valid dan bermanfaat bagi pengambilan keputusan nyata di lapangan.

2.5. Peneliti Terdahulu

Berikut ada 10 penelitian terdahulu yang mana penelitian ini mengacu pada beberapa studi terdahulu seperti Wahid et al. (2023) dan Kurniadi et al. (2023), yang berhasil menerapkan algoritma Naive Bayes untuk klasifikasi penerima bantuan sosial dengan tingkat akurasi yang cukup tinggi. Penelitian-penelitian tersebut menjadi landasan dalam pengembangan model klasifikasi pada penelitian ini, dengan melakukan penyesuaian terhadap konteks lokal melalui penggunaan data sosial-ekonomi masyarakat di wilayah Rantauprapat Labuhanbatu. Adaptasi dilakukan pada pemilihan fitur yang relevan seperti pendapatan, jumlah tanggungan, dan status pekerjaan, serta tahapan preprocessing data yang disesuaikan dengan kondisi data lapangan, sehingga diharapkan dapat menghasilkan model klasifikasi yang lebih akurat dan tepat sasaran.

Tabel 2. 1. Penelitian Terdahulu Terkait Klasifikasi Penerima Bantuan Sosial

No.	Judul Penelitian	Tahun	Penulis	Hasil Utama
1	Klasifikasi Penerima Bantuan Langsung Tunai Menggunakan Naive Bayes dan SMOTE	2023	Kurniadi et al.	Akurasi mencapai **87,5%**, efektif dalam menangani data tidak seimbang.

2	Penerapan Data Mining untuk Seleksi Penerima Bantuan Sembako Menggunakan Naive Bayes	2021	Damuri et al.	Akurasi klasifikasi sebesar **85,2%** pada data sembako.
3	Naive Bayes untuk Klasifikasi Penerima Bantuan Sosial di Kota Malang	2023	Wahid et al.	Efisien dan akurat dengan hasil **82%**, cocok untuk data sosial kompleks.
4	Verifikasi Data Penerima Bantuan COVID-19 Menggunakan Naive Bayes	2022	Kamali et al.	Akurasi tinggi sebesar **85%**, mampu memverifikasi data secara cepat.
5	Klasifikasi Penerima Bantuan Non-Tunai Menggunakan Metode Naive Bayes	2023	Anam et al.	Nilai akurasi sebesar **80,5%**, efektif untuk data sosial ekonomi.
6	Penerapan Naive Bayes untuk Analisis Kelayakan Penerima Bantuan Rumah Layak Huni	2022	Siregar et al.	Dapat mengidentifikasi penerima bantuan dengan **akurasi 84,3%**.
7	Sistem Pendukung Keputusan Penerima PKH Menggunakan Algoritma Naive Bayes	2020	Nurmala et al.	Model cepat dan akurat dengan hasil klasifikasi mencapai **83,6%**.

8	Penerapan Data Mining dalam Seleksi Penerima BLT Dana Desa	2021	Yuliani & Hidayat	Kombinasi preprocessing dan Naive Bayes menghasilkan akurasi **86,7%**.
9	Klasifikasi Layak Tidak Layak Penerima Bantuan BPJS Berdasarkan Sosial Ekonomi	2024	Saputri & Harahap	Model Naive Bayes berhasil membedakan kategori layak dengan presisi **75%**.
10	Naive Bayes untuk Prediksi Bantuan Pendidikan di Kabupaten Sumedang	2022	Putri et al.	Efektif dalam klasifikasi penerima bantuan, akurasi **88,9%** dicapai.

2.6. Flowchart Sistem

Flowchart sistem adalah representasi visual yang digunakan untuk menggambarkan alur proses dalam sistem yang sedang diteliti, terutama dalam konteks klasifikasi data sosial menggunakan teknik machine learning. Dalam penelitian ini, flowchart menyajikan tahapan-tahapan yang harus dilakukan dalam menerapkan algoritma Naive Bayes pada klasifikasi penerima bantuan BPJS, mulai dari pengumpulan data hingga evaluasi dan validasi hasil.

2.6.1. Langkah-Langkah dalam Flowchart Sistem

1. Mulai

Tahap awal penelitian dimulai dengan merumuskan masalah yang akan diteliti. Identifikasi masalah ini bertujuan agar penelitian memiliki arah yang jelas serta fokus pada isu yang relevan. Dalam konteks penelitian klasifikasi penerima

bantuan BPJS di Rantauprapat Labuhanbatu, masalah yang dihadapi adalah masih adanya kesulitan dalam menentukan penerima bantuan secara objektif dan tepat sasaran. Dengan identifikasi masalah yang baik, penelitian tidak akan meluas ke arah yang kurang penting, melainkan tetap konsisten pada tujuan utama.

Selain identifikasi masalah, peneliti juga merumuskan tujuan penelitian yang ingin dicapai. Tujuan ini akan menjadi pedoman dalam setiap tahapan penelitian, mulai dari pengumpulan data hingga evaluasi model. Tanpa adanya tujuan yang terarah, penelitian akan sulit memberikan manfaat nyata dan hasilnya bisa tidak sesuai dengan kebutuhan di lapangan. Oleh karena itu, tahap ini sangat penting untuk memberikan kerangka awal penelitian.

Tahap mulai juga mencakup penyusunan perencanaan metode yang akan digunakan. Misalnya, pemilihan algoritma Naïve Bayes sebagai metode klasifikasi harus diputuskan sejak awal karena metode ini dianggap sesuai dengan karakteristik data sosial-ekonomi masyarakat. Dengan begitu, setiap langkah yang diambil dalam penelitian dapat terstruktur dan sistematis sesuai dengan metodologi yang telah ditentukan.

2. Pengumpulan Data Penerima Bantuan BPJS

Setelah tahap awal, langkah berikutnya adalah mengumpulkan data penerima bantuan BPJS. Data ini merupakan fondasi penelitian karena kualitas hasil klasifikasi akan sangat bergantung pada kualitas data yang diperoleh. Pengumpulan data dilakukan dengan memperhatikan keterwakilan populasi agar hasil penelitian dapat digeneralisasi dengan baik. Sumber data bisa berupa survei langsung kepada

masyarakat, wawancara dengan calon penerima, atau memanfaatkan data sekunder yang sudah ada.

Proses pengumpulan data harus dilakukan secara sistematis agar tidak menimbulkan bias. Misalnya, jika hanya mengandalkan satu sumber data saja, maka kemungkinan besar hasil penelitian tidak akan mencerminkan kondisi sebenarnya di lapangan. Oleh karena itu, kombinasi beberapa teknik pengumpulan data lebih dianjurkan agar hasilnya lebih akurat dan valid.

Selain itu, data yang dikumpulkan harus mencakup variabel-variabel penting seperti usia, pekerjaan, pendapatan bulanan, jumlah tanggungan, dan status kepemilikan rumah. Variabel-variabel inilah yang nantinya menjadi fitur dalam proses klasifikasi. Dengan memastikan data yang terkumpul lengkap dan relevan, peneliti akan lebih mudah melakukan tahap berikutnya, yaitu preprocessing data.

3. Preprocessing Data

Setelah data terkumpul, tahap selanjutnya adalah preprocessing atau pra-pemrosesan data. Tujuan dari preprocessing adalah untuk memastikan data dalam kondisi optimal sebelum digunakan dalam pelatihan model. Data mentah biasanya mengandung berbagai masalah seperti nilai yang hilang (missing values), data yang tidak konsisten, atau adanya noise. Jika masalah ini tidak ditangani, maka model klasifikasi bisa menghasilkan prediksi yang keliru.

Langkah-langkah dalam preprocessing antara lain membersihkan data dari kesalahan, melakukan normalisasi nilai, serta mengubah data kategorikal menjadi bentuk numerik agar dapat diolah oleh algoritma. Misalnya, status pekerjaan “PNS”, “Petani”, dan “Buruh” dapat dikonversi menjadi angka tertentu sehingga

lebih mudah diproses oleh Naïve Bayes. Proses transformasi ini juga membantu dalam menyederhanakan data tanpa menghilangkan informasi penting.

Preprocessing menjadi krusial karena kualitas model sangat bergantung pada kualitas data latih yang digunakan. Dengan preprocessing yang baik, model akan lebih mudah mendeteksi pola yang ada dalam data sehingga hasil klasifikasinya menjadi lebih akurat. Oleh karena itu, meskipun tahap ini tampak teknis, namun dampaknya sangat besar terhadap keberhasilan penelitian.

4. Pelatihan Model Naïve Bayes

Pada tahap ini, algoritma Naïve Bayes digunakan untuk melatih model klasifikasi berdasarkan data yang sudah melalui preprocessing. Naïve Bayes dipilih karena mampu menangani data kategorikal dengan baik, efisien, dan tidak membutuhkan komputasi yang rumit. Prinsip kerja algoritma ini adalah menghitung probabilitas dari setiap kelas berdasarkan fitur-fitur yang ada, lalu menentukan kelas dengan probabilitas tertinggi sebagai hasil prediksi.

Pelatihan model dilakukan dengan membagi data menjadi dua bagian, yaitu data latih (training data) dan data uji (testing data). Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengukur performa model. Dalam penelitian ini, model akan dilatih menggunakan data latih yang sudah bersih dan terstruktur sehingga dapat mendeteksi pola yang konsisten.

Proses pelatihan ini sangat penting karena menentukan seberapa baik model dapat mempelajari hubungan antara variabel-variabel seperti pekerjaan, pendapatan, jumlah tanggungan, dan status rumah dengan label penerima bantuan.

Hasil dari tahap ini berupa model probabilistik yang siap digunakan untuk memprediksi penerima bantuan pada data baru.

5. Pengujian Model

Tahap pengujian dilakukan setelah model selesai dilatih. Tujuannya adalah untuk menguji sejauh mana model mampu memprediksi data baru yang belum pernah dilihat sebelumnya. Pada tahap ini, data uji yang telah disiapkan digunakan untuk mengukur kinerja model. Proses ini penting agar peneliti mengetahui kemampuan model dalam melakukan generalisasi.

Pengujian juga memberikan gambaran awal tentang keandalan model. Jika model hanya mampu memberikan hasil yang baik pada data latih, namun buruk pada data uji, maka berarti model mengalami overfitting. Dengan melakukan pengujian, peneliti bisa memastikan bahwa model benar-benar mampu bekerja pada data nyata di lapangan, bukan hanya pada data yang digunakan untuk pelatihan.

Selain itu, pengujian juga membantu dalam menemukan kelemahan model. Misalnya, jika akurasi rendah, bisa jadi karena jumlah data latih masih kurang atau preprocessing belum dilakukan dengan baik. Dengan demikian, pengujian menjadi tahap krusial sebelum model digunakan lebih lanjut untuk evaluasi dan validasi.

6. Evaluasi Model

Setelah pengujian, tahap selanjutnya adalah evaluasi model. Evaluasi dilakukan dengan menggunakan metrik-metrik tertentu seperti akurasi, presisi, recall, dan F1-score. Masing-masing metrik memiliki fungsi tersendiri dalam menilai kinerja model. Akurasi mengukur seberapa banyak prediksi yang benar, presisi melihat ketepatan model dalam mengklasifikasi positif, recall melihat

kemampuan model menemukan semua data positif, sementara F1-score menggabungkan presisi dan recall dalam satu nilai.

Evaluasi sangat penting karena memberikan gambaran menyeluruh mengenai efektivitas model. Misalnya, meskipun akurasi tinggi, namun jika presisi atau recall rendah, maka model bisa dianggap kurang baik. Hal ini terutama penting dalam data sosial yang biasanya tidak seimbang antara kelas "Dapat" dan "Tidak Dapat".

Selain itu, evaluasi juga membantu peneliti menentukan langkah selanjutnya. Jika hasil evaluasi menunjukkan model bekerja baik, maka model bisa digunakan dalam aplikasi nyata. Namun, jika hasilnya masih belum memuaskan, peneliti bisa melakukan penyesuaian seperti memperbaiki preprocessing atau menambah jumlah data latih.

7. Analisis Hasil

Tahap analisis hasil dilakukan setelah evaluasi selesai. Pada tahap ini, peneliti menafsirkan hasil yang diperoleh untuk mengetahui sejauh mana model dapat diterapkan dalam konteks nyata. Analisis hasil bukan hanya melihat angka-angka metrik, tetapi juga memahami kekuatan dan kelemahan model secara menyeluruh.

Analisis ini juga membantu peneliti untuk menentukan apakah model benar-benar layak digunakan dalam skala yang lebih luas. Misalnya, jika akurasi tinggi tetapi model tidak bisa menjelaskan alasan di balik prediksi, maka hal itu bisa menjadi kelemahan yang perlu diperhatikan. Sebaliknya, jika model dapat bekerja dengan baik sekaligus mudah dipahami, maka ini menjadi nilai tambah.

Dengan analisis hasil, peneliti juga bisa memberikan masukan terkait kebijakan penentuan penerima bantuan. Model yang baik tidak hanya membantu

dalam prediksi, tetapi juga dapat digunakan sebagai alat pendukung keputusan yang objektif dan transparan.

8. Validasi Hasil

Setelah analisis, langkah berikutnya adalah validasi hasil. Validasi bertujuan untuk memastikan bahwa model yang dihasilkan benar-benar sesuai dengan kondisi di lapangan. Proses validasi bisa dilakukan dengan membandingkan hasil klasifikasi dengan data nyata penerima bantuan.

Validasi ini penting karena tanpa validasi, hasil penelitian tidak bisa dipertanggungjawabkan. Dengan validasi, peneliti dapat mengetahui apakah model dapat digunakan secara praktis atau masih memerlukan perbaikan. Selain itu, validasi juga memberikan kepercayaan kepada pengguna model bahwa hasil yang diberikan benar-benar bisa diandalkan.

9. Kesimpulan dan Rekomendasi

Tahap kesimpulan adalah langkah akhir dari penelitian ini. Kesimpulan merangkum temuan-temuan utama, baik dari proses pengumpulan data, preprocessing, pelatihan, pengujian, hingga evaluasi model. Dari kesimpulan ini, peneliti dapat mengetahui apakah tujuan awal penelitian telah tercapai atau tidak.

Selain kesimpulan, tahap ini juga menghasilkan rekomendasi untuk penelitian selanjutnya. Rekomendasi dapat berupa saran untuk menggunakan algoritma lain, menambah jumlah data, atau memperbaiki teknik preprocessing. Dengan adanya rekomendasi, penelitian tidak berhenti hanya pada satu titik, melainkan menjadi dasar untuk pengembangan lebih lanjut.

Kesimpulan dan rekomendasi juga penting karena memberikan kontribusi praktis. Bagi masyarakat atau pihak terkait, hasil penelitian ini bisa menjadi acuan dalam pengambilan keputusan terkait distribusi bantuan. Dengan begitu, penelitian benar-benar memberikan manfaat nyata bagi semua pihak.

10. Selesai

Tahap terakhir adalah menyatakan penelitian selesai. Namun, selesai di sini tidak berarti penelitian berhenti total. Justru dari penelitian ini, bisa lahir penelitian-penelitian lanjutan yang lebih mendalam. Misalnya, peneliti lain dapat menggunakan algoritma berbeda untuk membandingkan hasil atau memperluas cakupan penelitian ke wilayah lain.

Tahap selesai juga berarti seluruh proses, mulai dari perencanaan hingga validasi, telah dijalankan secara sistematis. Dengan demikian, hasil penelitian bisa dipublikasikan dan dipertanggungjawabkan secara ilmiah.

2.6.2. Tools Pendukung Penelitian

Dalam menunjang proses klasifikasi data sosial menggunakan algoritma Naive Bayes, penelitian ini menggunakan beberapa perangkat lunak dan tools pendukung untuk memastikan akurasi perhitungan, efisiensi pemrosesan, serta kemudahan dalam visualisasi dan analisis data. Adapun tools yang digunakan meliputi:

1. Microsoft Excel

Digunakan dalam tahap awal untuk proses input data, pengolahan awal, serta perhitungan manual seperti probabilitas dasar, distribusi frekuensi, dan pembuatan confusion matrix.

2. Rapid Miner

Merupakan platform data science yang digunakan untuk membangun, melatih, dan mengevaluasi model klasifikasi Naive Bayes. RapidMiner dipilih karena memiliki antarmuka visual yang mudah digunakan serta mendukung berbagai proses seperti preprocessing data, validasi model, dan perhitungan metrik evaluasi (akurasi, presisi, recall, dan F1-score).