BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2. 1 Penelitian Terdahulu

Penulis memulai penelitian ini dengan terlebih dahulu melakukan studi kepustakaan dari penelitian-penelitian dan sumber-sumber lain. Dari studi kepustakaan itu penulis menemukan beberapa penelitian yang mendorong penulis untuk mengangkat tema seperti di atas. Penelitian tersebut membahas tentang topik yang terkait dengan penelitian penulis, antara lain adalah penelitian mengenai algoritma yang akan digunakan penulis.

Penelitian pertama yang berkaitan dengan laporan tugas akhir ini berjudul "Penerapan Algoritma Klasifikasi Data Mining ID3 untuk Menentukan Penstatus Siswa SMA N 6 Semarang" yang ditulis oleh Obbie Kristanto[7]. Penelitian tersebut menggunakan metode algoritma klasifikasi data mining Iterative Dichotomiser 3 atau yang lebih populer dengan sebutan ID3, ID3 membentuk pohon keputusan yang nantinya dapat digunakan sebagai dasar pertimbangan dalam menentukan penstatus di SMA, dan tools yang digunakan dalam pembuatan aplikasi tersebut adalah pemrograman Java. Kesimpulan dari penelitian tersebut adalah aplikasi telah berhasil dirancang sesuai dengan kebutuhan – kebutuhan yang menjadi tujuan dari perancangan, yaitu aplikasi dapat bekerja sebagai media pembantu dalam penstatus pada SMA N 6 Semarang dan mengetahui tingkat akurasi dengan menggunakan data dari guru BP sebagai perbandingan. Dengan menggunakan 371 dataset dan 20 data uji yang diinputkan terdapat 4 kasus yang meleset dan 16 kasus berhasil, sehingga didapat akurasi sebesar 80%.

Melalui pengujian – pengujian yang sudah dilakukan setelah aplikasi dapat diimplementasikan pada Java, semua pengujian input dan output aplikasi telah sesuai dengan yang diharapkan. Aplikasi telah berhasil memproses semua input yang diberikan dan menghasilkan output yang sesuai dengan

kebutuhan, sehingga aplikasi sudah layak untuk digunakan dan diterapkan sesuai dengan kebutuhan.

Penelitian kedua berjudul "Perbandingan Algoritma ID3 dan C5.0 dalam Indentifikasi Penstatus Siswa SMA" yang ditulis oleh Holisatul Munawaroh, Bain Khusnul K, S. T., M.Kom, Yeni Kustiyahningsih, S.Kom., M.Kom[1]. Penelitian tersebut menggunakan algoritma ID3 dan C5.0, yang kemudian dilakukan pembandingan terhadap kinerja dari kedua algoritma tersebut dalam melakukan identifikasi penstatus siswa SMA. Penelitian ini menggunakan 200 data siswa kelas X tahun ajaran 2011/2012, data tersebut dipecah menjadi 2 yaitu 150 data training dan 50 data testing. Kesimpulan dari penelitian tersebut adalah Hasil uji coba pengukuran kinerja kedua algoritma menggunakan 3 skenario yang telah dilakukan, dapat disimpulkan bahwa pada skenario 3 merupakan ujicoba paling efektif karena akurasi yang dihasilkan mencapai 95% pada algoritma C5.0 post pruning 100:100. Algoritma pohon keputusan yang terbaik adalah algoritma C5.0 karena memiliki kinerja (precision, recall, accuracy dan error rate) yang lebih baik dibandingkan algoritma ID3. Ini terlihat dari nilai akurasi C5.0 post pruning 100:100 sebesar 95% sedangkan untuk ID3 100:100 sebesar 93%. Hasil penilaian kinerja yang telah diketahui dapat disimpulkan juga bahwa semakin banyak data testing yang digunakan semakin tinggi tingkat akurasi yang dihasilkan. Ini terlihat dari hasil skenario 1 menggunakan 50 data testing algoritma ID3 sebesar 86% dan C5.0 post pruning sebesar 90%. Sedangkan menggunakan 100 data testing hasil kinerjanya meningkat pada algoritma ID3 sebesar 93% dan C5.0 post pruning sebesar 95%.

Penelitian ketiga berjudul "Penentuan Status Sekolah Menengah Atas dengan Algoritma Fuzzy C-Means" yang ditulis oleh Bahar[10]. Penelitian tersebut menggunakan metode *Fuzzy C-Means* dalam penentuan status di Sekolah Menengah Atas. Kesimpulan dari penelitian tersebut adalah sebagai berikut, dari hasil pengujian algoritma Fuzzy C-Means (FCM) dalam penentuan status di Sekolah Menengah Atas pada 81 sampel data siswa yang

diuji dalam penelitian ini, menunjukkan bahwa Algoritma FCM memiliki tingkat akurasi yang lebih tinggi (yaitu rata-rata 78,39%), jika dibandingkan dengan metode penentuan status secara manual yang selama ini dilakukan (hanya memiliki tingkat akurasi rata-rata 56,17 %). Dari data yang dilatih, diperoleh tiga kelompok berdasarkan nilai rata-rata mata pelajaran pejurusanan, yaitu :

a. Kelompok pertama, terdiri atas siswa yang memiliki nilai rata-rata mata pelajaran pejurusanan KHUSUS sekitar 72,0635; nilai rata-rata mata pelajaran pejurusanan TIDAK sekitar 76,3067; dan nilai rata-rata mata pelajaran pejurusanan Siap sekitar 71,5032.
b. Kelompok kedua, terdiri atas siswa yang memiliki nilai rata-rata mata pelajaran pejurusanan KHUSUS sekitar 73,5371; nilai rata-rata mata pelajaran pejurusanan TIDAK sekitar 74,7951; dan nilai rata-rata mata pelajaran pejurusanan Siap sekitar 79,7301.
c. Kelompok ketiga, terdiri atas siswa yang memiliki nilai rata-rata mata pelajaran pejurusanan KHUSUS sekitar 80,0742; nilai rata-rata mata pelajaran pejurusanan TIDAK sekitar 75,0224; dan nilai rata-rata mata pelajaran pejurusanan Siap sekitar 74,4123.
d. Proses klastering dalam penelitian ini dilakukan dengan menentukan jumlah klaster yang terbentuk diawal proses sesuai dengan jumlah kelompok (Status) yang diinginkan. Dengan demikian, tidak dapat dkhususstikan berapa sesungguhnya jumlah klaster ideal yang terbentuk dari data nilai siswa yang ada, sehingga akurasi hasil pengelompokkan tidak dapat terukur.

Penelitian keempat berjudul “Penerapan Algoritma Naive Bayes untuk Penentuan Status Turn-Over Pegawai” yang ditulis oleh Yeffriansjah Salim[11]. Penelitian tersebut bermaksud untuk mengukur akurasi, presisi, dari penentuan status turn-over pegawai menggunakan algoritma *naive bayes* pada PT. Rig Tenders Indonesia Banjarmasin. Kesimpulan dari penelitian tersebut adalah sebagai berikut, pengujian menggunakan perhitungan

manual Naive Bayes dengan melibatkan 807 data training dan 17 data uji yang diambil berdasarkan data pegawai yang berhenti bekerja pada bulan desember 2011 menghasilkan nilai akurasi 70,6%, nilai akurasi ini lebih rendah 2,83% darkhususda pengujian menggunakan 824 data dengan perhitungan rapid miner tanpa optimisasi PSO yakni 73,43%, sedangkan pengujian menggunakan 824 data dengan perhitungan rapid miner dengan optimisasi PSO menghasilkan nilai akurasi sebesar 76,08%, dimana nilai akurasi ini lebih tinggi sebesar 2,65% dibandingkan perhitungan tanpa menggunakan optimisasi PSO yang hanya mendapatkan tingkat akurasi sebesar 73,43%.

Penelitian kelima berjudul "Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa" yang ditulis oleh Arief Jananto[12]. Penelitian tersebut menggunakan teknik data mining khususnya klasifikasi untuk prediksi dengan algoritma naive bayes. Kesimpulan dari penelitian tersebut adalah sebagai berikut :

a. Lama masa studi atau dalam hal ini ketepatan masa studi setiap mahasiswa dapat diprediksi berdasarkan faktor-faktor yang berkaitan dengan latar belakang sekolah sebelumnya dan data akademik serta pribadi saat berada di perguruan tinggi.
b. Fungsi prediksi dengan memanfaatkan teknik data mining menggunakan algoritma naive bayes telah dapat dibuat dan digunakan untuk memprediksi (menentukan kelas) dari masa studi atau ketepatan masa studi dari mahasiswa dengan data training dan data testing yang telah diperoleh.
c. Tingkat kesalahan dari fungsi klasifikasi yang digunakan untuk prediksi masih berkisar pada 20% hingga 34% yang hal ini dimungkinkan dapat dipengaruhi oleh jumlah data training maupun testingyang digunakan serta tingkat konsisten data yang digunakan.

Penelitian keenam berjudul “Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah Asuransi” yang ditulis oleh Bustami[13]. Penelitian ini menggunakan algoritma Naive Bayes untuk menambang data nasabah sebuah perusahaan asuransi untuk mengetahui lancar, kurang lancar atau tidak lancarnya nasabah tersebut. Kesimpulan dari penelitian tersebut adalah sebagai berikut :

a. Sistem klasifikasi data nasabah ini digunakan untuk menampilkan informasi klasifikasi lancar, kurang lancar atau tidak lancarnya calon nasabah dalam membayar premi asuransi dengan menggunakan algoritma Naive Bayes.
b. Dengan adanya sistem ini maka mempermudah pihak asuransi dalam memperkirakan nasabah yang bergabung, sehingga perusahaan bisa mengambil keputusan untuk menerima atau menolak calon nasabah tersebut.
c. Algoritma Naive Bayes di dukung oleh ilmu Probabilistik dan ilmu statistika khususnya dalam penggunaan data petunjuk untuk mendukung keputusan pengklasifikasian. Pada algoritma Naive Bayes, semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain.
d. Variabel penentu yang digunakan dalam penelitian ini adalah jenis kelamin, usia, status, pekerjaan, penghasilan per tahun, masa pembayaran asuransi, dan cara pembayaran asuransi.

Tabel 1.1: Tabel Ringkasan Penelitian

No	Judul	Penulis	Tahun	Metode	Hasil
1	Penerapan algoritma klasifikasi data mining ID3 untuk menentukan penstatus siswa sman 6 semarang.	Obbie Kristanto	2014	ID3	Dengan menggunakan 371 dataset dan 20 data uji yang diinputkan terdapat 4 kasus yang meleset dan 16 kasus berhasil, sehingga didapat akurasi sebesar 80%.
2	Perbandingan algoritma ID3 dan c5.0 dalam identifikasi penstatus siswa sma	Holisatul Munawaroh dkk	2012	ID3 dan C5.0	Kedua algoritma memiliki tingkat akurasi yang baik seiring dengan bertambahnya data testing namun hasil algoritma C5.0 lebih akurat dibanding dengan ID3.
3	Penentuan status sekolah menengah atas dengan algoritma fuzzy c-means	Bahar	2011	Clustering Fuzzy C-Means	Dari hasil pengujian algoritma Fuzzy C-Means (FCM) dalam penentuan status di Sekolah Menengah Atas pada 81 sampel data siswa yang diuji dalam penelitian ini, menunjukkan bahwa Algoritma FCM memiliki tingkat akurasi yang lebih tinggi (yaitu rata-rata 78,39%), jika dibandingkan dengan metode penentuan status secara manual yang selama ini dilakukan (hanya memiliki tingkat akurasi rata-rata 56,17 %).
4	Penerapan Algoritma Naive Bayes untuk Penentuan Status Turn-Over Pegawai	Yeffriansjah Salim	2012	Bayesian Classification	Pengujian menggunakan perhitungan manual Naive Bayes dengan melibatkan 807 data training dan 17 data uji yang diambil

					berdasarkan data pegawai yang berhenti bekerja pada bulan desember 2011 menghasilkan nilai akurasi 70,6%, nilai akurasi ini lebih rendah 2,83% darkhususda pengujian menggunakan 824 data dengan perhitungan rapid miner tanpa optimisasi PSO yakni 73,43%,sedangkan pengujian menggunakan 824 data dengan perhitungan rapid miner dengan optimisasi PSO menghasilkan nilai akurasi sebesar 76,08%,dimana nilai akurasi ini lebih tinggi sebesar 2,65% dibandingkan perhitungan tanpa menggunakan optimisasi PSO yang hanya mendapatkan tingkat akurasi sebesar 73,43%
5	Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa	Arief Jananto	2013	Algoritma Naive Bayes	Dari hasil uji coba diperoleh tingkat kesalahan prediksi berkisar 20% sampai dengan 50% dengan data training dan testing yang diambil secara random. Namun rata-rata tingkat kesalahan berkisar 20 % hingga 34%. Tinggi rendahnya tingkat kesalahan dapat disebabkan oleh jumlah record data dan tingkat konsistensi dari data training

					yang dgunakan. Sedangkan hasil prediksi dari ketepatan lama studi dari mahasiswa angkatan 2008 adalah sebesar 254 mahasiswa diprediksi ”Tepat Waktu” dan sisanya yaitu 4 orang diprediksi ”Tidak Tepat Waktu”.
6	Penerapan Algoritma Naive Bayes untuk Mengklasifikasi Data Nasabah Asuransi	Bustami	2014	Algoritma Naive Bayes	Algoritma Naive Bayes bertujuan untuk melakukan klasifikasi data pada kelas tertentu, kemudian pola tersebut dapat digunakan untuk memperkirakan nasabah yang bergabung, sehingga perusahaan bisa mengambil keputusan menerima atau menolak calon nasabah tersebut.

Adapun yang menjadi perbedaan dari penelitian penulis dengan penelitian sebelumnya adalah sebagai berikut :

a. Penulis menggunakan datasheet yang diambil dari hasil ujian tryout yang di laksanakan oleh lembaga kursus SMK Negeri 1 Dukuturi
b. Datasheet yang penulis analisa ini difokuskan untuk menentukan kesiapan siswa dalam menghadapi ujian nasional dengan parameter yang digunakan meliputi nama, dan nilai tryout dari masing-masing mata pelajaran.
c. Metode yang digunakan untuk membuat sistem penstatus ini adalah algoritma klasifikasi *Naive Bayes*, dengan penggunaan Rapidminer sebagai piranti perangkat lunak yang berguna untuk melihat hasil akurasi dari algoritma yang digunakan terhadap datasheet yang sedang diteliti, kemudian digunakannya tools bantu Matlab sebagai piranti perangkat

lunak yang digunakan untuk mengolah datasheet dalam klasifikasi penstatus siswa menggunakan metode data mining.

2.1 Tinjauan Pustaka

2.1.1 Data Mining

Pengertian data mining, berdasarkan beberapa orang:

1. Data mining (penambangan data) adalah suatu proses untuk menemukan suatu pengetahuan atau informasi yang berguna dari data berskala besar. Sering juga disebut segabai bagian proses KDD (Knowledge Discovery in Databases).
   (Santosa, 2007) [19].
2. proses menemukan korelasi-korelasi penuh arti, pola-pola dan trend dengan penyaringan melalui sejumlah data yang besar pada tempat penyimpanan, dan menggunakan teknologi pengenalan pola seperti yang terdapat pada teknik-teknik di statistika dan matematika (Larose, 2005) [16].
3. Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, ataupenyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu–ilmu lain, seperti database system, data warehousing, statistik, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing (Han, et al., 2006) [20].
4. Data mining didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan haruspenuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar (Witten, et al.,2005)[5].

Karakteristik data mining sebagai berikut [21]:

a. Data mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
b. Data mining biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih percaya.
c. Data mining berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

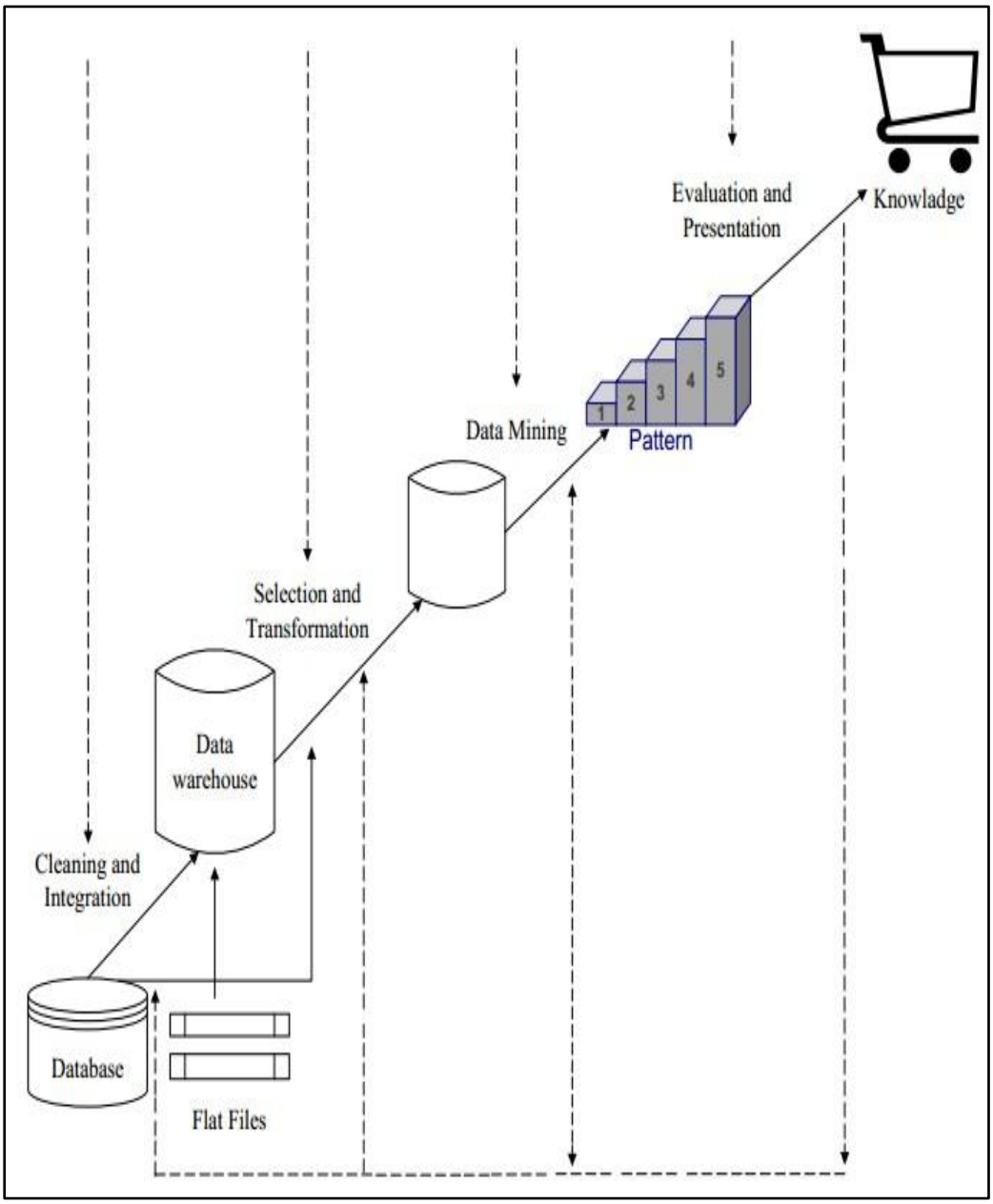
2.2.1.1 Tahap-tahap Data Mining

Salah satu tuntutan dari data mining ketika diterapkan pada data berskala besar adalah diperlukan metodologi sistematis tidak hanya ketika melakukan analisa saja tetapi juga ketika mempersiapkan data dan juga melakukan interpretasi dari hasilnya sehingga dapat menjadi aksi ataupun keputusan yang bermanfaat. Karenanya data mining seharusnya dkhusushami sebagai suatu proses, yang memiliki tahapan-tahapan tertentu dan juga ada umpan balik dari setiap tahapan ke tahapan sebelumnya. Pada umumnya proses data mining berjalan interaktif karena tidak jarang hasil data mining pada awalnya tidak sesuai dengan harapan analisnya sehingga perlu dilakukan desain ulang prosesnya.

Disini akan diuraikan tahap-tahap umum dari data mining tapi perlu diingat sebelum seorang analis menerapkan tahapan-tahapan data mining tersebut, sebagai prasyarat penerapan data mining, diperlukan pemahaman terhadap data dan proses diperolehnya data tersebut. Yang lebih mendasar lagi adalah diperlukannya pemahaman mengapa menerapkan data mining dan target yang ingin dicapai. Sehingga secara garis besar sudah ada hipotesa mengenai aksi-aksi yang dapat diterapkan dari hasilnya nanti. Pemahaman-pemahaman tersebut akan sangat membantu

dalam mendesain proses data mining dan juga pemilihan teknik data mining yang akan diterapkan[14].

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan knowledge base.



Gambar 2.1 : Tahapan Data Mining

Keterangan:

1. Pembersihan data

   Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan. Pada umumnya data yang diperoleh, baik dari database suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa data mining yang dimiliki. Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data juga akan mempengaruhi performasi dari teknik data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Integrasi data

   Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Tidak jarang data 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-aribut yang mengidentifikasikan entitasentitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebagai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada.

3. Seleksi data

   Data yang ada pada database sering kali tidak semuanya dkhususkai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database. Sebagai contoh,

sebuah kasus yang meneliti faktor kecenderungan orang membeli dalam kasus market basket analysis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

4. Transformasi data

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering disebut transformasi data.

5. Proses mining

   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

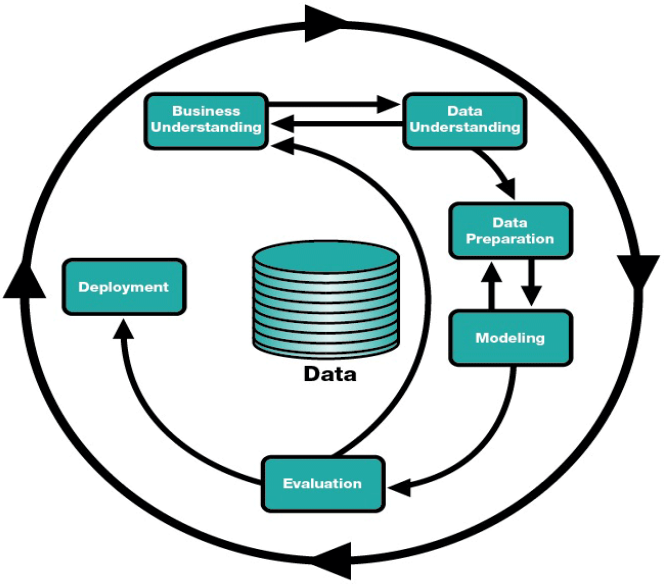
6. Presentasi pengetahuan

   Untuk mengidentifikasi pola-pola menarik kedalam 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 tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya umpan balik untuk memperbaiki proses data mining, mencoba metode data mining lain yang lebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaan yang mungkin bermanfaat[15].

2.2.2 CRISP-DM (Cross Industry Standart Process for Data Mining)

CRISP-DM (CRoss-Industry Standard Process for Data Mining) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai

proses standar dalam data mining yang dapat diaplikasikan di berbagai sektor industri. Berikut ini adalah gambar proses siklus hidup pengembangan dari CRISP-DM[16]:



Gambar 2.2 : CRISP-DM

Keterangan gambar :

1. Business Understanding

   Tahap pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemakan pengetahuan ini ke dalam pendefinisian masalah dalam data mining. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.

2. Data Understanding

   Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. Data Preparation

   Tahap ini meliputi semua kegiatan untuk membangun dataset akhir (data yang akan diproses pada tahap pemodelan/modeling) dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan tabel, record, dan atribut-atribut data, termasuh proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan (modeling).

4. Modeling

   Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal. Secara khusus, ada beberapa teknik berbeda yang dapat diterapkan untuk masalah data mining yang sama. Di pihak lain ada teknik pemodelan yang membutuhan format data khusus. Sehingga pada tahap ini masih memungkinan kembali ke tahap sebelumnya.

5. Evaluation

   Pada tahap ini, model sudah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisa data. Pada tahap ini akan dilakukan evaluasi terhadap keefektifan dan kualitas model sebelum digunakan dan menentukan apakah model dapat mencapat tujuan yang ditetapkan pada fase awal (Business Understanding). Kunci dari tahap ini adalah menentukan apakah ada masalah bisnis yang belum dipertimbangkan. Di akhir dari tahap ini harus ditentukan penggunaan hasil proses data mining.

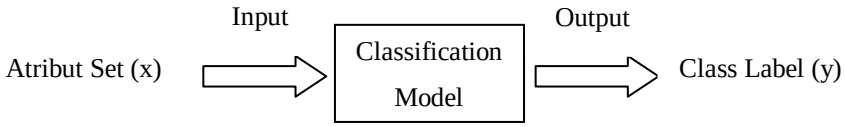
6. Deployment

   Pada tahap ini, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap deployment dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses data mining yang berulang dalam perusahaan. Dalam banyak kasus, tahap deployment melibatkan konsumen, di samping analis

data, karena sangat penting bagi konsumen untuk memahami tindakan apa yang harus dilakukan untuk menggunakan model yang telah dibuat.

2.2.3 Klassifikasi

Klasifikasi adalah proses 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. Dalam mencapai tujuan tersebut, proses klasifikasi membentuk suatu model yang mampu membedakan data kedalam kelas-kelas yang berbeda berdasarkan aturan atau fungsi tertentu. Model itu sendiri bisa berupa aturan "jika-maka", berupa pohon keputusan, atau formula matematis[13].



Gambar 2.3 : Blok Diagram Model Klasifikasi

2.2.4 Algoritma Naive Bayes

Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa 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[13].

Persamaan dari teorema Bayes adalah :

$P(H|X) = \frac{P\ (X|H).P(H)}{P(X)}$ ..... (1)

Keterangan :

X : Data dengan class yang belum diketahui

H : Hipotesis data *X* merupakan suatu class spesifik

P(H|X) :Probabilitas hipotesis *H* berdasar kondisi *X* (posteriori probability)

P(H) : Probabilitas hipotesis *H* (prior probability)

P(X|H) : Probabilitas *X* berdasarkan kondisi pada hipotesis *H*

P(X) : Probabilitas *X*

Untuk menjelaskan teorema Naive Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, teorema bayes di atas disesuaikan sebagai berikut :

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots . F_n|C)}{P(F_1 . \dots F_n)} \quad \text{..... (2)}$$

Dimana Variabel *C* merepresentasikan kelas, sementara variabel *F 1... Fn* merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas *C (Posterior)* adalah peluang munculnya kelas *C* (sebelum masuknya sampel tersebut, seringkali disebut prior), dikali dengan peluang kemunculan karakteristik – karakteristik sampel pada kelas *C* (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik – karakteristik sampel secara global (disebut juga *evidence*). Karena itu, rumus diatas dapat pula ditulis secara sederhana sebagai berikut :

Posterior = $\frac{Prior\ \ x\ likeli\ hood}{evidence}$

Nilai Evidence selalu tetap untuk setiap kelas pada satu sampel. Nilai dari posterior tersebut nantinya akan dibandingkan dengan nilai – nilai posterior kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus Bayes tersebut dilakukan dengan menjabarkan *(C|F1, ... , Fn)* menggunakan aturan perkalian sebagai berikut :

$$
\begin{aligned}
P(C|F_1,\dots,F_n) &= P(C)\,P(F_1,..,F_n|C) \\
&= P(C)P(F_1|C)P(F_2,\dots,F_n|C,F_1) \\
&= P(C)P(F_1|C)P(F_2|C,F_1)P(F_3,\dots,F_n|C,F_1,F_2) \\
&= P(C)P(F_1|C)P(F_2|C,F_1)P(F_3|C,F_1,F_2),P(F_4,\dots,F_n|C,F_1,F_2,F_3) \\
&= P(C)P(F_1|C)P(F_2|C,F_1)P(F_3|C,F_1,F_2)\dots P(F_n|C,F_1,F_2,F_3,..,F_{n-1})
\end{aligned}
$$

..... **(3)**

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor – faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi
independensi yang sangat tinggi (naif), bahwa masing – masing petunjuk (F1, F2 ... Fn) saling bebas (independen) satu sama lain. Dengan asumsi maka berlaku suatu kesamaan sebagai berikut :

$$P(P_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(\bar{F_j})}{P(F_j)} = P(F_i)$$

Untuk $i \neq j$, sehingga

$$P(F_i|C,F_j) = P(F_i|C) \qquad \text{..... (4)}$$

Dari persamaan diatas dapat disimpulkan bahwa asumsi independensi naif tersebut membuat syarat peluang menjadi

$$P(C|F_1,..,F_n) = P(C)P(F_1|C)P(F_2|C)P(F_3|C)\ldots$$
$$= P(C)\prod_{i=1}^{n} P(F_i|C) \quad \text{..... (5)}$$

sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran *P(C|F1, ... , Fn)* dapat disederhanakan menjadi :

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

$$P(X_i = x_i|Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2_{ij}}} \quad \text{..... (6)}$$

Keterangan :

P : Peluang

X : Atribut ke i

xi : Nilai atribut ke i

Y : Sub kelas Y yang dicari

yi : Sub kelas Y yang dicari

μ : Mean, menyatakan rata-rata dari seluruh atribut

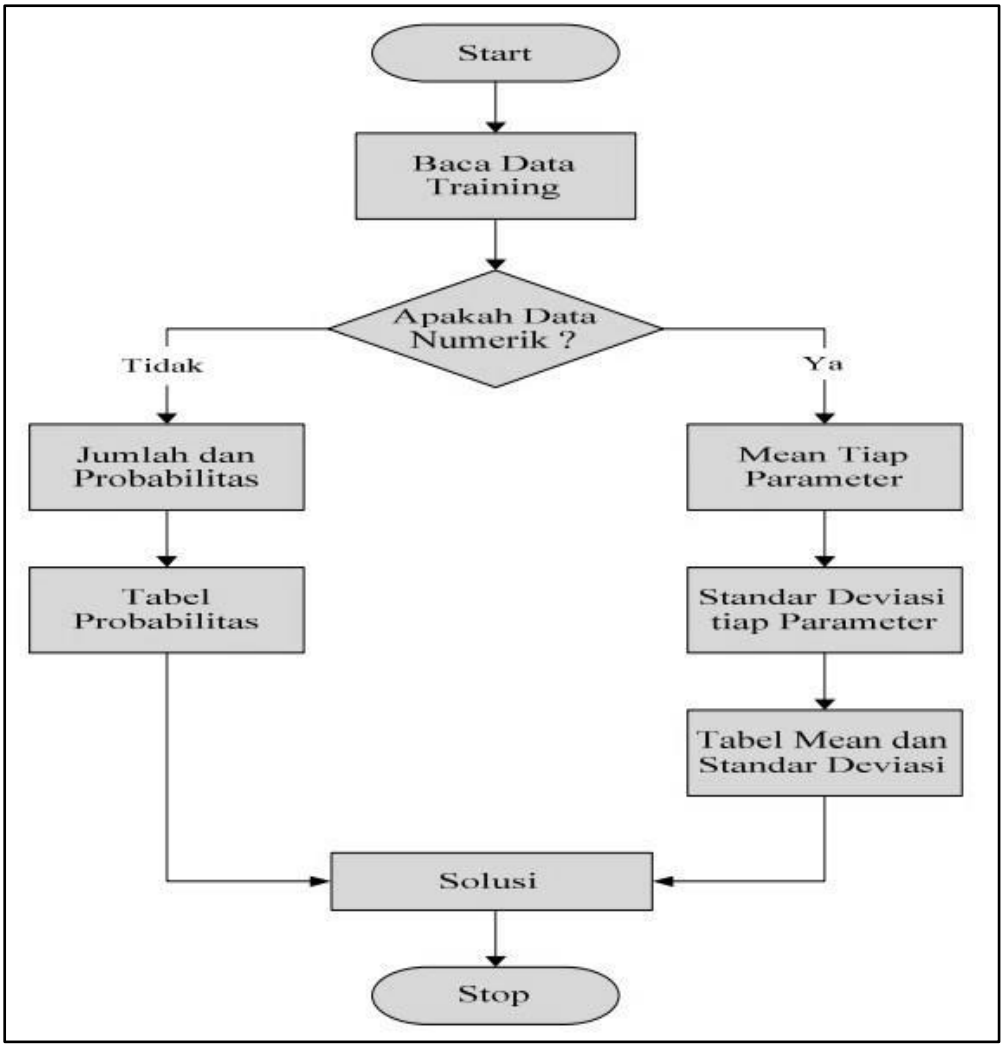
σ : Deviasi standar, menyatakan varian dari seluruh atribut

Adapun alur dari metode Naive Bayes adalah sebagai berikut :

1. Baca data training
2. Hitung Jumlah dan probabilitas, namun apabila data numerik maka:
   a. Cari nilai mean dan standar deviasi dari masing – masing parameter yang merupakan data numerik.

b. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.

3. Mendapatkan nilai dalam tabel mean, standart deviasi dan probabilitas.



Gambar 2.4 : Skema Naive Bayes

2.2.5 Pengujian *Cross Validation*

Validation adalah proses untuk mengevaluasi keakuratan prediksi dari model. Validasi digunakan untuk memperoleh prediksi menggunakan model yang ada dan kemudian membandingkan hasil tersebut dengan hasil yang sudah diketahui, ini mewakili langkah paling penting dalam proses membangun sebuah model[10].

Cross Validation adalah teknik validasi dengan membagi data secara acak ke dalam *k* bagian dan masing-masing bagian akan dilakukan proses klasifikasi. Dalam *Cross Validation*, jumlah tetap lkhususan atau partisi dari data ditentukan sendiri. Cara standar untuk memprediksi *error rate* dari teknik pembelajaran dari sebuah sampel data tetap adalah dengan menggunakan *tenfold cross validation.*

Dengan *tenfold cross validation,* data akan dibagi secara acak menjadi 10 bagian, dimana *class* diwakili (kurang lebih) proporsi yang sama seperti pada dataset yang penuh. Setiap bagian mendapatkan gilirannya dan skema pembelajaran dilatih pada sisa sembilan persepuluh; kemudian *error rate* dihitung pada *holdout set*. Dengan demikian, prosedur pembelajaran dilaksanakan sebanyak 10 kali di *training set* yang berbeda (setiap set memiliki banyak kesamaan dengan yang lain). Akhirnya, 10 estimasi error dirata-rata untuk menghasilkan perkiraan kesalahan keseluruhan.

2.2.6 Evaluasi dan Validasi Klasifikasi Data Mining

Untuk melakukan evaluasi pada algoritma *naïve bayes* maka diilakukan beberapa pengujian menggunakan *confusion matrix*.

2.2.6.1 Confusion Matrix

Confusion matrix memberikan keputusan yang diperoleh dalam *traning* dan *testing*, *confusion matrix* memberikan penilaian *performance* klasifikasi berdasarkan objek dengan benar atau salah. *Confusion matrix* berisi informasi aktual (*actual*) dan prediksi (*predicted)* pada sistem klasifikasi.

Table 2.1: Tabel Confusion Matrix untuk 2 Kelas

Classification	*Predicted Class*		
Observed Class		*Class = Yes*	*Class = No*
	Class = Yes	*A (true positif – tp)*	*B (false negative – fn)*
	Class = No	*C (false positif – fp)*	*D (true negative – tn)*

Keterangan:

a. *True Positive* (tp) = proporsi positif dalam data set yang diklasifikasikan positif.
b. *True Negative* (tn) = proporsi negative dalam data set yang diklasifikasikan negative.
c. *False Positive* (fp) = proporsi negatif dalam data set yang diklasifikasikan potitif.
d. *FalseNegative*(fn) = proporsi negative dalam data set yang diklasifikasikan negative

Tabel 2.2 : Tabel Confusion Matrix untuk 3 Kelas

		Predicted Class		
		Class 1	Class 2	Class 3
Actual Class	Class 1	Count11	Count12	Count13
	Class 2	Count21	Count22	Count23
	Class 3	Count31	Count32	Count33

Berikut adalah persamaan model *confusion matrix* untuk 3 kelas*:*

a. Nilai akurasi

$$\text{Accuracy} = \frac{Count11+Count22+Count33}{Count11+Count12+Count13+Count\ 21+Count22+Count23+Count31+Count32+Count33}$$

b. Error rate

$$\text{Error Rate} = \frac{Count12+Count13+Count21+Count23+Count31+Count32}{Count11+Count12+Count13+Count\ 21+Count22+Count23+Count31+Count32+Count33}$$

2.3 RapidMiner

RapidMiner merupakan perangakat lunak yang bersifat terbuka (open source). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan munggunakan siap java sehingga dapat bekerja di semua sistem operasi.

RapidMiner sebelumnya bernama YALE (Yet Another Learning Environment), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund. RapidMiner didistribusikan di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan mengunakan RapidMiner di lebih dari 40 negara. RapidMiner sebagai software open source untuk data mining tidak perlu diragukan lagi karena software ini sudah terkemuka di dunia. RapidMiner menempati peringkat pertama sebagai Software data mining pada polling oleh KDnuggets, sebuah portal data-mining pada 2010-2011.

RapidMiner menyediakan GUI (Graphic User Interface) untuk merancang sebuah pipeline analitis. GUI ini akan menghasilkan file XML (Extensible Markup Language) yang mendefenisikan proses analitis keingginan pengguna untuk

diterpkan ke data. File ini kemudian dibaca oleh RapidMiner untuk menjalankan analis secara otomatis.

RapidMiner memiliki beberapa sifat sebagai berikut:

a. Ditulis dengan siap pemrograman Java sehingga dapat dijalankan di berbagai sistem operasi.
b. Proses penemuan pengetahuan dimodelkan sebagai operator trees.
c. Representasi XML internal untuk memastikan format standar pertukaran data.
d. Siap scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
e. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
f. Memiliki GUI, command line mode, dan Java API yang dapat dkhususnggil dari program lain.

Beberapa Fitur dari RapidMiner, antara lain:

a. Banyaknya algoritma data mining, seperti decision treee dan self-organization map.
b. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots.
c. Banyaknya variasi plu gin, seperti text plugin untuk melakukan analisis teks.
d. Menyediakan prosedur data mining dan machine learning termasuk: ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi
e. Proses data mining tersusun atas operator-operator yang nestable, dideskrtidakikan dengan XML, dan dibuat dengan GUI
f. Mengintegrasikan proyek data mining Weka dan statistika R[17].

2.4 Matlab

Matlab merupakan siap canggih untuk komputasi teknik. Matlab merupakan integrasi dari komputasi, visualisasi dan pemograman dalam suatu lingkungan yang mudah digunakan, karena permasalahan dan pemecahannya dinyatakan dalam notasi matematika biasa. Kegunaan Matlab secara umum adalah untuk :

a. Matematika dan komputasi
b. Pengembangan dan algoritma
c. Pemodelan, simulasi dan pembuatan prototype
d. Analisa data, eksplorasi dan visualisasi
e. Pembuatan aplikasi termasuk pembuatan *graphical user interface*

Matlab adalah sistem interaktif dengan elemen dasar array yang merupakan basis datanya. Array tersebut tidak perlu dinyatakan khusus seperti di siap pemograman yang ada sekarang. Matlab merupakan sekumpulan fungsi – fungsi yang dapat dkhususnggil dan dieksekusi. Fungsi – fungsi tersebut dibagi – bagi berdasarkan kegunaannya yang dikelompokan didalam toolbox yang ada pada matlab[18].

2.5 Contoh Kasus Penerapan Algortima Naive Bayes

Model statistik merupakan salah satu model yang efisien sebagai pendukung pengambilan keputusan. Konsep probabilistik merupakan salah satu bentuk model statistik. Salah satu metode yang menggunakan konsep probabilistik adalah Naive Bayes. Algoritma Naive Bayes adalah salah satu algoritma dalam teknik klasifikasi yang mudah diimplementasikan dan cepat prosesnya. Pada metode ini, semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atibut yang sama penting dan setiap atribut saling bebas satu sama lain. Apabila diberikan k atribut yang saling bebas (independence), nilai probabilitas dapat diberikan sebagai berikut :

$P\ (x_1\ ,...,\ x_k \mid C) = P\ (x_1 \mid C)\ x\ ...\ xP\ (x_k \mid C)$

Tahap awal cara kerja dari proses perhitungan Naive Bayes adalah dengan melakukan pengambilan data training dari data nasabah asuransi. Adapun variabel penentu yang digunakan dalam mengklasifikasikan data nasabah yaitu[13]:

a. Jenis Kelamin

   Merupakan variabel jenis kelamin nasabah yang dikelompokkan dalam dua kategori yaitu laki – laki dan perempuan.

b. Usia

   Merupakan variabel usia nasabah yang di kelompokkan dalam tiga kategori yaitu 20-29 tahun, 30-40 tahun, dan diatas 40 tahun.

c. Status

   Merupakan variabel status nasabah yang dikelompokkan dalam dua kategori yaitu kawin dan belum kawin.

d. Pekerjaan

   Merupakan variabel pekerjaan nasabah yang dikelompokkan dalam tiga kategori yaitu PNS, Pegawai Swasta, Wiraswasta.

e. Penghasilan

   Merupakan variabel penghasilan dari nasabah yang dikelompokkan dalam tiga kategori yaitu 0-25 juta, 25-50 juta, dan diatas 50 juta.

f. Cara pembayaran premi

   Merupakan variabel cara pembayaran premi yang dikelompokkan dalam empat kategori yaitu bulanan, triwulan, semesteran, dan tahunan.

g. Masa pembayaran premi

   Merupakan variabel masa pembayaran premi yang dikelompokkan dalam tiga kategori yaitu 5 – 10 tahun, 11 – 15 tahun, dan diatas 15 tahun.

Tabel 2.3: Data Pelatihan

No	Nama	Jenis Kelamin	Usia	Status	Pekerjaan	Penghasilan	Masa Asuransi	Cara Pembayaran	Klasifikasi
1	Dani Lukman	Laki-Laki	30 - 40 Tahun	Kawin	Pns	<25 Juta	>15 Tahun	Tahunan	Tidak Lancar
2	Evaliana	Perempuan	30 - 40 Tahun	Kawin	Pns	<25 Juta	5 – 10 Tahun	Semesteran	Lancar
3	Rasyidah	Perempuan	20 - 29 Tahun	Kawin	Pegawai Swasta	<25 Juta	5 – 10 Tahun	Triwulan	Tidak Lancar
4	Dina Saufika	Perempuan	30 - 40 Tahun	Belum Kawin	Pns	<25 Juta	5 – 10 Tahun	Triwulan	Lancar
5	Wilsa Rizki	Laki-Laki	30 - 40 Tahun	Kawin	Wiraswasta	<25 Juta	5 – 10 Tahun	Tahunan	Kurang Lancar
6	Irwanto	Laki-Laki	30 - 40 Tahun	Belum Kawin	Wiraswasta	>50 Juta	11 – 15 Tahun	Semesteran	Lancar
7	Ade Gunawan	Laki-Laki	30-40 Tahun	Kawin	Pns	25 - 50 Juta	11 – 15 Tahun	Semesteran	Tidak Lancar
8	Fauziah	Perempuan	20 - 29 Tahun	Kawin	Wiraswasta	25 - 50 Juta	11 – 15 Tahun	Tahunan	Lancar
9	Zulaikha	Perempuan	20 - 29 Tahun	Kawin	Wiraswasta	<25 Juta	11 – 15 Tahun	Triwulan	Tidak Lancar
10	Zulfahmi	Laki-Laki	20 - 29 Tahun	Kawin	Pns	<25 Juta	11 – 15 Tahun	Triwulan	Kurang Lancar
11	Hidayatullah	Laki-Laki	30 - 40 Tahun	Belum Kawin	Wiraswasta	25 - 50 Juta	11 – 15 Tahun	Tahunan	Lancar
12	Nilam Sari	Perempuan	30 - 40 Tahun	Kawin	Wiraswasta	25 - 50 Juta	>15 Tahun	Tahunan	Kurang Lancar
13	Nahari Arifin	Laki-Laki	30 - 40 Tahun	Kawin	Wiraswasta	>50 Juta	11 – 15 Tahun	Triwulan	Lancar
14	Yusnidar	Perempuan	>40 Tahun	Kawin	Pns	<25 Juta	>15 Tahun	Semesteran	Kurang Lancar
15	Rizwan Hadi	Laki-Laki	20-29 Tahun	Belum Kawin	Pns	<25 Juta	11 – 15 Tahun	Tahunan	Lancar
16	Rahmat Saputra	Laki-Laki	30 -40 Tahun	Belum Kawin	Wiraswasta	<25 Juta	11 – 15 Tahun	Semesteran	Lancar
17	M. Sahril	Laki-Laki	>40 Tahun	Kawin	Pegawai swasta	<25 Juta	11 – 15 Tahun	Tahunan	Tidak Lancar
18	M. Irfan	Laki-Laki	30 - 40 Tahun	Kawin	Pegawai swasta	25 - 50 Juta	11 – 15 Tahun	Tahunan	Tidak Lancar
19	Tutri Wulandari	Perempuan	30 - 40 Tahun	Kawin	Wiraswasta	<25 Juta	11 – 15 Tahun	Triwulan	Lancar
20	Leni Syamsiah	Perempuan	20 - 29 Tahun	Belum Kawin	Wiraswasta	25 - 50 Juta	5 – 10 Tahun	Bulanan	Tidak Lancar
21	Syafi Arkan	Laki-Laki	30 - 40 Tahun	Kawin	wiraswasta	25 - 50 Juta	11 – 15 Tahun	Semesteran	????

Berdasarkan tabel diatas dapat dihitung klasifikasi data nasabah apabila diberikan input berupa jenis kelamin, usia, status, pekerjaan, penghasilan/tahun, masa asuransi dan cara pembayaran menggunakan algoritma Naive Bayes. Apabila diberikan input baru, maka klasifikasi data nasabah asuransi dapat ditentukan melalui langkah berikut :

1. Menghitung jumlah class/label.

P (Y=Lancar) = 9/20 “Jumlah data lancar pada data pelatihan dibagi dengan jumlah keseluruhan data”
P (Y=Kurang Lancar) = 4/20 “Jumlah data kurang lancar pada data pelatihan dibagi dengan jumlah keseluruhan data”
P (Y= Tidak Lancar) = 7/20 “Jumlah tidak lancar pada data pelatihan dibagi dengan jumlah keseluruhan data”

2. Menghitung jumlah kasus yang sama dengan class yang sama.
   P(Jenis Kelamin = Laki-laki| Y=Lancar) = 5/9
   P(Jenis Kelamin = Laki-laki| Y=Kurang Lancar) = 2/4
   P(Jenis Kelamin = Laki-laki| Y=Tidak Lancar) = 4/7
   P(Usia = 30 - 40 Tahun| Y=Lancar) = 7/9
   P(Usia = 30 - 40 Tahun| Y=Kurang Lancar) =2/4
   P(Usia = 30 - 40 Tahun| Y=Tidak Lancar) = 3/7
   P(Status = Kawin| Y=Lancar) = 4/9
   P(Status = Kawin| Y=Kurang Lancar) = 4/4
   P(Status = Kawin| Y=Tidak Lancar) = 6/7
   P(Pekerjaan = Wiraswasta| Y=Lancar) = 6/9
   P(Pekerjaan = Wiraswasta| Y=Kurang Lancar) = 2/4
   P(Pekerjaan = Wiraswasta| Y=Tidak Lancar) = 2/7
   P(Penghasilan = 25 – 50 Juta| Y=Lancar) = 2/9
   P(Penghasilan = 25 – 50 Juta| Y=Kurang Lancar) =1/4
   P(Penghasilan = 25 – 50 Juta| Y=Tidak Lancar) = 3/7
   P(Masa_Asuransi = 11 – 15 Tahun| Y=Lancar) = 7/9
   P(Masa_Asuransi = 11 – 15 Tahun| Y=Kurang Lancar) = 1/4
   P(Masa_Asuransi = 11 – 15 Tahun| Y=Tidak Lancar) = 4/7
   P(Cara Pembayaran = Semesteran| Y=Lancar) = 3/9
   P(Cara Pembayaran = Semesteran| Y=Kurang Lancar) = ¼
   P(Cara Pembayaran = Semesteran| Y=Tidak Lancar) = 1/7

3. Kalikan semua hasil variabel Lancar, Kurang Lancar dan Tidak Lancar.

   P(Laki–Laki\Lancar) * P(30–40 Tahun\Lancar) * P(Kawin\Lancar). P(Wiraswasta\Lancar) * P(25–50 Juta\Lancar) * P(11–15Tahun\Lancar). P(Semesteran\Lancar) * P(Lancar)

   $= \frac{5}{9} x \frac{7}{9} x \frac{4}{9} x \frac{6}{9} x \frac{2}{9} x \frac{7}{9} x \frac{3}{9} x \frac{9}{20}$

   = 0,5556 × 0,7778 × 0,4444 × 0,6667 × 0,2222 × 0,7778 × 0,3333 × 0,45

   = 0,0033

   P (Laki – Laki \ Kurang Lancar) * P(30 – 40 Tahun \ Kurang Lancar) * P(Kawin\Kurang Lancar) * P(Wiraswasta\ Kurang Lancar) * P(25 – 50 Juta\Kurang Lancar) * P(11 – 15 Tahun\Kurang Lancar). P(Semesteran\Kurang Lancar) * P(Kurang Lancar)

   $= \frac{2}{4} x \frac{2}{4} x \frac{4}{4} x \frac{2}{4} x \frac{1}{4} x \frac{1}{4} x \frac{1}{4} x \frac{4}{20}$

   = 0,5×0,5×1×0,5×0,25×0,25×0,25×0,2

   = 0,0004

   P(Laki – Laki\Tidak Lancar) * P(30 – 40 Tahun\Tidak Lancar) * P(Kawin\Tidak Lancar) * P(Wiraswasta\Tidak Lancar) * P(25 – 50 Juta\Tidak Lancar) * P(11 – 15 Tahun\Tidak Lancar) * P(Semesteran\Tidak Lancar). P(Tidak Lancar)

   $= \frac{2}{4} x \frac{2}{4} x \frac{4}{4} x \frac{2}{4} x \frac{1}{4} x \frac{1}{4} x \frac{1}{4} x \frac{4}{20}$

   = 0,5714×0,4286×0,857×0,2857×0,4286×0,5714×0,1429×0,35

   = 0,0007

4. Bandingkan hasil class Lancar, Kurang Lancar dan Tidak Lancar.Dari hasil diatas, terlihat bahwa nilai probabilitas tertinggi ada pada kelas (P|Lancar) sehingga dapat disimpulkan bahwa status calon nasabah tersebut masuk dalam klasifikasi “Lancar”[13].

2.6 Perhitungan Algoritma Naive Bayes pada Studi Kasus kesiapan siswa dalam menghadapi ujian nasional

Dasar pengambilan keputusan untuk kesiapan siswa dalam menghadapi ujian nasional adalah untuk memudahkan sekolah atau lembaga kursus dalam memantau kesiapan siswa dalam menghadapi ujian nasional, aspek-aspek yang dilihat meliputi nilai tryout yang di adakan oleh lembaga kursus,

Jika nilai siswa tersebut dibawah KKN maka siswa tersebut diwajibkan untuk mengikuti jam pelajaran tambahan guna menunjang kesiapan siswa dalam menghadapi ujian nasional

Berikut contoh perhitungan manual penerapan algoritma *naïve bayes* untuk menentukan kesiapan siswa dalam menghadapi ujian nasional menggunakan sample data training sebanyak 20 data dan data testing 3 data.

NAMA	N_IND	N_ING	N_MAT	N_IPA	Jurusan	Status
Aldo Banida Rafid	3	2	3	3	1	Siap
Baihaqy Hadi T	2	3	3	3	1	TIDAK
Maharnum Pramitya Lilimadani	2	2	2	2	1	TIDAK
Dea Nabila Rahmadika	2	1	2	3	1	TIDAK
Sofie Antania Hanjani	2	2	3	3	3	Siap
Yuliantina Bunga Kinanti	3	2	3	2	1	TIDAK
Irfan Rifaldi	2	3	3	3	1	TIDAK
Moch. Galuh Raga Prameswara	2	2	3	2	1	TIDAK
Muhammad Hutomo Adi N.	3	3	3	3	2	TIDAK
Zahra Putsumazki	2	2	2	2	1	TIDAK
Wildan Prakoso	2	2	3	2	2	TIDAK
Andhika Arya Perdana Suyana Putra	2	2	2	2	1	KHUSUS

Aulia Kemal Muhammad	2	2	3	2	1	TIDAK
Ika Riski Lestari	2	3	3	3	1	TIDAK
Vera Febriyanti	2	3	3	2	1	TIDAK
Sheila Zalfatika	1	2	3	2	1	TIDAK
Debby Maylinda Virtraries	2	3	4	3	2	TIDAK
Almira Jovankova Yunan	1	2	2	2	1	Siap
Mochamad Rafli Ramadhan	2	2	1	1	1	KHUSUS
Dimasta Wardhana	2	2	3	2	1	TIDAK
A	**2**	**2**	**1**	**2**	**1**	**?**
B	**3**	**3**	**2**	**2**	**2**	**?**
C	**3**	**2**	**2**	**2**	**3**	**?**

Table 2.4: Tabel Sample Training dan Testing

Soal Pertama

a. Tahap pertama yang dilakukan adalah mengitung jumlah *class/ label* :

   P(Y=KHUSUS.) = 2/20

   P(Y=TIDAK) = 15/20

   P(Y=SIAP)= 3/20

b. Tahap kedua menghitung jumlah kasus yang sama dengan *class* yang sama :

   P(N_IND=2|Y=KHUSUS.) = 2/2

   P(N_IND=2|Y=TIDAK) = 12/15

   P(N_IND=2|Y=SIAP) = 1/3

   P(N_ING=2|Y=KHUSUS.) = 2/2

   P(N_ING=2|Y=TIDAK) = 8/15

   P(N_ING=2|Y=SIAP) = 3/3

   P(N_MAT=1|Y=KHUSUS.) = 1/2

   P(N_MAT=1|Y=TIDAK) = 0/15

P(N_MAT=1|Y=SIAP) = 0/3

P(N_IPA=1|Y=KHUSUS.) = 1/2
P(N_IPA=1|Y=TIDAK) = 0/15
P(N_IPA=1|Y=SIAP) = 0/3

P(JURUSAN=1|Y=KHUSUS.) = 2/2
P(JURUSAN=1|Y=TIDAK) = 12/15
P(JURUSAN=1|Y=SIAP) = 2/3

c. Tahap ketiga kalikan semua hasil variable KHUSUS., TIDAK, dan Siap
= {P(P(N_IND=2|Y=KHUSUS.)) . P(N_ING=2|Y=KHUSUS) . P(N_MAT=1|Y=KHUSUS) . P(N_IPA=1|Y=KHUSUS.) . P(JURUSAN=1|Y=KHUSUS)
= 2/2 . 2/2 . 1/2 .1/2 . 2/2
= 1 . 1 . 0,5 . 0,5 . 1
= 0,25
TIDAK
= {P(P(N_IND=2|Y=TIDAK)) . P(N_ING=2|Y=TIDAK) . P(N_MAT=1|Y=TIDAK) . P(N_IPA=1|Y=TIDAK) . P(JURUSAN=1|Y=TIDAK)
= 12/15 . 8/15 . 0/15 . 0/15 . 12/15
= 0,8 . 0,53 . 0 . 0 . 0,8
= 0
SIAP
={P(P(N_IND=2|Y=SIAP)).P(N_ING=2|Y=SIAP).P(N_MAT=1|Y=SIAP) .P(N_IPA=1|Y=SIAP)P(JURUSAN=1|Y=SIAP)
= 1/3 . 3/3 . 0/3 . 0/3 . 2/3
= 0,3 . 1 . 0 . 0 . 0,6
= 0

d. Tahap keempat bandingkan hasil *class*KHUSUS., TIDAK, dan Siap.
Karena Hasil (P|KHUSUS) lebih besar dari (P|TIDAK) dan (P|Siap) maka keputusannya adalah KHUSUS.
0,25 > 0 maka "KHUSUS"

A	**2**	**2**	**1**	**2**	**1**	**KHUSUS**

Soal Kedua

a. Tahap pertama yang dilakukan adalah mengitung jumlah *class/ label* :
P(Y=KHUSUS) = 2/20
P(Y=TIDAK) = 15/20
P(Y=SIAP)= 3/20

b. Tahap kedua menghitung jumlah kasus yang sama dengan *class* yang sama :
P(N_IND=3|Y=KHUSUS) = 0/2
P(N_IND=3|Y=TIDAK) = 2/15
P(N_IND=3|Y=SIAP) = 1/3

P(N_ING=3|Y=KHUSUS) = 0/2
P(N_ING=3|Y=TIDAK) = 6/15
P(N_ING=3|Y=SIAP) = 0/3

P(N_MAT=2|Y=KHUSUS) = 1/2
P(N_MAT=2|Y=TIDAK) = 3/15
P(N_MAT=2|Y=SIAP) = 1/3

P(N_IPA=2|Y=KHUSUS) = 1/2
P(N_IPA=2|Y=TIDAK) = 8/15
P(N_IPA=2|Y=SIAP) = 1/3

P(JURUSAN=2|Y=KHUSUS) = 0/2
P(JURUSAN=2|Y=TIDAK) = 3/15
P(JURUSAN=2|Y=SIAP) = 0/3

c. Tahap ketiga kalikan semua hasil variable KHUSUS, TIDAK, dan Siap.

KHUSUS

={P(P(N_IND=3|Y=KHUSUS)).P(N_ING=3|Y=KHUSUS).P(N_MAT=2|Y=KHUSUS).P(N_IPA=2|Y=KHUSUS).P(JURUSAN=2|Y=KHUSUS)

= 0/2 . 0/2 . 1/2 .1/2 . 0/2

= 0 . 0 . 0,5 . 0,5 . 0

= 0

TIDAK

={P(P(N_IND=3|Y=TIDAK)).P(N_ING=3|Y=TIDAK).P(N_MAT=2|Y=TIDAK).P(N_IPA=2|Y=TIDAK).P(JURUSAN=2|Y=TIDAK)

= 2/15 . 6/15 . 3/15 . 8/15 . 3/15

= 0,13 . 0,4 . 0,2 . 0,53 . 0,2

= 0,0011024

SIAP

={P(P(N_IND=3|Y=SIAP)).P(N_ING=3|Y=SIAP).P(N_MAT=2|Y=SIAP).P(N_IPA=2|Y=SIAP)P(JURUSAN=2|Y=SIAP)

= 1/3 . 0/3 . 1/3 . 1/3 . 0/3

= 0,3 . 0 . 0,3 . 0,3 . 0

= 0

d. Tahap keempat bandingkan hasil *class*KHUSUS, TIDAK, dan Siap.

Karena Hasil (P|TIDAK) lebih besar dari (P|KHUSUS) dan (P|Siap) maka keputusannya adalah TIDAK.

0,0011024 > 0 maka “TIDAK”

B	**3**	**3**	**2**	**2**	**2**	**TIDAK**

Soal Ketiga

a. Tahap pertama yang dilakukan adalah mengitung jumlah *class/ label* :

   P(Y=KHUSUS) = 2/20

   P(Y=TIDAK) = 15/20

   P(Y=SIAP)= 3/20

b. Tahap kedua menghitung jumlah kasus yang sama dengan *class* yang sama :

   P(N_IND=3|Y=KHUSUS) = 0/2

   P(N_IND=3|Y=TIDAK) = 2/15

   P(N_IND=3|Y=SIAP) = 1/3

   P(N_ING=2|Y=KHUSUS) = 2/2

   P(N_ING=2|Y=TIDAK) = 8/15

   P(N_ING=2|Y=SIAP) = 3/3

   P(N_MAT=2|Y=KHUSUS) = 1/2

   P(N_MAT=2|Y=TIDAK) = 3/15

   P(N_MAT=2|Y=SIAP) = 1/3

   P(N_IPA=2|Y=KHUSUS) = 1/2

   P(N_IPA=2|Y=TIDAK) = 8/15

   P(N_IPA=2|Y=SIAP) = 1/3

   P(JURUSAN=3|Y=KHUSUS) = 0/2

   P(JURUSAN=3|Y=TIDAK) = 3/15

   P(JURUSAN=3|Y=SIAP) = 1/3

c. Tahap ketiga kalikan semua hasil variable KHUSUS, TIDAK, dan Siap.

KHUSUS

={P(P(N_IND=3|Y=KHUSUS)).P(N_ING=2|Y=KHUSUS).P(N_MAT=2|Y=KHUSUS).P(N_IPA=2|Y=KHUSUS).P(JURUSAN=3|Y=KHUSUS)

= 0/2 . 2/2 . 1/2 .1/2 . 0/2

= 0 . 1 . 05 . 0,5 . 0

= 0

TIDAK

={P(P(N_IND=3|Y=TIDAK)).P(N_ING=2|Y=TIDAK).P(N_MAT=2|Y=TIDAK).P(N_IPA=2|Y=TIDAK).P(JURUSAN=3|Y=TIDAK)

= 2/15 . 8/15 . 3/15 . 8/15 . 3/15

= 0,13 . 0,53 . 0,2 . 0,53 . 0,2

= 0,00146068

SIAP

={P(P(N_IND=3|Y=SIAP)).P(N_ING=2|Y=SIAP).P(N_MAT=2|Y=SIAP).P(N_IPA=2|Y=SIAP)P(JURUSAN=3|Y=SIAP)

= 1/3 . 3/3 . 1/3 . 1/3 . 1/3

= 0,3 . 1 . 0,3 . 0,3 . 0,3

= 0,027

d. Tahap keempat bandingkan hasil *class*KHUSUS, TIDAK, dan Siap.

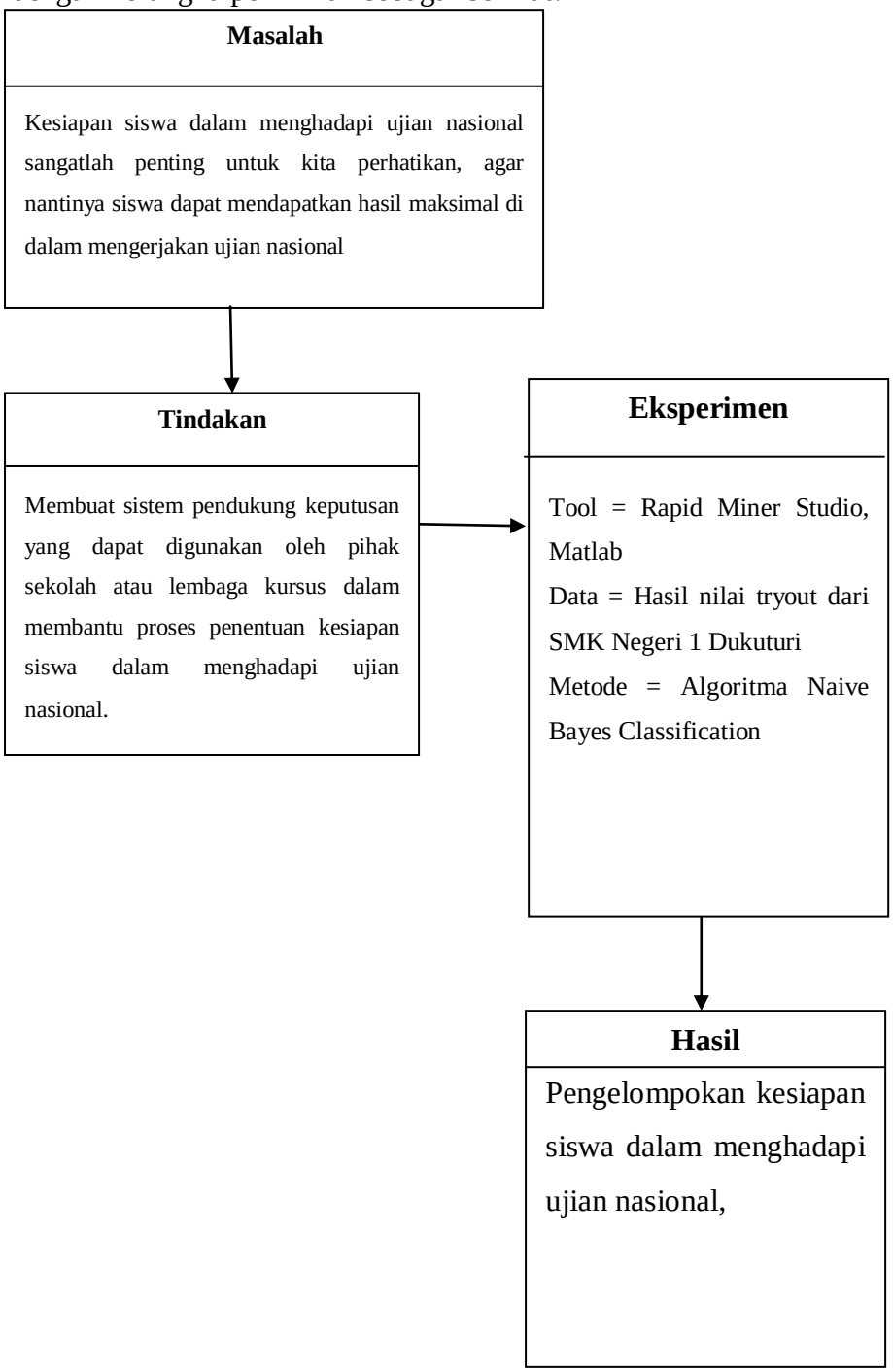
Karena Hasil (P|TIDAK) lebih besar dari (P|KHUSUS) dan (P|Siap) maka keputusannya adalah TIDAK.

0,027 > 0,00146068 > 0 maka “Siap”

B	**3**	**2**	**2**	**2**	**3**	**Siap**

2.7 Kerangka Pemikiran

Penulis perlu membuat gambaran singkat sebagai alur penyusunan laporan ini dengan kerangka pemikiran sebagai berikut:



Gambar 2.5 : Kerangka Pikiran