BAB II
KAJIAN LITERATUR

1.1. Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database. Untuk dapat melakukan analisis data menggunakan teknik-teknik dalam *data mining*, sebaiknya pemahaman tentang defenisi *data mining* itu sendiri perlu diberikan. Terdapat beberapa defenisi *data mining* menurut para ahli.

Secara sistematis, langkah utama untuk melakukan *data mining* terdiri dari tiga tahap, yaitu sebagai berikut:

1. Eksplorasi atau pemrosesan awal data

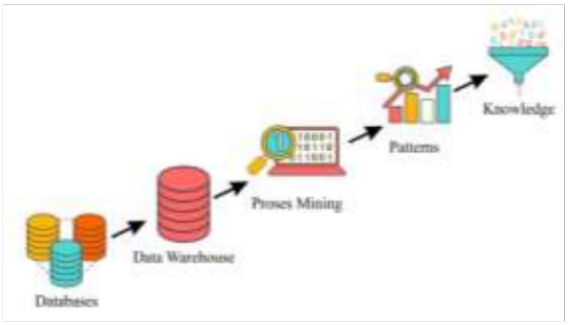
   Eksplorasi atau pemrosesan awal data terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan missing value, reduksi dimensi, pemilihan subset atribut, dan sebagainya.
2. Membangun model dan validasi

   Membangun model dan validasi, yaitu melakukan analisis dari berbagai model dan memilih model sehingga menghasilkan kinerja yang terbaik. Pembangunan model dilakukan menggunakan metode-metode seperti klasifikasi, regresi, analisis *cluster*, dan asosiasi.
3. Penerapan

   Penerapan dilakukan dengan menerapkan model yang dipilih pada data yang baru untuk menghasilkan kinerja yang baik pada masalah yang diinvestigasi.

1.1.1. Tahap-Tahap Data Mining

Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap sebagaimana seperti yang digambarkan pada gambar.



Gambar 2. 1. Tahap-tahap data mining

1. *Cleaning*

Data cleaning (Pembersihan data) adalah proses yang dilakukan untuk menghilangkan *noise* pada data yang tidak konsisten atau bisa disebut tidak relevan. Data yang diperoleh dari *database* suatu perusahaan maupun hasil eksperimen yang sudah ada, tidak semuanya memiliki isian yang sempurna misalnya data yang hilang, data yang tidak valid, atau bisa juga hanya sekedar salah ketik. Data yang tidak relevan itu dapat ditangani dengan cara dibuang atau sering disebut dengan proses *cleaning*. Proses *cleaning* dapat berpengaruh terhadap performa dari teknik *data mining*. Berikut merupakan contoh *cleaning* dataset.

No	Job_level	Job_duration_in current_job_level	Person_level	Employee_type	Gender	...	Best_performance
1	JG04	1.17	PG03	RM_type_A	Male	...	0
2	JG04	1.83	PG03	RM_type_A	Male	...	1
3	JG03	0.75	PG01	RM_type_B	Male	...	0
4	JG03	0	PG01	RM_type_B	Male	...	0
5	JG04	1.17	PG03	RM_type_A	Male	...	0
6	JG04	0.75	PG03	RM_type_B	Male	...	0
7	JG04	1.83	PG03	RM_type_B	Female	...	1
8	JG03	0.75	PG01	RM_type_B	Male	...	0
9	JG04	1.83	PG03	RM_type_B	Male	...	0
10	JG04	1.17	PG03	RM_type_A	Male	...	0

Gambar 2. 2. contoh sampel dataset[9]

Tahapan ini memiliki peran cukup penting dari keseluruhan proses penelitian ini, pengisian missing *values* dimaksudkan untuk menjaga jumlah data yang bisa digunakan untuk proses data mining selanjutnya. Dataset yang digunakan memiliki jumlah *missing values* yang tidak sedikit. Rangkuman jumlah *missing values* pada setiap atribut yang terdapat pada data ini dapat dilihat pada gambar.

job_level	0
job_duration_in_current_job_level	0
person_level	0
job_duration_in_current_person_level	0
job_duration_in_current_branch	0
Employee_type	12
Employee_status	0
gender	0
age	0
marital_status_maried(1/0)	0
number_of_dependences	0
number_of_dependences (male)	0
number_of_dependences (female)	0
Education_level	3608
GPA	7452
year_graduated	3946
job_duration_as_permanent_worker	2055
job_duration_from_training	0
branch_rotation	0
job_rotation	0
assign_of_otherposition	0
annual leave	0
sick_leaves	0
Avg_achievement_%	6289
Last_achievement_%	6302
Achievement_above_100%_during3quartal	6302
achievement_target_1	6727
achievement_target_2	6727
achievement_target_3	6727
Best Performance	0

Gambar 2. 3. Jumlah *missing values* per atribut dari dataset contoh

Missing values pada dataset ini akan diselesaikan menggunakan pendekatan *imputation*, yaitu proses mengisi *missing values* dengan nilai-nilai hasil estimasi menggunakan metode tertentu. Selain *imputation* pendekatan penanganan *missing values* dapat berupa penghapusan baris data yang memiliki *missing values*, akan tetapi pendekatan tersebut akan mengakibatkan jumlah dataset semakin berkurang drastis jika jumlah *missing values* yang terdapat pada dataset berjumlah banyak.

Metode yang digunakan untuk melakukan *missing values imputation* pada dataset ini adalah *Missforest*. *Missforest* merupakan teknik *missing values imputation* berbasis pada metode Random Forest di mana metode ini menggunakan model Random Forest pada data yang tersedia pada dataset untuk kemudian melakukan estimasi pada missing values pada dataset tersebut. *MissForest* dapat digunakan untuk pengisian missing values pada dataset dengan tipe data yang *heterogeny* (*mixed data-type*). Dataset hasil *missing values imputation* menggunakan metode *Missforest* dapat dilihat pada Tabel berikut.

No	Atribut Last_achievement_% (sebelum *missing values imputation*)	Atribut Last_achievement_% (setelah *missing values imputation*)
1	?	33.0385
2	46.8	46.8
3	?	32.1895
4	?	31.4021
5	?	38.1723
6	?	22.6998
7	?	33.0385

Gambar 2. 4. Sampel atribut sebelum dan sesudah missing values imputation

2. *Transformation*

Transformasi data merupakan proses pengubahan data dan penggabungan data ke dalam format tertentu. *Data mining* membutuhkan format data khusus sebelum diaplikasikan. Misalnya metode standar seperti analisis asosiasi dan *clustering* hanya bisa menerima input data yang bersifat kategorikal. Karenanya data yang berupa angka numerik apabila mempunyai sifat kontinu perlu dibagi-bagi menjadi beberapa interval.

Pada tahap ini, data-data yang memiliki bentuk data nominal akan di transformasi menjadi bentuk numerik. Bentuk transformasi data menggunakan mekanisme sederhana yaitu mengubah semua data nominal dengan kode numerik 0-n, di mana n adalah varian terakhir dari data nominal pada atribut tersebut. Contoh hasil transformasi data dapat dilihat pada tabel.

No	Job_level	Job_duration_in current_job_level	Person_level	Employee_type	Gender	...	Best_performance
1	4	1.17	3	0	1	...	0
2	4	1.83	3	0	1	...	1
3	3	0.75	1	1	1	...	0
4	3	0	1	1	1	...	0
5	4	1.17	3	0	1	...	0
6	4	0.75	3	1	1	...	0
7	4	1.83	3	1	0	...	1
8	3	0.75	1	1	1	...	0
9	4	1.83	3	1	1	...	0
10	4	1.17	3	0	1	...	0

Gambar 2. 5. Transformasi dataset sampel

3. *Mining Process*

Proses *mining* dapat disebut juga sebagai proses penambangan data. Proses *mining* merupakan proses utama yang menggunakan metode untuk menemukan pengetahuan berharga yang tersembunyi dari data.

4. *Evaluation and Precentation*

a. Evaluasi Pola (*Pattern Evaluation*)

Evaluasi pola bertugas untuk mengidentifikasi pola-pola yang menarik ke dalam *knowledge based* yang ditemukan. Pada tahap ini dihasilkan pola-pola yang khas dari model klasifikasi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai dengan hipotesa, terdapat beberapa alternatif yang bisa diambil seperti menjadikannya umpan balik untuk memperbaiki proses *data mining*, atau mencoba metode *data mining* lain yang lebih sesuai.

b. Presentasi Pengetahuan (*Knowledge Presentation*)

Knowledge presentation merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan atau informasi yang telah digali oleh pengguna. Tahap terakhir dari proses *data mining* adalah memformulasikan keputusan dari hasil analisis yang didapat.[10]

1.1.2. Fungsi Data Mining

Data mining mengidentifikasi kesimpulan-kesimpulan atau fakta-fakta yang disarankan berdasarkan penyaringan melalui data untuk menjelajahi pola-pola atau anomali-anomali data. *Data mining* mempunyai 5 fungsi:

1. *Clustering*

   Clustering adalah mengindetifikasi kelompok-kelompok dari produk-produk atau barang-barang yang memiliki karakteristik khusus (*clustering* berbeda dengan *classification*, dimana pada clustering tidak ada defenisi-defenisi karakteristik yang diberikan pada waktu *classification*).
2. *Classification*

   Classification adalah menyimpulkan defenisi-defenisi karakteristik sebuah grup. Contohnya: pelanggan-pelanggan perusahaan yang telah berpindah kesaingan perusahaan yang lain.
3. *Association*

   Association adalah mengidentifikasin hubungan antara kejadian-kejadian yang terjadi pada suatu waktu, contohnya: isi-isi dari keranjang belanja
4. *Sequencing*

   Sequencing adalah mengidentifikasikan hubungan-hubungan yang berbeda pada suatu periode waktu tertentu (hampir sama dengan *association*), contohnya: pelanggan-pelanggan yang mengunjungi supermarket secara berulang-ulang

5. *Forecasting*

   Forescasting adalah perencanaan pengendalian nilai pada masa yang akan datang berdasarkan pola-pola dengan sekumpulan data yang besar, seperti peramalan permintaan pasar.[11]

1.1.3. Faktor-Faktor Data Mining

Beberapa faktor yang mendukung perlunya dilakukan *data mining* adalah:

1. Data telah mencapai jimlah dan ukuran yang sangat besar

   Hasil dan proses *data mining* merupakan suatu informasi yang akan mendasari tindakan tertentu sehingga tingkat kebenaran informasi tersebut menjadi sangat signifikan, dan makin besar serta makin banyak data yang digunakan maka akan semakin valid hasilnya. Perkembangan data dalam hal jumlah dan ukuran telah mencapai kecepatan yang sangat cepat, sehingga ukuran basis data yang dimiliki oleh sebuah perusahaan bisa mencapai kisaran *gigabyte* atau bahkan *terabyte*.
2. Telah dilakukan proses *data warehousing*

   Untuk mencapai hasil yang memuaskan, maka sumber data yang digunakan dalam proses *data mining* sering kali merupakan data gabungan dari banyak departemen, daerah operasi bahkan dari sumber-sumber lain seperti data kependudukan. Oleh karena itu maka disarankan perlunya proses *data warehousing* untuk menjaga konsistensi, memberikan prespektif yang lebih baik terhadap data dan menjaga integritas data.
3. Kemampuan Komputasi yang semakin terjangkau

   Pada dasarnya prosre *data mining* melakukan banyak akses terhadap data yang sangat besar. Selain itu juga melakukan proses komputasi yang membutuhkan sumber daya sangat besar. Penurunan harga yang cukup cepat terhadap perangkat keras komputer serta semakin tingginya kinerja yang berhasil dicapai oleh perangkat computer maupun teknologi pengolahan data seperti teknologi paralel proses saat, menjadikan proses saat ini, mejadikan proses *data mining* sudah cukup layak untuk dilakukan secara komersial.
4. Persaingan bisnis semakin ketat

   Tekanan persangingan bisnis yang semakin ketat mendorong perusahaan-perusahaan untuk selalu berinovasi agar mampu meningkatkan daya saingnya dipasar global. Beberapa tren yang berkembang saat ini adalah:

   a. Setiap bisnis adalah bisnis pelayanan

b. Adanya fenomena kustomisasi oleh masyarakat
c. Informasi adalah produk. [12]

1.1.4. Komponen Data Mining

Sebagai salah satu bagian dari sebuah sistem informasi, *data mining* menyediakan perencanaan mulai dari ide hingga implementasi akhir. Komponen-komponen dari rencana *data mining* adalah:

a. Analisa Masalah (*Analyzing the Problem*)

   Data sumber atau data asal harus ditaksir untuk dilihat apakah data tersebut memenuhi kriteria *data mining* atau tidak. Kualitas kelimpahan data yaitu faktor utama untuk memutuskan apakah data tersebut cocok dan tersedia sebagai tambahan. Hasil yang diharapkan dari dampak *data mining* harus dengan hati-hati dipastikan dan menegerti bahwa data yang diperlukan membawa informasi yang bisa diekstrak.

b. Mengekstrak dan membersikan data (*Extracting and Cleansing The Data*)

   Data pertama kali diektrak dari data aslinya, seperti dari OLP basis data, *text file*, *Mircosoft Acces Database*, dan bahkan dari *spreadsheet*, lalu data tersebut diletakkan dalam data *warehouse* yang mempunyai struktur yang sesuai dengan data model secara khas. *Data Transformation Service* (DTS) dipakai untuk mengekstrak dan membersihkan data dari tidak konsistennya dan tidak kompatibelnya dengan format yang sesuai.

c. Validasi Data (*Validating The Data*)

   Sekali data telah diekstrak dan dibersihkan, ini adalah latihan yang bagus untuk menelusuri model yang telah kita ciptakan untuk memastikan bahwa semua data yang ada adalah data sekarang dan tetap

d. Membuat dan Melatih Model (*Creating and Training the Model*)

   Ketika algoritmaa diterapkan pada model, struktur telah dibangun. Hal ini sangatlah penting pada saat ini untuk melihat data yang telah dibangun untuk memastikan bahwa data tersebut menyerupai fakta didalam data sumber.

e. Query Data dari Model *Data mining* (*Querying the Model Data*)

   Ketika model yang telah cocok diciptakan dan dibangun, data yang telah dibuat tersedia untuk mendukung keputusan. Hal ini biasanya melibatkan penulisan front end query aplikasi dengan program aplikasi atau suatu program basis data.

f. Evaluasi Validitas dari Mining Model (*Maintaning the validity of the data mining model*)

Sebuah model *data mining* terkumpul lewat beberapa waktu karakteristik data awal seperti granularitas dan validitas mungkin berubah. Karena model *data mining* dapat terus berubah seiring perkembangan waktu.[13]

1.2. Klasifikasi

Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target yang memetakan setiap set atribut (atribut) kesatu jumlah label kelas yang tersedia, bisa juga diartikan pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari jumlah kelas yang tersedia. Klasifikasi melakukan pembangunan model berdasarkan data latih yang ada, kemudian menggunakan model tersebut untuk mengklasifikasikan pada data yang baru, dengan tujuan sistem yang dibangun nantinya dapat melakukan klasifikasi semua data set dengan benar, keberhasilannya perlu dilakukan pengukuran setelah pola terbentuk. Terdapat beberapa klasifikasi pada penelitian ini, antara lain:

1. *Random Forest*, merupakan salah satu jenis algoritma klasifikasi yang terdiri dari lebih satu pohon keputusan yang setiap pohon keputusan dibentuk bergantung pada nilai-nilai vektor acak sampel secara independen dan identik didistribusikan yang sama untuk semua pohon. Metode ini merupakan salah satu metode klasifikasi yang sangat akurat digunakan dalam melakukan prediksi, bisa menangani inputan variabel yang sangat besar jumlahnya tanpa *overfitting*, dan membantu menghilangkan korelasi antara pohon keputusan seperti karakteristik *ensemble methods*.
2. *Decision Tree*, merupakan algoritma pengambilan keputusan yang melakukan partisi rekursif atas ruang *instance*, sebuah pohon keputusan tipikal terdiri dari simpul internal, tepi dan simpul daun. Setiap simpul internal disebut simpul keputusan yang mewakili tes pada atribut atau subset atribut, dan masing-masing *edge* diberi label dengan nilai spesifik atau rentang nilai atribut input. Pengklasifikasi *Decision Tree* memperoleh *accuracy* yang serupa dan terkadang lebih baik jika dibandingkan dengan metode klasifikasi lainnya.
3. *Naïve Bayes*, merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan, juga dikemukakan oleh ilmuwan Inggris

Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

4. *Artificial Neural Network*, atau jaringan syaraf tiruan adalah model *non-linear* yang rumit dibangun dari komponen yang secara individu berperilaku mirip seperti model regresi yang dapat merepresentasikan sebuah grafik, dan beberapa sub-grafik tampaknya ada integritas yang sama dengan gerbang logika. Struktur dari jaringan *neuron* atau saraf secara terperinci dirancang terlebih dahulu.

Proses klasifikasi didasarkan pada empat komponen:

a. Kelas Variabel dependen berupa kategori yang mempresentasikan "label" yang terdapat pada objek.
b. Prediktor Variabel independen yang direpresentasikan oleh karakteristik data.
c. *Training dataset* Satu set data yang mempunya nilai dari kedua komponen yang digunakan untuk menentukan kelas yang cocok berdasarkan prediktor.
d. *Testing dataset* Berupa data baru yang diklasifikasian oleh model data yang telah di buat dan *accuracy* klasifikasi di evaluasi. [14]

Tools yang digunakan dalam proses klasifikasi adalah Rapidminer Studio 10.1.

1.2.1. Algoritma Naïve Bayes

Naïve bayes adalah sebuah algoritma *supervised learning* berdasarkan teorema Bayes yang digunakan untuk memecahkan masalah klasifikasi dengan mengikuti pendekatan probabilistik. *Naïve bayes* dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu memprediksi peluang dimasa depan berdasarkan pengalaman sebelumnya sehingga dikenal sebagai Teorema Bayes. Klasifikasi naïve bayes diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak berhubungan dengan ciri dari kelas yang lainnya. Sehingga *Naïve Bayes Classifier* dapat diartikan sebagai 9 metode klasifikasi yang berdasarkan teorema bayesian dengan asumsi bahwa setiap variabel atau parameter penentu keputusan bersifat bebas (*independence*) sehingga ada atau tidaknya variabel atau parameter sama sekali tidak terkait dengan keberadaan atribut yang lainnya. Algoritma *naïve bayes* menggunakan dua bentuk data untuk proses prediksinya yaitu dataset dan data tes. *Dataset* digunakan sebagai data latih untuk menentukan peluang yang akan terjadi. Sedangkan data tes sebagai data uji atau data yang akan diprediksi dari peluang yang sudah terbentuk tersebut. Menurut Effrida dan Fricles keuntungan penggunaan *naïve bayes* dalam penelitiannya adalah bahwa metode ini hanya membutuhkan data training yang kecil untuk

menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian dan juga dapat bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan. Teorema *Naïve Bayes* dirumuskan sebagai berikut:

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

Keterangan:

X : Data dengan class yang belum diketahui

H : Hipotesis data merupakan suatu *class* spesifik

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posterior probabilitas)

P(H) :Probabilitas hipotesis H (prior probabilitas)

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

P(X) : Probabilitas X

Data yang dapat diolah pada Algoritma Naïve Bayes Classifier terdapat dua macam yaitu data jenis kategori dan data jenis *numeric*.[15] Untuk menjelaskan metode *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *Naïve Bayes* di atas disesuaikan sebagai berikut:

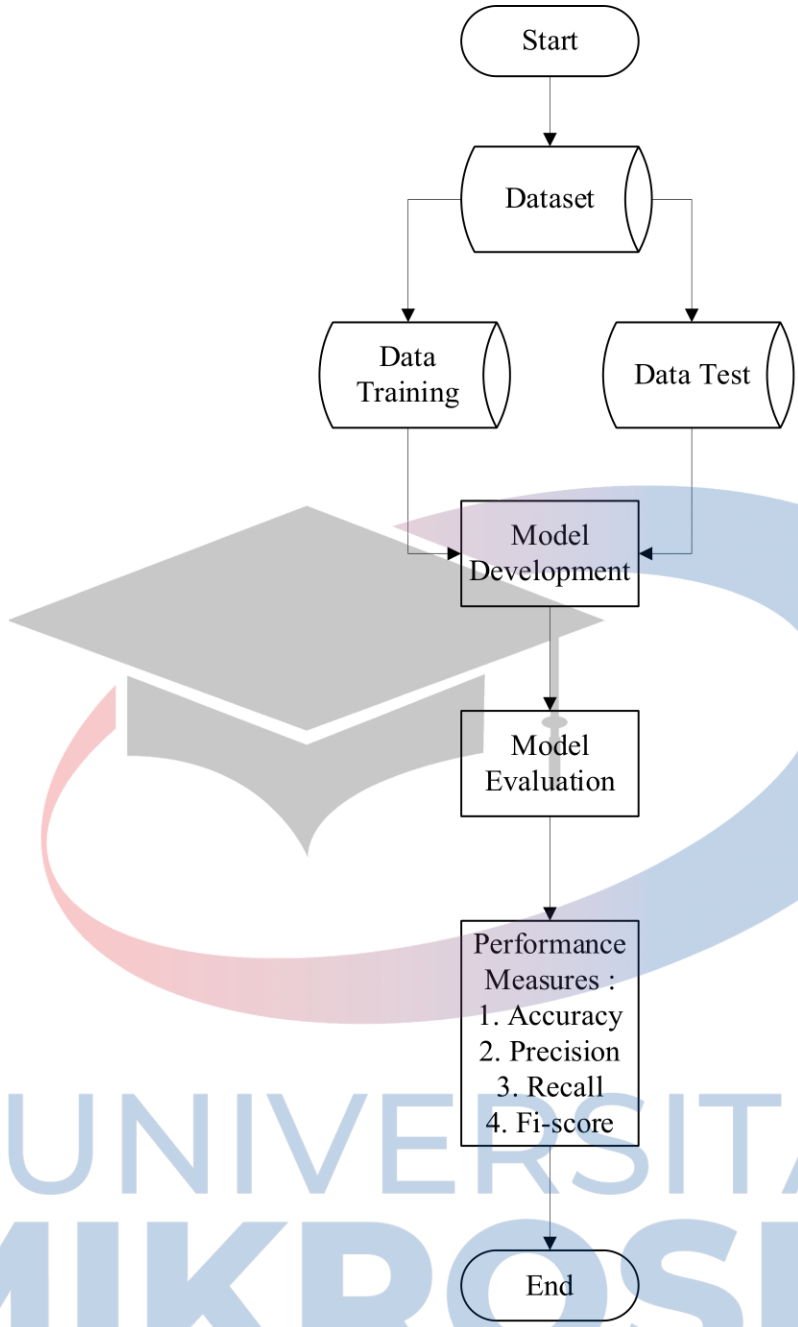
$$(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots . Fn|C)}{P(F1 \dots Fn)} \quad (2)$$

Di mana Variabel C merepresentasikan kelas, sementara variabel F1...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 di atas dapat pula ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{Prior\ x\ likehood}{evideance} \quad (3)$$

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.[16]



Gambar 2. 6. Alur Proses Algoritma Naive Bayes

1.3. Evaluasi

Secara umum evaluasi adalah suatu proses dalam menyediakan informasi untuk mengetahui sejauh mana kegiatan tersebut telah dicapai. Evaluasi mengukur suatu pekerjaan atau hal-hal yang dilakukan, sangat bermanfaat karena dapat mengetahui tingkatan pekerjaan dan juga sebagai penilaian terhadap apa yang telah dikerjakan. Pengertian evaluasi yang lain adalah suatu proses sistematis dalam menentukan atau membuat keputusan terhadap sejauh mana program atau sistem suatu aplikasi telah dicapai.

Evaluasi adalah kegiatan untuk mengumpulkan informasi tentang bekerjanya sesuatu, yang selanjutnya informasi tersebut digunakan untuk menentukan alternatif yang tepat dalam mengambil keputusan. Fungsi utama evaluasi dalam hal ini adalah menyediakan informasi-informasi yang berguna bagi pihak decision maker untuk menentukan kebijakan yang akan diambil berdasarkan evaluasi yang telah dilakukan.

Dari dua pendapat diatas dapat disimpulkan evaluasi adalah kegiatan yang di lakukan untuk mengukur dan mengumpulkan informasi tentang sejauah mana tingkat kemudahan pengguna dalam mempelajari website ketika pertama kali menggunakan website. Mengetahui tingkat kecepatan dan kemudahan pengguna dalam mencari informasi yang dibutuhkan, berapa jumlah kesalahan, serta mengukur tingkat kepuasan pengguna.[17]

1.3.1. K-Fold Cross Validation

K-Fold Cross Validation merupakan suatu metode untuk membagi data ke dalam beberapa bagian (*fold*) sebanyak k untuk menentukan data *training*, dan data *testing*. Kemudian menggunakan data *traning* untuk melatih model dan dataset *testing* untuk menguji model. Setelah mengevaluasi kinerja model berdasarkan matrik kesalahan untuk menentukan keakuratan model. Namun metode ini, tidak terlalu dapat diandalkan karena akurasi yang diperoleh untuk satu data *testing* bisa sangat berbeda dengan akurasi yang diperoleh untuk data *testing* yang berbeda. *K-fold cross validation* memberikan solusi untuk masalah ini dengan membagi data menjadi fold dan memastikan bahwa setiap fold digunakan sebagai dataset *testing* di beberapa titik *cross validation*. [18], [19]

1.3.2. Confusion Matrix

Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya *confusion matrix* mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. Pada pengukuran kinerja menggunakan *confusion matrix*, terdapat 4 (empat) istilah sebagai representasi hasil proses klasifikasi. Keempat istilah tersebut adalah *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Nilai *True Negative* (TN) merupakan jumlah data negatif yang terdeteksi dengan benar, sedangkan *False Positive* (FP) merupakan data negatif namun terdeteksi sebagai data positif. Contoh *confusion matrix* untuk klasifikasi biner ditunjukkan pada gambar berikut.[20]

Tabel 2. 1. Tabel Confusion Matrix

		Kelas Prediksi	
		1	0
Kelas Sebenarnya	1	TP	FN
	0	FP	TN

Keterangan:

TP (*True Positive*) : Jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 1

TN (*True Negative*) : Jumlah dokumen dari kelas 0 yang benar diklasifikasikan sebagai kelas 0

FP (*False Positive*) : Jumlah dokumen dari kelas 0 yang salah diklasifikasikan sebagai kelas 1

FN (*False Negative*) : Jumlah dokumen dari kelas 1 yang benar diklasifikasikan sebagai kelas 0

Melalui 4 data tersebut, dapat diperoleh data-data lain yang sangat berguna untuk mengukur performa sebuah model, diantaranya:

1. *Accuracy* = Total keseluruhan seberapa sering model benar mengklasifikasi.

$$\frac{TP + TN}{Total} \quad (4)$$

2. *Precision* = Ketika model memprediksi positif, seberapa sering prediksi itu benar.

$$\frac{TP}{FP + TP} \quad (5)$$

3. *Recall* = Ketika kelas aktualnya positif, seberapa sering model memprediksi positif.

$$\frac{TP}{FN + TP} \quad (6)$$

4. *F1-score* = rata-rata harmonik dari *Precision* dan *Recall*.

$$2 * \frac{precision * recall}{precision + recall} \quad (7)$$

Setelah itu, *Receiver Operating Charateristics* (ROC) akan digunakan untuk memvisualisasikan secara dua dimensi performa dari setiap dari klasifikasi yang diujikan, dimana garis horizontal merupakan nilai *false positive*, sedangkan garis *vertical* berupa *true positive*. Nilai *Area Under Curve* (AUC) merupakan area dibawah grafik ROC. Untuk pengkategorian hasil AUC, nilai kualitas suatu klasifikasi berdasarkan nilai AUC-nya bisa dilihat pada tabel dibawah ini.[21]

Tabel 2. 2. Kategori nilai AUC

Nilai AUC	Kategori
0,90 - 1.00	Excellent
0,80 - 0,90	Good
0,70 - 0,80	Fair
0,60 - 0,70	Poor
<0,60	Failure

1.4. Performa Mahasiswa

Kata performa diterjemahkan dari bahasa Inggris *performance* yang juga memiliki arti prestasi atau kinerja akademik. *Performance* dalam bahasa Inggris memiliki arti pertunjukan, pembuatan, daya guna, prestasi dan hasil. Dari terjemahan tersebut dapat dilihat bahwa *performance* mempunyai penekanan pada kemampuan seseorang dalam melaksanakan tugasnya. Kemampuan akademik yaitu sejumlah kapasitas yang dimiliki seseorang dalam proses belajar baik itu berupa materi bidang studi, wawasan, berbahasa dan diskusi. Jadi bisa diartikan bahwa kemampuan akademik merupakan suatu kapasitas yang dimiliki seseorang dalam melakukan suatu pembelajaran.[22]

1.5. Information Gain

Information Gain sering digunakan untuk meranking atribut yang paling berpengaruh terhadap kelasnya. Nilai *gain* dari suatu atribut, diperoleh dari nilai entropi sebelum pemisahan dikurangi dengan nilai entropi setelah pemisahan. Tujuan pengurangan atribut pengukuran nilai informasi diterapkan sebagai tahap sebelum pengolahan awal. Hanya atribut memenuhi kriteria yang ditentukan dipertahankan untukdigunakan oleh algoritma klasifikasi.[23]

Langkah-langkah dalam perhitungan bobot *information gain* sebagai berikut:

1. Menghitung nilai *entropy* pada dataset :

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (8)$$

Keterangan :

S : Himpunan kasus

n : Jumlah partisi

pi : Proporsi dari Si terhadap S

2. Menghitung *information gain* pada dataset:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (9)$$

Keterangan :

S : Himpunan Kasus

A : Atribut

n : Jumlah partisi atribut

|Si| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam s

1.6. Peneliti Terdahulu

Tabel 2. 3. Penelitian Terdahulu[7], [14], [24]–[26]

No	Penulis	Judul	Hasil
1	M. Riski Qisthiano dkk (2021)	Pengembangan Model Untuk Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu dengan Metode Naïve Bayes	Hasil data uji yang didapat memiliki tingkat *Accuracy* sebesar 0.810 dengan menggunakan Algoritma Naïve Bayes. Untuk nilai percision Kelas Tepat Waktu memiliki nilai "0.81" sedangkan Tidak Tepat Waktu sebesar "0.81". Untuk nilai *recall* Kelas Tepat Waktu memiliki nilai "0.82" sedangkang kelas Tidak Tepat Waktu sebesar "0.80".

2	Asrul Azhari Muin dan Andi Abdillah (2020)	Perancangan Sistem Klasifikasi Mahasiswa untuk Prediksi Performa Mahasiswa Menggunakan Naïve Bayes Classifier	Hasil Penelitian terhadap 275 data alumni menggunakan algoritma *Naïve Bayes Classifier* yang digunakan untuk klasifikasi performa mahasiswa menghasilkan model klasifikasi dengan rata-rata nilai *accuracy, precision, recall* dan *f1-score* masing-masing sebesar 94%, 93%, 94%, dan 94% yang dihitung menggunakan metode *10-Fold Cross Validation* dan Confusion Matrix.
3	Lila Setiyani dkk (2020)	Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode *Data Mining Naïve Bayes*	Hasil identifikasi literatur yang digunakan, didapat bahwa metode *data mining* naïve bayes dapat membuat suatu prediksi mengenai kelulusan mahasiswa tepat waktu dengan memperhitungkan atribut – atribut dari database perguruan tinggi yang digunakan. Sedangkan untuk tingkat *accuracy* ketiga literatur menghasilkan *accuracy* di atas 90% walaupun dengan menggunakan jumlah atribut dan aplikasi *data mining* yang berbeda.
4	Yopi Apridiansyah, Nuri David Maria Putra Veronika dan Erwin Dwika Putra (2021)	Prediksi Kelulusan Mahasiswa Fakultas Teknik Informatika Universitas Muhammadiyah Bengkulu Menggunakan Metode Naive Bayes	Dengan menerapkan algoritma Naïve Bayes dapat memprediksi tingkat kelulusan mahasiswa tepat waktu atau tidak yang dapat berguna dalam memberikan informasi dan masukan bagi pihak perguruan tinggi dalam membuat kebijakan kedepannya. Serta Tingkat keberhasilan prediksi algoritma Naïve Bayes yang ditentukan oleh atribut nama, NPM, dan nilai IPK menghasilkan nilai

			persentase precession sebesar 90 %, *Recall* 100 % dan *accuracy* sebesar 90 %.
5	Nurul Khasanah dkk (2022)	Prediksi Kelulusan Mahasiswa dengan Metode *Naïve Bayes*	Penelitian prediksi tingkat kelulusan mahasiswa menggunakan 379 data, dengan rincian data training 303 data dan data testing 76 data. Atribut yang digunakan, antara lain: nama, status mahasiswa, status perkawinan, IPS, IPK, dan status kelulusan. Tahapan yang dilakukan dalam penelitian, antara lain: identifikasi masalah, pengumpulan data, *data cleaning, data transformation* (dibagi menjadi *data training* dan *data tesing*), klasifikasi dengan KNN, validasi, evaluasi dan hasil. Hasil penelitian yang diperoleh yaitu *accuracy* = 88,16%, *precision* = 93,62% dan *recall* = 88%. Hasil klasifikasi termasuk dalam kategori *good classification.*