BAB 2
TINJAUAN PUSATAKA

2.1 Landasan Teori

Pada bagian landasan teori akan dijelakan beberapa teori yang berkaitan dengan penelitian yang dilakukan. Dimana hal tersebut bermanfaat bagi peneliti sebagai acuan dan dasar pemahaman dalam penelitian yang akan dilakukan.

2.1.1 Tuberkulosis

Tuberkulosis (TB) merupakan penyakit yang disebabkan oleh bakteri *mycobacterium tuberculosis*. Proses penularan infeksi oleh bakteri *mycobacterium tuberculosis* terjadi secara inhalasi (inhalasi ialah proses saat menghirup oksigen melalui hidung dan masuk ke paru-paru) basil yang mengandung *droplet nuclei* (percikan halus), khususnya yang didapat dari pasien TB paru dengan batuk berdarah atau berdahak yang mengandung basil tahan asam (BTA). Gejala yang dialami penderita TB paru antar lain : batuk berdahak lebih dari dua minggu, batuk serta mengeluarkan darah, dada terasa nyeri dan sesak. Faktor – faktor penyebab penyebab penyakit TB ialah : merokok, kelembaban hawa atau udara, keadaan tempat tinggal, kepadatan hunian kamar tidur.

2.1.2 Data

Data menurut Indrajani adalah fakta – fakta mentah yang kemudian dikelola sehingga menghasilkan informasi yang penting bagi sebuah perusahaan atau organisasi [1]. Sedangkan menurut menurut Fathansyah data adalah suatu nilai yang merepresentasikan fakta di dunia nyata dan mewakili suatu objek seperti manusia (pegawai, siswa, pembeli, pelanggan), barang, hewan, peristiwa, konsep, keadaan, dsb. Nilai tersebut lalu direkam dalam bentuk angka, huruf, simbol, teks, gambar, atau kombinasinya [2].

2.1.3 Basis Data

Menurut Fathansyah [2], basis data merupakan sekumpulan data yang saling berkaitan untuk memenuhi kebutuhan tertentu yang disimpan secara elektronik agar dapat dimanfaatkan dengan mudah.

Menurut Connolly dan Begg, basis data adalah sebuah kumpulan data yang secara logis terkait dan dirancang untuk memenuhi suatu kebutuhan informasi dari sebuah organisasi [3].

Menurut Indrajani, basis data adalah kumpulan data yang saling berhubungan secara logis dan didesain untuk mendapatkan data yang dibutuhkan oleh suatu organisasi [1].

2.1.4 *Data Mining*

Data mining merupakan suatu kegiatan kegiatan mengekstraksi atau menambang pengetahuan dari data yang berukuran atau berjumlah besar. *Data mining* juga dapat diartikan sebagai pengekstrak informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan [5].

Data mining mengacu pada proses untuk menambang (*mining*) pengetahuan dari sekumpulan data yang sangat besar. Maka dari itu *data mining* diperlukan adalah karena adanya sejumlah besar data yang dapat digunakan untuk menghasilkan informasi dan pengetahuan yang berguna.

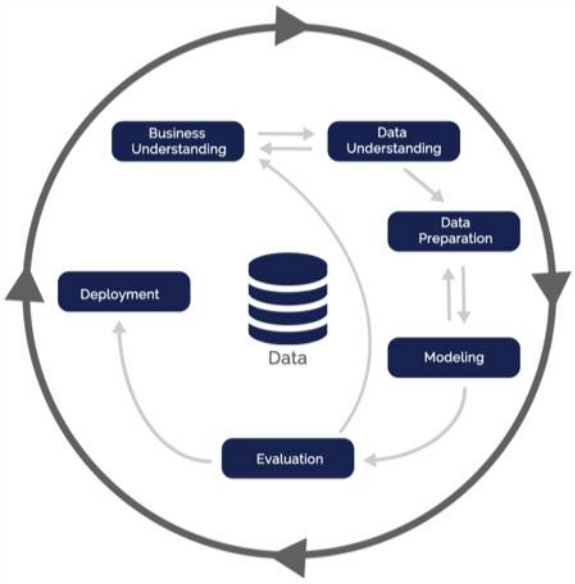
Dalam *data mining*, pengelompokan data juga bisa dilakukan. Tujuannya untuk mengetahui keputusan yang akan diambil selanjutnya atau tindak lanjut berikutnya yang dapat diambil. Semua hal tersebut bertujuan mendukung kegiatan operasional perusahaan sehingga tujuan akhir perusahaan diharapkan dapat tercapai [5].

2.1.5 Metode CRISP – DM

Pada penelitian ini metode yang digunakan dalam penyelesaian *data mining* adalah kerangka kerja *CRoss- Industry Standard Process for Data Mining* (CRSIP – DM). Metode ini dipilih karena menurut Prajitno dan Purwarianti dalam

penelitiannya yang berjudul "Prediksi Kinerja Penjualan Karya Musik Menggunakan Framework CRISP-DM (Studi Kasus: X Music Indonesia)", CRISP-DM menyediakan standar proses baku untuk data mining yang dapat diterapkan kedalam strategi pemecahan masalah umum pada bisnis ataupun pada unit peneltian. CRISP-DM membandingkan metodologi *data mining* lain lebih lengkap dan terdokumentasi dengan baik. Setiap fase terstruktur dan terdefinisi dengan jelas sehingga mudah diaplikasikan bahkan bagi pemula sekalipun. Dan penulis dimudahkan dalam pencarian referensi relevan karena penggunaan metode CRSIP – DM.

Cross–Industry Standard of Data Mining (CRSIP–DM) merupakan salah satu metodologi *data mining* yang digambarkan dengan model proses dan di dalamnya terdapat tugas – tugas yang harus dilakukan untuk melakukan proyek *data mining* [6]. Proses CRISP – DM terdiri dari enam tahap yang fleksibel. Beikut ini adalah tahap – tahap pada model proses CRSIP – DM :



Gambar 2.1 Model proses CRISP-DM

1. *Business Understanding*

Pada tahap *business understanding* berfokus pada pemahaman tujuan dan persyaratan proyek dari perspektif bisnis, kemudian mengubah pengtahuan

tersebut ke dalam definisi masalah *data mining* dan rencana awal dirancang untuk mencapai tujuan tersebut. Beberapa tahap kegiatan pada *business understanding* adalah sebagai berikut :

1. Melakukan identifikasi tujuan bisnis

   Tahap ini dihasilkan output berupa latar belakang dan tujuan bisnis.

2. Melakukan penilaian situasi

   Tahap ini dihasilkan output berupa inventori sumber daya, kendala dan resiko, serta biaya dan manfaat.

3. Menentukan tujuan data mining

   Tahap ini dihasilkan output berupa sasaran data mining.

4. Menyusun rencana proyek yang akan dilakukan

   Tahap ini dihasilkan output berupa rencana proyek dan penilaian awal untuk penggunaan alat dan teknik.

2. *Data Understanding*

Pada tahap *data understanding* dimulai dengan pengumpulan data awal, memahami data – data tersebut, dan menilai kualitas data. Kemudian dilakukan pendeteksian bagian dari data yang mungkin mengandung informasi tersembunyi. Tahap ini dibagi menjadi beberapa langkah sebagai berikut :

1. Pengumpulan data awal

   Pada tahap ini dikumpulkan data – data yang akan digunakan dalam *data mining*.

2. Mendeskripsikan data yang diperoleh

   Pada tahap ini akan dijelaskan data yang telah diperoleh di tahap pertama, Dalam penjelasan data akan dijelaskan format dari data, kuantitas data, jumlah *record* dan *field* dalam setiap tabel, dan sebagainya.

3. Mengeksplorasi data

Pada tahap ini akan dilakukan analisis data yang telah dijelaskan pada tahap sebelumnya. Salah satu cara mengeksplorasi data adalah menggunakan analisis statistik deskriptif [7].

4. Melakukan verifikasi terhadap kualitas data

   Pada tahap ini akan dilakukan pemeriksaan data dengan cara melihat apakah terdapat *missing* ataupun *outlier*. Metode yang digunakan dapat digunakan untuk mendeteksi *outlier* salah satunya yaitu dengan menggunakan *interquartile range* (IQR) [8].

3. *Data Preparation*

Pada tahap ini mencakup semua kegiatan yang diperlukan untuk membangun data final dari sekumpulan data mentah yang akan digunakan ke dalam pemodelan. Tahap data preparation data terdapat bebeberapa langkah sebagai berikut :

1. Melakukan pemilihan data

   Tahap ini bertujuan untuk menentukan data yang akan digunakan. Pemilahan data ini meliputi pemilihan atribut ataupun pemilihan *record*.

2. Melakukan pembersihan data

   Tahap ini bertujuan untuk menghilangkan data berdasarkan data yang telah diverifikasi pada tahap verifikasi kualitas data.

3. Melakukan pembangunan data

   Tahap ini akan disiapkan data yang telah dipilih dan dibersihkan untuk digunakan pada tahap pemodelan.

4. Mengintegrasikan data dari berbagai sumber

   Tahap ini akan digabungkan sesuai dengan kebutuhan untuk pemodelan.

5. Mentransformasikan data sehingga siap untuk diproses

Tahap ini struktur data akan diformat disesuaikan dengan data yang dibutuhkan untuk pemodelan.

4. *Modelling*

Pada tahap ini akan dilakukan pemilihan dan penerapan model yang sesuai berdasarkan tujuan yang akan dicapai. Tahap ini dibagi menjadi beberapa langkah sebagai berikut :

1. Memilih teknik pemodelan yang sesuai dengan data

   Tahap ini digunakan untuk memilih teknik pemodelan yang sesuai dengan permasalahan dan tujuan yang ingin dicapai.

2. Menjelaskan prosedur teknik pemodelan yang digunakan

   Tahap ini akan dijelaskan teknik pemodelan yang telah dipilih pada tahap pemilihan teknik pemodelan.

3. Melakukan penerapan teknik pemodelan

   Tahap ini akan dilakukan penerapan teknik pemodelan yang telah dipilih.

4. Menilai model yang dihasilkan

   Tahap ini dilakukan untuk penilaian terhadap pemodelan yang telah dipilih.

5. *Evaluation*

Setelah model terbentuk, perlu dilakukan evaluasi terhadap langkah – langkah yang dilakukan sebelumnya. Hal tersebut dilakukan untuk memastikan model yang dipilih telah sesuai dengan tujuan bisnis yang ditetapkan. Tahap ini dibagi menjadi beberapa langkah sebagai berikut :

1. Mengevaluasi model yang dihasilkan

   Tahap ini akan dievaluasi hasi dari pemodelan apakah sudah sesuai dengan tujuan bisnis atau tidak.

2. Mengkaji ulang proses – proses yang dilakukan

Tahap ini akan dilakukan pengkajian ulang yang menyeluruh proses *data minin*g yang telah dilakukan untuk memastikan adanya faktor penting atau tugas yang terabaikan.

3. Menentukan keputusan penggunaan hasil *data mining*

   Tahap ini akan diputuskan langkah selanjutnya, apakah hasil pemodelan akan dilanjutkan ke tahap deployment atau tidak.

6. *Deployment*

Pada tahap ini merupakan tahap implementasi hasil proses *data mining*. Model yang telah dihasilkan perlu diorganisir dan disajikan kepada aktor. Hal tersebut berguna untuk proses pengambil keputusan perusahaan. Tahap ini dibagi menjadi beberapa langkah sebagai berikut :

1. Menentukan rencana penerapan hasil *data mining*

   Tahap ini akan dihasilkan *output* berupa perencanaan yang akan dilakukan pada tahap *deployment*.

2. Menentukan rencana pengawasan dan pemeliharaan

   Tahap ini rencana yang telah dibuat akan dipantau dan dipelihara.

3. Membuat laporan akhir

   Tahap ini akan dihasilkan *output* berupa laporan akhir dan presentasi akhir.

4. Melakukan ulasan terhadap proyek yang telah dilakukan

   Tahap ini akan menghasilkan ulasan dokumentasi terhadap proyek yang telah dilakukan.

2.1.6 *Data Preprocessing*

Dalam *data mining*, kualitas data yang akan digunakan perlu diperhatikan. Ada beberapa faktor yang mempengaruhi kualitas data, antara lain keakuratan, keutuhan, konsisten, aktualitas, dan penafsiran. *Data preprocessing* dapat memperbaiki kualitas data, sehingga dapat meningkatkan keakuratan dan efisiensi

hasil dari *data mining*[9]. Beberapa kegiatan *data preprocessing* yang akan dilakukan dalam penelitian ini adalah sebagai berikut :

1. Penanganan *missing value*

Mising value pada data adalah masalah yang sering terjadi. *Missing value* pada data biasanya disebabkan oleh kesalahan *input* atau suatu atribut yang memang tidak memiliki sebuah nilai. *Missing value* pada dasarnya tidak bermasalah bagi keseluruhan data, apalagi jika jumlahnya hanya sedikit. Namun jika persentase data yang hilang tersebut cukup besar, maka perlu dilakukan pengujian apakah data yang mengandung banyak *missing value* masih layak diproses lebih lanjut ataukah tidak. Cara untuk menangani *missing value* adalah dengan menghitung nilai rata – rata dari objek yang hilang, nilai rata-rata tersebut dijadikan nilai penganti dari nilai yang hilang. Selain itu missing value juga bisa ditangani dengan memberikan nilai default pada setiap nilai yang hilang.

2. Penanganan *Outlier*

Outlier adalah data yang berbeda secara signifikan dari data yang lainnya atau dapat direpresentasikan sebagai *noise*. *Outlier* juga akan mempengaruhi hasil pemodelan data mining. Untuk itu *outlier* perlu dihapus agar menghasilkan model yang berkualitas. Cara yang dapat digunakan untuk menghapus *outlier* adalah dengan melakukan *smoothing*. Salah satu teknik untuk melakukan *smoothing* adalah dengan menggunakan metode *binning*. Metode *binning* membagi kumpulan data ke dalam beberapa partisi atau *bin*. Dimulai dengan mengurutkan setiap nilai pada sebuah atribut. Kemudian data yang sudah diurutkan dibagi ke dalam beberapa partisi atau *bin* yang memiliki frekuensi yang sama (*equal-frequency partitioning*). Setelah itu, data pada setiap *bin* diganti dengan nilai batas *bin* terdekat (*smoothing by bin boundaries*). Nilai batas *bin* merupakan nilai minimum dan maksimum pada setiap *bin*. Ada dua cara *smoothing* dalam metode *binning*, yaitu s*moothing by bin means* dan *smoothing by bin boundaries*.

Dalam *smoothing by bin means* dilakukan dengan mengubah setiap nilai dalam *bin* dengan nilai rata-rata dari *bin* tersebut. Sedangkan dalam *smoothing by bin*

boundaries setiap nilai dalam *bin* diubah menjadi batas bawah (minimum) dan batas bawah (maksimum) pada setiap *bin*.

2.1.7 Klasifikasi

Klasifikasi data adalah suatu proses yang menemukan properti-properti yang sama pada sebuah himpunan objek di dalam sebuah basis data, dan mengklasifikasikannya ke dalam kelas – kelas yang berbeda menurut model klasifikasi yang ditetapkan. Tujuan dari klasifikasi adalah untuk menemukan model dari *training set* yang membedakan atribut ke dalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan atribut yang kelasnya belum diketahui sebelumnya. Berikut adalah tahap – tahap yang harus dilakukan dalam penggunaan klasifikasi :

1. Konstruksi model

Dalam tahap ini dibutuhkan suatu *dataset* yang disebut *dataset training*, dan dijelaskan himpunan kelas yang telah terdefinisi sebelumnya. Tiap baris atau *sample* diasumsikan memiliki kelas yang telah didefinisikan sebelumnya. Tiap *sample* merupakan atribut kelas yang bersangkutan. Model merepresentasikan sebagai aturan klasifikasi, pohon keputusan atau formula matematika.

Data training proses konstruksi model memiliki beberapa variabel *predikor*. Variabel *predictor* yaitu variabel yang mempunyai kekuatan memprediksi suatu data baru.

2. Pengunaaan Model

Dalam tahap ini dilakukan pengklasifikasian objek yang belum diketahui kelasnya. Hal yang harus diperhatikan dalam pengklasifikasian objek di penggunaan model adalah sebagai berikut :

1. Estimasi keakuratan model, model yang telah dibangun akan digunakan untuk memprediksi data yang belum diketahui labelnya. Kemudian label dari *sample test* yang telah diketahui sebelumnya dibandingkan dengan hasil klasifikasi dari model terhadap *sample test* yang sama. Akan di hitung rata-rata akurasi dari model tersebut. Rata-rata akurasi adalah presentase dari

himpunan sampel yang secara benar digolongkan oleh model. Hal panting yang harus diperhatikan adalah himpunan tes merupakan himpunan *training* yang *independent*, karena jika tidak maka *over-fitting* akan terjadi. *Over-fitting* yaitu kecenderungan menganggap seluruh data hampir sama dengan *data training*. Hal ini akan mengakibatkan hilangnya nilai prediktif.

2. Jika keakuratan diterima, gunakan model untuk mengklasifikasi data yang label kelasnya belum diketahui.

2.1.8 *Naïve Bayes*

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teaorema *Bayes*. Teorema tersebut dikombinasikan dengan *naïve* yang mana diasumsikan kondisi antar atribut saling bebas.

2.1.9 *Naïve Bayes Classifier*

Naïve Bayes Classifier merupakan sebuah metode klasifikasi yang berakar pada teorema *Bayes*. Ciri utama dari *Naïve Bayes Classifier* adalah asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi atau kejadian. Berikut akan dijelaskan teorema *bayes* yang menjadi dasar dari metode tersebut.

1. Pada teorema Bayes, bila terdapat dua kejadian yang terpisah (misalkan A dan B), maka teorema *bayes* dirumuskan sebagai berikut:

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

2. Teorema *bayes* sering juga dikembangkan mengingat berlakunya hukum probabilitas total, menjadi seperti berikut:

$$P(A|B) = \frac{P(A)P(B|A)}{\sum_{i=1}^{n} P(A_i|B)}$$

3. Untuk menjelaskan teorema *naïve bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang

cocok bagi sampel yang akan dianalisis tersebut. Karena itu, teorema *bayes* di atas disesuaikan sebagai berikut:

$$P(C|F_1,..,F_n) = \frac{P(C)P(F_1,..,F_n|C)}{P(F_1,..,F_n)}$$

4. Dimana variable C merepresentasikan kelas, sementara variable F_1,……., F_n merepresentasikan karakteristik – karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Maka rumus tersebut menjelaskan bahwa peluang masuknya sampel dengan karakteristik tertentu dalam kelas C (posterior) adalah peluang munculnya kelas C (sebelum masuk 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 (3) dapat ditulis secara sederhana sebagai berikut:

$$Posterior = \frac{prior\ x\ likehood}{evidence}$$

5. Nilai $evidence$ selalu tetap untuk setiap kelas pada satu sampel. Nilai dari $Posterior$ tersebut yang 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 $P(F_1,..,F_n|C)$ menggunakan aturan perkalian, menjadi sebagai berikut :

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

6. 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. Di sinilah digunakan asumsi independensi yang sangat tinggi (naif), bahwa masing-masing petunjuk

$(F_1, F_2 \ldots . . F_n)$ saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut :

$$P(F_i - F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

Untuk i ≠ j, sehingga

$$P(F_i|C, F_j) = P(F_i C)$$

7. Dari persamaan di atas dapat disimpulkan bahwa asumsi independensi naïf tersebut membuat syarat peluang menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran $(F_1, F_2 \ldots . . F_n|C)$ dapat disederhanakan menjadi seperti berikut :

$$P(F_1 \ldots . F_n|C) = P(F_1|C)P(F_2|C) \ldots \ldots P(F_n|C)$$

$$P(F_1 \ldots . F_n|C) = \prod\nolimits_{i=1}^{n} P(F_i|C)$$

8. Dengan kesamaan di atas, persamaan teorema bayes dapat dituliskan sebagai berikut :

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

$$P(C|F_1 \ldots . F_n) = \frac{P(C)}{Z} \prod\nolimits_{i=1}^{n} P(F_i|C)$$

Persamaan di atas merupakan model dari teorema naïve bayes yang selanjutnya akan digunakan dalam proses klasifikasi. Adapun Z mempresentasikan $evidence$ yang nilainya konstan untuk semua kelas pada satu sampel.

2.1.10 Klasifikasi dengan *Naïve Bayes Classifier*

Klasifikasi adalah proses untuk menentukan model atau fungsi yang dapat menjalankan atau membedakan konsep, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek. Maka dari itu kelas yang ada tentu lebih dari satu. Penentuan kelas dilakukan dengan cara membandingkan nilai probabilitas suatu sampel

berbeda di kelas yang satu dengan nilai probabilitas suatu sampel berada di kelas yang lain.

Dengan persamaan teorema *naïve bayes* yang telah diturunkan di subbab 2.2.10, didapatkan nilai $P(F_1, \dots., F_n)$, yaitu nilai peluang suatu sampel dengan karakteristik $F_1, \dots. F_n$ berada dalam kelas C, atau dikenal dengan istilah $Posterior$. Umumnya kelas yang ada tidak hanya satu, melainkan lebih dari satu.

Sebagai contoh, seorang ahli statistik sedang mengklasifikasi sampel tikus jenis kelaminnya. Maka kelas yang terbentuk yaitu jantan atau betina. Suatu sampel ayam akan diklasifikasi ke dalam satu kelas saja, entah itu jantan atau betina, dengan melihat petunjuk – petunjuk yang ada.

Penentuan kelas yang cocok bagi suatu sampel dilakukan dengan cara membandingkan nilai $Posterior$ untuk masing-masing kelas. Dan mengambil kelas dengan nilai $Posterior$ yang tinggi. Secara matematis klasifikasi dirumuskan sebagai berikut :

$$C_{NB} = argmax_{c \in C} P(c) \prod_{i=1}^{n} P(F_i|c)$$

Dengan c yaitu variable kelas yang tergabung dalam suatu himpunan kelas C.

Dapat dilihat bahwa rumusan di atas tidak memuat nilai $Evidence$ (Z). hal ini disebabkan karena $evidence$ memiliki nilai yang positif dan tetap untuk semua kelas sehingga tidak mempengaruhi perbandingan nilai $Posterior$. Karena itu, faktor Z ini dapat dihilangkan. Perlu menjadi perhatian pula bahwa metode *Naïve Bayes Classifier* ini dapat digunakan bila sebelumnya telah tersedia data yang dijadikan acuan untuk melakukan klasifikasi.

2.1.11 Evaluasi dan Validasi Model

Pengukuran akurasi dalam penelitian ini diperlukan untuk mengetahui seberapa besar persentase akurasi prediksi yang dilakukan. Untuk mengukur akurasi model maka dilakukan evaluasi dan validasi menggunakan teknik :

1. *Confusion matrix*

Confusion matrix merupakan perhitungan yang digunakan untuk mengevaluasi performa sebuah algoritma. *Confusion matrix* akan menghasilkan perhitungan prediksi dan kondisi sesungguhnya berdasarkan data hasil algoritma yang digunakan. Akurasi dalam klasifikasi adalah presentase ketetapan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. *Confusion matrix* berisi informasi aktual (*actual*) dan prediksi (*predicted*) pada sisitem klasifikasi [16].

Tabel 2.1 *Confussion Matrix*

Kelas Aktual	**Kelas hasil prediksi**	
	Ya	**Tidak**
Ya	TP	FN
Tidak	FP	TN
Jumlah	P’	N’

Empat istilah sangat penting untuk memahami semua ukuran evaluasi dalam table tersebut adalah sebagai berikut :

1. TP atau *True Positives* adalah jumlah tuple positif yang dilabeli dengan benar oleh *classifier*. Yang dimaksud tuple positif adalah tuple aktual yang berlabel positif.
2. TN atau *True Negatives* adalah jumlah tuple negatif yang dilabeli dengan benar oleh *classifier*. Yang dimaksud tuple negatif adalah tuple aktual yang berlabel negatife
3. FP atau *False Positives* adalah tuple negatif yang salah dilabeli oleh *classifier*.
4. FN atau *False Negatives* adalah jumlah tuple positif yang salah dilabeli oleh *classifier*.

Untuk evaluasi yang akan digunakan dengan *error rate* atau *misclassification rate* yang juga dapat menghasilkan akurasi model klasifikasi.

$$Akurasi = \frac{FP + FN}{P + N}$$

2. Kurva ROC (*Reciever Operating Characteristic*)

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertical [13]. *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan. AUC dihitung menggunakan rumus [14]:

$$\theta^r = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\,(Xi^r, Xj^r)$$

Dimana

$$\psi(X,Y) = \begin{cases} 1\; Y < X \\ \frac{1}{2}\; Y = X \\ 0\; Y > X \end{cases}$$

Performance keakurasian AUC dapat diklasifikasikan menjadi lima kelompok yaitu [12] :

0.90 – 1.00 = *Exellent Clasification*

0.80 – 0.90 = *Good Clasification*

0.70 – 0.80 = *Fair Clasification*

0.60 – 0.70 = *Poor Clasification*

0.50– 0.60 = *Failure*