BAB II
TINJAUAN PUSTAKA

2.1. Program Bedah Rumah

Program bedah rumah untuk memenuhi kebutuhan dasar terutama perumahan sebagai tempat tinggal, melalui peningkatan kesejahteraan sosial masyarakat dan perbaikan atau rehabilitasi rumah tidak layak huni (bedah rumah), sehingga keluarga miskin dapat menempati rumah yang layak huni dalam lingkungan yang sehat dan sejahtera. Dengan terpenuhinya salah satu kebutuhan dasar berupa rumah yang layak huni, diharapkan tercapai ketahanan keluarga. Rumah yang baik adalah rumah yang sehat atau sering disebut layak huni, yang harus diupayakan keberadaannya, kebutuhan rumah yang layak huni diharapkan sebagai upaya mencapai ketahanan keluarga, sebaliknya jika tidak terpenuhi akan menimbulkan permasalahan, seperti keterlantaran ataupun permasalahan kesejahteraan sosial keluarga[6]

2.2. Data Mining

Data Mining adalah kegiatan menemukan pola yang menarik dari data yang jumlahnya besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lain[7].

Data Mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Informasi yang dihasilkan dan diperoleh dengan cara mengekstraksi dan mengenali pola yang penting dan menarik dari data yang terdapat pada basis data. Data mining digunakan untuk mencari pengetahuan yang terdapat dalam basis data yang besar sehingga sering disebut Knowledge Discovery Database (KDD[4]

Data Mining adalah langkah analisis terhadap proses penemuan pengetahuan di dalam basis data atau Knowledge Discovery Database (KDD). Pengetahuan bisa berupa pola data atau relasi antar data yang valid yang artinya tidak diketahui sebelumnya. Data mining juga merupakan gabungan sejumlah disiplin ilmu komputer yang didefinisikan sebagai proses penemuan pada pola-pola baru dari

kumpulan data yang sangat besar, meliputi metode yang merupakan irisan dari artificial intelligence, machine learning, statistics dan database system. Data mining ditujukan untuk mengekstrak (mengambil intisari) pengetahuan dari sekumpulan data sehingga didapatkan struktur yang bisa dimengerti manusia serta meliputi basis data dan management data, sebelum pemprosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, setelah pemrosesan terhadap struktur yang ditemukan, visualisasi, dan online updating[8]

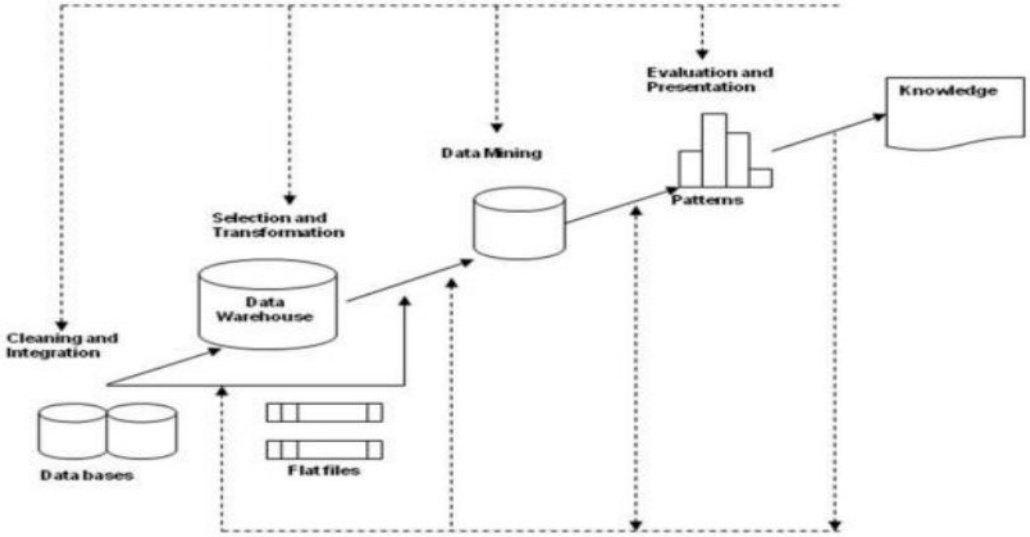
2.3. Proses Tahapan Data Mining

Data Mining merupakan inti dari proses Knowledge Discovery Database (KDD), yaitu proses untuk menggali dan menganalisis sejumlah data dan mengekstrak informasi dan pengetahuan yang berguna. Hasil pengetahuan yang diperoleh dalam proses tersebut dapat digunakan sebagai basis pengetahuan untuk keperluan pengambilan keputusan. Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahapan proses yang bersifat interaktif, yaitu pengguna terlibat langsung dengan perantara Knowledge Discovery Database (KDD). Tahapan - tahapan data mining adalah[9]

1. Pembersihan Data (Data Cleaning) Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning antara lain membuang duplikasi data, memeriksa data yang tidak konsisten, memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi), juga dilakukan proses anchriment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.
2. Integrasi Data (Data Integration) Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database

baru. Tidak jarang yang diperlukan untuk data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau file teks. Integrasi data dilakukan pada atribut-atribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya.

3. Seleksi Data (Data Selection) Pemilihan data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database
4. Transformasi Data (Data Transformation) Data yang diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum diaplikasikan.
5. Proses Mining Suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Beberapa metode yang dapat digunakan berdasarkan pengelompokan data mining.
6. Evaluasi Pola (Pattern Evaluation) Untuk mengidentifikasi pola-pola menarik ke dalam knowledge based yang ditemukan. Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang sudah tercapai.
7. Presentasi Pengetahuan (Knowledge Presentation) Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau 5 aksi dari hasil analisis yang telah didapat. Presentasi hasil data mining dalam bentuk pengetahuan yang bisa dipahami oleh semua orang adalah satu tahapan yang diperlukan dalam proses data mining. Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil dari data mining.



Gambar 2. 1 Tahapan Knowledge Discovery Database (KDD)[10]

2.4. Algoritma Naive Bayes.

Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi yang menggunakan metode probabilitas yang sederhana berdasarkan pada teorema bayes dengan asumsi tidak ketergantungan (independent) yang tinggi. Beberapa studi mengenai algoritma klasifikasi menunjukan bahwa naive bayes memiliki performa yang sebanding dengan decision tree dan neural network classifier tertentu. Selain itu, metode ini juga menunjukan akurasi dan kecepatan yang tinggi ketika digunakan dalam basis data yang berukuran besar[11]

Naive Bayes merupakan sebuah pengklasifikasian probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya, sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan naive dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi naive bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya[12]

Pengertian lain dari Naive Bayes yaitu sebuah klasifikasi probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan, algoritma menggunakan teorema bayes dan mengasumsikan semua atribut independent yang diberikan oleh nilai pada variabel kelas. Keuntungan penggunaan metode Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Data Training) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian, naive bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan[13]

Rumus *Naïve Bayes* nya adalah:

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \qquad (2.2)$$

Keterangan:

X = data dengan kelas yang belum diketahui

H = hipotesis data X, merupakan suatu kelas yang spesifik

P(H|X) = probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(H) = probabilitas hipotesis H (posteriori probability)

P(X|H) = probabilitas X berdasar kondisi H

P(X) = probabilitas dari X

atau

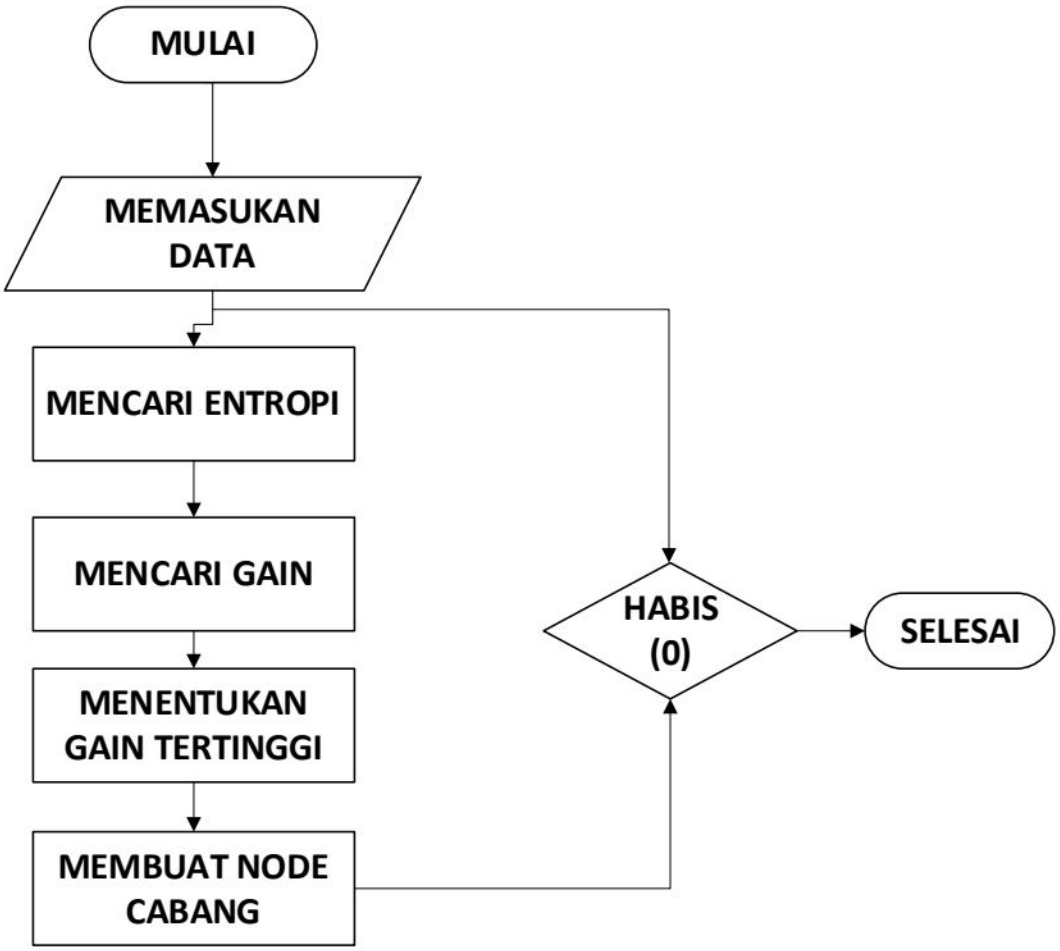
$$Posterior\ Probability = \frac{Prior\ Probability\ X\ likelihood}{evidance}$$

2.5. Algoritma C4.5

Algoritma C4.5 adalah sebuah algoritma yang berfungsi untuk membangun decision tree (pohon keputusan). Algoritma C4.5 dan pohon keputusan merupakan dua model yang tidak terpisahkan. Algoritma C4.5 adalah salah dari satu algoritma klasifikasi yang kuat dan banyak digunakan atau diimplementasikan untuk pengklasifikasian dalam berbagai hal. Algoritma C4.5 diperkenalkan oleh J. Ross

Quinlan (1996) sebagai versi perbaikan dari algoritma Iterative Dichotomiser 3 (ID3). Serangkaian perbaikan dilakukan pada algoritma ID3 mencapai puncaknya dengan menghasilkan sebuah sistem praktis dan simple yang berpengaruh untuk pembentukan pohon keputusan. Perbaikan tersebut meliputi metode untuk menangani data kontinu, mengatasi missing data, dan melakukan pemangkasan pohon [14]

Berikut adalah flowchart dari Algoritma C4.5 untuk membentuk sebuah pohon keputusan yang dapat dilihat pada Gambar 2.1



Gambar 2. 2 flowchart dari Algoritma C4.5

Pada Gambar 2.1 memasukan data yang telah dimasukkan ke beberapa atribut, kemudian melakukan perhitungan nilai entropy dan gain untuk mendapat gain tertinggi. Nilai tersebut yang akan menjadi atribut akar atau root dari pohon keputusan. Kemudian dalam proses pembuatan node cabang untuk masing – masing nilai. Jika setiap kasus dalam cabang tersebut telah berada di dalam satu kelas yang sama maka proses perhitungan sudah selesai, tapi jika kasus berbeda kelas maka kembali ke perhitungan entropy dan begitu seterusnya hingga semua kasus berada di dalam kelas yang sama. Dalam memilih satu atribut menjadi akar, dilakukan perhitungan nilai dari atribut yang ada. Nilai gain yang paling tinggi
dijadikan root di pohon keputusan. Untuk menghitung nilai gain rumus yang digunakan adalah [15] persamaan 2.1

$$Gaint\ (S.A\ ) = Entropy(S)\frac{Si}{S}*\text{ Entropy (Si)}$$

Keterangan:
S: Himpunan kasus
A: Data Atribut
n: Jumlah partisi di dalam atribut
|Si|: Jumlah kasus pada partisi ke-i
|S|: Jumlah kasus

Sedangkan untuk menghitung nilai entropy dapat dihitung dengan rumus [16], persamaan 2.2.

$$Entropy\ (S) = \sum_{pi}^{n} \text{-pi=log}_2\ \text{pi}$$

Keterangan:
S: Himpunan kasus
n: Jumlah partisi dalam atribut
pi: Proporsi dari Si terhadap S

Information gain adalah

salah satu attribute selection measure yang digunakan untuk memilih test attribute tiap node pada tree. Atribut dengan informasi gain tertinggi dipilih sebagai test atribut dari suatu node.

Dalam prosesnya perhitungan gain bisa terjadi missing value dan juga bisa tidak terjadi missing value. Rumus yang digunakannya juga berbeda.

Misalkan S berisi s data samples. Anggap atribut untuk class memiliki m nilai yang berbeda, (untuk i=1, I). Anggap menjadi jumlah samples S pada class. Maka besar informasinya dapat dihitung dengan:

$$1(S^1S^2 \ldots\ldots S_{\mathrm{m}} = -\sum_{j=1P1}^{m} * Log_2(P^1)$$

dimana adalah probabilitas dari sample yang mempunyai class C

Misalkan atribut A mempunyai v nilai yang berbeda, {…}. Atribut A dapat digunakan untuk mempartisi S menjadi v subset, {…}, dimana berisi samples pada S yang mempunyai nilai dari A. jika A terpilih menjadi test atribut (yaitu, best atribut untuk splitting), maka subset-subset akan berhubungan dengan pertumbuhan node-node cabang yang berisi S. Anggap sebagai jumlah samples class pada subset. Entropy, atau nilai information dari subset A adalah: [1]

$$E(a) = \sum_{f=1}^{y} \frac{S^1j^+ \ldots\ldots_+S_{\mathrm{m}}}{S} 1(S^1, S^2, \ldots S_{\mathrm{m}})$$

$$\frac{S^1j^+ \ldots\ldots_+S_{\mathrm{m}}}{S}$$

adalah bobot dari subset jth dan jumlah samples pada subset (yang mempunyai nilai dari A) dibagi dengan jumlah total samples pada S. Untuk subset,

$$1(S^{1J}, S^{2J} \dots S_m{}^J) = \sum_{J=1}^{m} p^i * \text{Log}_2(P^{iJ})$$

Dimana = adalah probabilitas sample yang mempunyai kelas Ci.
Maka nilai information gain atribut A pada subset S adalah Gain (A) = I …) – E(A)

2.6. Gain

Gain adalah Ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dari suatu variabel merupakan selisih antara nilai entropy total dengan *entropy* dari variabel tersebut. *Gain* dapat dirumuskan dengan: [1]

$$Gain\ (A) = \text{Entropy (S)} - entropy_A$$

Pada algoritma C4.5, nilai *gain* digunakan untuk menentukan variabel mana yang menjadi *node* dari suatu pohon keputusan. Suatu variabel yang memiliki *gain* tertinggi akan dijadikan *node* di pohon keputusan.

Split Info

Split info digunakan sebagai pembagi dari *Gain(*A) yang akan menghasilkan *Gain Ratio*.

$$SplitInfo_A\ (D) = \sum_{J=1}^{V} \binom{DJ}{D} \text{Log}_2 \binom{DJ}{D}$$

2.7. Gain Ratio

Gain Ratio merupakan salah satu ukuran lain yang digunakan untuk mengatasi masalah pada atribut yang memiliki nilai sangat bervariasi. *Gain Ratio* tertinggi dipilih sebagai atribut test untuk simpul. [1]

$$\text{GainRatio}\ \frac{Gain(A)}{SplitInfo_{A(D)}}$$

2.8. Pemilihan Variable

Pemilihan variabel yang juga disebut sebagai pemilihan atribut, digunakan pada *dataset* untuk menemukan pola yang penting dalam *data mining*. Pemilihan variabel digunakan untuk pengurangan dimensi pada *dataset*. Pemilihan variabel digunakan untuk melakukan eliminasi variabel yang tidak *relevan* dan *redundan*, yang dapat menyebabkan kebingungan dalam penggunaan variable.

Pemilihan variabel dapat mengurangi dimensi data, hal ini memungkinkan lebih efektif dalam operasi agar lebih cepat dari beberapa algoritma data mining. Dengan adanya pemilihan variabel membuat algoritma data mining lebih cepat dan lebih efektif.

Penggunaan pemilihan variabel pada *dataset* yang menggunakan variabel bebas dapat meningkatkan performa model. Pemilihan variabel juga merupakan proses yang cukup memakan biaya, dan juga bertentangan dalam asumsi awal, bahwa semua informasi diperlukan untuk mencapai akurasi yang maksimal.

Metode yang dapat digunakan untuk pemilihan variabel antara lain *Backward Elimination, Forward Selection*, *Genetic Algorithm*, dan yang lainnya. Metode-metode tersebut digunakan dalam penelitian *data mining* agar dapat menghasilkan variabel yang relevan dalam penelitian.

Pemilihan variabel dengan filter model ini lebih murah dalam komputasi karena tidak melibatkan induksi algoritma dalam prosesnya.

2.9. Aplikasi Rapid Miner

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blachardstown dan Ralf Klinkenberg dari rapidi.com dengan tampilan GUI (Graphical User Interface) yang dapat mempermudah pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat open source dan dibuat dengan menggunakan program java dibawah lisensi GNU Public Licence dan Rapid Miner dapat dijalankan pada sistem operasi manapun. Dengan menggunakan rapid miner, tidak dibutuhkan kemampuan coding khusus,

karena semua fasilitas kemampuan sudah tersedia. Rapid miner dikhususkan untuk penggunaan data mining, model yang telah disediakan juga cukup banyak dan lengkap, seperti Model Bayesian, Modelling, Tree Induction, Neural Network dan lain-lain[17]

Rapid Miner adalah salah satu software/platform perangkat lunak untuk melakukan pengolahan data dengan menggunakan metode-metode data mining, rapid miner akan mengekstrak pola-pola dari dataset yang besar dengan mengkombinasikan metode statistika, database dan kecerdasan buatan. Rapid Miner memudahkan pengguna dalam melakukan perhitungan data yang sangat banyak dengan menggunakan operator yang berfungsi untuk memodifikasi data. Data akan dihubungkan dengan node-node pada operator kemudian pengguna hanya tinggal menghubungkannya ke node hasil untuk diketahui hasilnya[18]

2.10. Pengertian Desa

Pengertian Desa menurut Undang-Undang Nomor 32 Tahun 2004 tentang pemerintahan Daerah (UU Pemda) adalah sebagai kesatuan masyarakat hukum yang memiliki batas wilayah berwenang untuk mengatur dan mengurus kepentingan 19 masyarakat setempat, berdasarkan asal-usul dan adat istiadat setempat yang diakui dan dihormati dalam sistem pemerintahan Negara Kesatuan Republik Indonesia

Desa merupakan sebuah komunitas kecil yang terikat, untuk mengatur kepentingan masyarakat baik sebagai tempat tinggal dan tempat pemenuhan kebutuhan hidup sesuai dengan kondisi dan sosial budaya. Pengertian dari masyarakat adalah sekelompok manusia yang saling berinteraksi sehingga dalam masyarakat tersebut terdapat sebuah kesepakatan yang telah ditentukan agar ditaati dan dilaksanakan oleh setiap anggota masyarakat [17].

Berdasarkan Peraturan Pemerintah Nomor 72 Tahun 2005 tentang Desa pada pasal 68 ayat 1 poin c, menyebutkan bahwa bagian dari dana perimbangan pusat dan daerah yang diterima oleh kabupaten/kota untuk desa, paling sedikit 10% secara proporsional pembagiannya untuk setiap desa, dana ini dalam bentuk Alokasi Dana

Desa atau sering disebut ADD. Alokasi Dana Desa (ADD) merupakan dana yang dialokasikan oleh pemerintah Kabupaten untuk Desa, yang bersumber dari bagian dana perimbangan keuangan pusat dan daerah yang diterima oleh Kabupaten. Alokasi Dana Desa merupakan dana yang cukup signifikan bagi Desa untuk menunjang program-program Desa, salah satunya adalah program bantuan bedah rumah

2.11. Program Bantuan bedah rumah

Bansos adalah pemberian bantuan stimulan berupa uang untuk pembelian bahan bangunan guna pemugaran Rumah Tidak Layak Huni dari pemerintah daerah kepada individu, keluarga, kelompok dan masyarakat yang sifatnya tidak secara terus menerus dan selektif yang bertujuan untuk melindungi dari kemungkinan terjadinya resiko sosial

2.12. Tinjauan Studi

Berikut adalah ringkasan dari beberapa penelitian sebelumnya yang terkait dengan klasifikasi *data mining*

2.13. Kajian Jurnal Pertama

“Klasifikasi Metode Naive Bayes Dalam Data Mining Untuk Menentukan Konsentrasi Siswa (Studi Kasus di MAS PAB 2 MEDAN)” dalam penelitian ini berisi tentang memanfaatkan data training untuk menghasilkan probabilitas setiap kriteria untuk class yang berbeda, sehingga nilai-nilai probabilitas dari kriteria tersebut dapat dioptimalkan untuk memprediksi konsentrasi siswa berdasarkan proses klasifikasi yang dilakukan dan berdasarkan data akademik siswa yang dijadikan data training, metode naive bayes berhasil mengklasifikasikan 109 data siswa dari 120 data yang diuji. Sehingga dengan demikian metode naive bayes ini

berhasil memprediksi konsentrasi siswa dengan persentase keakuratan sebesar 90,8333%

2.14. Kajian Jurnal Pertama

Aplikasi Klasifikasi Penerima Kartu Indonesia Sehat Menggunakan Algoritma Naive Bayes Classifier" dalam penelitian ini berisi tentang perlunya membangun sistem aplikasi klasifikasi penentuan penerima Kartu Indonesia Sehat pada Dinas Sosial Kabupaten Sukoharjo. Dan data yang digunakan sebagai bahan pertimbangan untuk penentuan penerima Kartu Indonesia Sehat adalah usia, pendidikan terakhir, pekerjaan, pendapatan perbulan dan tanggungan anak. Berdasarkan pengujian data testing sebanyak 13 kali percobaan dengan menghasilkan rata-rata nilai accuracy sebesar 94,78% precision 98,86% dan recall 90,98%

2.15. Kajian Jurnal Pertama

"Penentuan Kelayakan Penerima Bantuan Renovasi Rumah Warga Miskin Menggunakan Naive Bayes" membahas tentang tingkat akurasi perhitungan algoritma 6 naive bayes menggunakan tools WEKA dengan jumlah data yang digunakan adalah 50 data dengan 7 kriteria didapatkan hasil bahwa 90% algoritma naive bayes tepat digunakan untuk membantu dalam pengambilan keputusan seleksi penerima bantuan renovasi rumah, sedangkan 10% tidak dapat membantu dalam pengambilan keputusan