BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Untuk perbandingan dan referensi berkaitan dengan penelitian mengenai algoritma pengklasifikasian data, dibutuhkan acuan berkaitan dengan penelitian sebelumnya. Di bawah ini adalah penelitian terdahulu yang berhubungan dengan penelitian "Perbandingan Algoritma Klasifikasi *Naïve Bayes*, C4.5 dan KNN Untuk Menentukan Perokok Aktif dan Perokok Pasif ".

Penelitian pertama oleh Fuad Nurhasan dkk tahun 2018 berjudul **"Perbandingan Algoritma C4.5, KNN, dan *Naïve Bayes* untuk Penentuan Model Klasifikasi Penanggung Jawab BSI Enterpreneur Center"**. Masalah yang ada pada penelitian tersebut adalah karyawan yang berada di BSI Enterpreneur Center pada saat itu belum dapat memenuhi kebutuhan sebagai penanggung jawab BSI Entrepreneur Center untuk ditempatkan pada masing-masing kampus Universitas Bina Sarana Informatika. Oleh sebab itu, dibutuhkan sistem dalam menemukan sumber daya manusia yang sesuai sebagai penanggung jawab BSI Entrepreneur Center pada tiap-tiap kampus Universitas Bina Sarana Informatika. Pengujian ini menggunakan algoritma C4.5, KNN, dan *Naïve Bayes* untuk menentukan mana akurasi klasifikasi yang terbaik diantara ketiga algoritma tersebut. Dalam penelitian ini dihasilkan metode yang sesuai dengan mengetahui nilai akurasi tertinggi, dengan menggunakan algoritma C4.5 memperoleh nilai akurasi 73,33 % dan metode KNN memperoleh nilai akurasi 70 % dan metode *Naïve Bayes* memperoleh nilai akurasi sebesar 80%. Sehingga hasil kesimpulan dari ketiga metode tersebut maka algoritma yang paling sesui untuk digunakan mengklasifikasikan menjadi penanggung jawab BSI enterpreneur center pada tiap-tiap kampus Universitas Bina Sarana Informatika adalah memakai algoritma *Naïve Bayes* yang menghasilkan nilai akurasi tertinggi.

Penelitian kedua oleh Annisa Putri Ayudhitama dan Utomo Pujianto pada tahun 2020 dengan judul **"Analisa 4 Algoritma dalam Klasifikasi Penyakit Liver Menggunakan Rapidminer"**. Masalah yang ada pada penelitian ini adalah

seseorang seringkali tidak sadar atau terlambat menyadari terkena penyakit liver sebagai akibatnya saat diperiksa penyakit liver telah parah, akan lebih baik bila dilakukan penanganan lebih awal dengan mengetahui gejala-gejala yang diderita. *Data mining* bisa membantu deteksi penyakit liver dengan lebih mudah terutama untuk membantu dokter dalam menentukan apakah pasien terkena penyakit liver atau tidak, dengan tanda-tanda mendekati penyakit liver. Proses diagnosa penyakit liver dilakukan menggunakan proses klasifikasi dan hasilnya berupa pasien tadi menderita liver atau tidak. Maka, dengan demikian dilakukanlah penelitian tersebut untuk menentukan mana diantara 4 algoritma yang mampu mendapatkan akurasi yang terbaik dalam klasifikasi penyakit Liver ini. Kemudian didapatkan lah metode yang terbaik untuk mengklasifikan penyakit Liver ini adalah *Decision Tree* dengan hasil akurasi sebesar 72,89%. Selain akurasinya paling tinggi, *Decision Tree* juga mampu mengklasifikasi para pasien yang menderita penyakit liver dengan jumlah lebih banyak sehingga terhitung akurat. Selama ini banyak algoritma yang memiliki nilai akurasi tinggi tetapi tidak mampu untuk mengklasifikasi dengan benar bahkan banyak yang mendeteksi pasien tidak mengalami liver padahal data aslinya terkena liver. Pada kurva ROC hanya algoritma *Decision Tree* yang memiliki grafik sumbu Y mendekati nilai 1.00 yang dikategorikan sebagai "*Excellent*" klasifikasi.

Penelitian Ketiga ini ditulis oleh Choirul Anam dan Harry Budi Santoso pada tahun 2018 berjudul **"Perbandingan Kinerja Algoritma C4.5 dan *Naïve Bayes* untuk Klasifikasi Penerima Beasiswa".** Permasalahan di penelitian ini adalah Penentuan penerima beasiswa wajib mempertimbangkan banyak faktor sebagai penentu dalam memastikan pihak penerima memang berhak mendapatkan beasiswa. Metode *data mining* untuk klasifikasi bisa digunakan dalam membantu mempercepat dan meningkatkan ketepatan dalam penentuan penerima beasiswa. Perbandingan kinerja algoritma C4.5 dan *Naïve Bayes* bertujuan mengukur tingkat akurasi dan lama waktu proses (*execution time*) dari setiap algoritma agar mendapatkan algoritma paling baik yang dapat diterapkan dalam membantu proses penentuan penerima beasiswa. Hasil pengujian terhadap model klasifikasi menghasilkan tingkat akurasi dari model algoritma C4.5 dengan nilai akurasi 96.4% dan tingkat akurasi model algoritma *Naïve Bayes* dengan nilai akurasi

95.11%. Berdasarkan hasil penerapan algoritma C4.5 dan *Naïve Bayes* untuk model klasifikasi penerima beasiswa untuk penelitian ini menghasilkan model algoritma C4.5 memiliki hasil kinerja yang lebih baik dari *Naïve Bayes*.

Penelitian keempat ditulis oleh Rony Asmara, Jefri Setiawan dan Meilany Nonsi Tentua tahun 2020 berjudul **"Komparasi Algoritma C4.5, *Naïve Bayes* dan K-*Nearest Neighbor* Pada Pasien yang Terkena Penyakit Diabetes"**. Sebanyak 197 pasien dalam dataset ini artinya 50% sampel acak dari pasien dengan penyakit retinopati diabetik yang mempunyai resiko tinggi seperti pendefinisian *Diabetic Retinopathy Study* (DRS). Setiap pasien mempunyai satu mata acak perawatan laser dan mata lainnya tidak menerima pengobatan, dan mempunyai dua pengamatan dalam sekumpulan data. Untuk mengatasi permasalahan ini penulis menggunakan perbandingan 3 algoritma klasifikasi untuk menangani gangguan yang ada pada data ini yaitu algoritma *Naïve Bayes,* C4.5 dan KNN. Diantara ketiga algoritma tersebut C4.5 mendapatkan akurasi tertinggi dengan tingkat akurasi 67%, dibandingkan *Naïve Bayes* 58% dan KNN 65% dikarenakan C4.5 memprediksi lebih banyak/tinggi dibandingkan algoritma lainnya. Kemudian, begitupula dengan precision yang dihasilkan kedudukan masih tetap sama C4.5 memperoleh tingkat presisi 66% lebih tinggi dibandingkan *Naïve Bayes* yang mendapatkan 62% dan KNN 63%. Lalu tingkat Recall C4.5 lebih tinggi berada di angka 93% dibandingkan dengan *Naïve Bayes* dan KNN yang berada di 80%. Maka Metode C4.5 adalah yang paling akurat dalam mengklasifikasikan data terhadap Pasien Yang Terkena Penyakit Diabetes dengan tingkat *accuracy* mempunyai nilai sebesar 67%, tingkat *precision* mempunyai nilai sebesar 66% dan *Recall* mempunyai nilai sebesar 93%.

Penelitian kelima ditulis oleh Syamsul Bahri, Dwi Marisa Midyanti dan Rahmi Hidayati pada tahun 2018 dengan judul **"Perbandingan Algoritma *Naïve Bayes* dan C4.5 Untuk Klasifikasi Penyakit Anak "**. Rentannya penyakit yang dapat menyerang anak membuat orang sekitar terutama orang tua sering tidak mengetahui gejala penyakit yang muncul pada anaknya, karena itu peneliti membuat sebuah terobosan untuk mengetahui gejala gejala ini dengan membuat sebuah klasifikasi berdasarkan gejala-gejala yang sering timbul. Metode klasifikasi yang dipakai oleh peneliti adalah algoritma C4.5 dan *Naïve Bayes*. Pada penelitian

yang dilakukan oleh penulis diperoleh hasil perbandingan algoritma terbaik antara kedua algoritma ini yaitu menggunakan Algoritma C4.5 dengan hasil akurasi sebesar 90.00% sedangkan algoritma *Naïve Bayes* mendapatkan hasil akurasi 89.58%, metode yang digunakan pada penelitian ini adalah menggunakan metode K-*Fold Cross Validation*.

Tabel 2.1 Penelitian Terkait

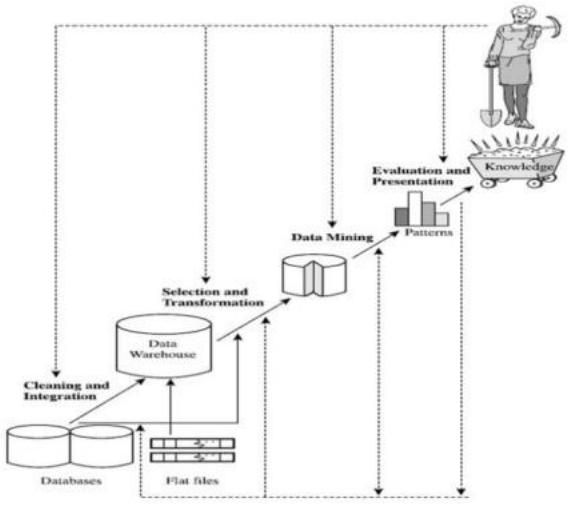
No	Penulis	Judul Penelitian	Tahun	Metode	Masalah	Hasil
1	Fuad Nurhasan, Noer Hikmah, Dwi Yuni Utami [12]	Perbandingan Algoritma C4.5, KNN, dan *Naïve Bayes* Untuk Penentuan Model Klasifikasi Penanggung Jawab BSI ENTERPRENEUR CENTER	2018	CRISP-DM, C4.5, KNN, *Naïve Bayes*, T-*Test*, *Cross Validation*	Membandingkan 3 Algoritma untuk menentukan Model Klasifikasi terbaik pada Kelas penanggung jawab BSI Enterpreneur Center	Metode yang sesuai dengan menentukan nilai akurasi yang lebih tinggi, adalah dengan menggunakan metode *Naïve Bayes* dengan Akurasi 80%, C4.5: 73.33%, KNN: 70%.
2	Choirul Anam & Harry Budi Santoso [14]	Perbandingan Kinerja Algoritma C4.5 dan *Naïve Bayes* Untuk Klasifikasi Penerima Beasiswa	2018	CRISP-DM, *Naïve Bayes*, C4.5	Penentuan penerima beasiswa wajib mempertimbangkan banyak faktor sebagai penentu dalam memastikan pihak penerima memang berhak mendapatkan beasiswa	Hasilnya didapatkan tingkat akurasi menggunakan C4.5 :96.4% dan *Naïve Bayes* :95.11%. Maka Metode C4.5 adalah metode yang lebih cocok dibandingkan *Naïve Bayes* pada permasalahan ini.
3	Syamsul Bahri, Dwi Marisa Midyanti,	Perbandingan Algoritma *Naïve Bayes* dan C4.5 Untuk Klasifikasi Penyakit Anak	2018	C4.5, *Naïve Bayes*, K-*Fold Cross Validation*	Belum diketahuinya Algoritma yang akurat dalam	Penelitian ini menggunakan metode k-*fold cross validation* didapatkan tingkat akurasi 90% pada C4.5 dan 89.58% menggunakan *Naïve Bayes*.

	Rahmi Hidayati [16]				mendeteksi penyakit anak	
4	Annisa Putri Ayudhitama, Utomo Pujianto [13]	Analisa 4 Algoritma dalam Klasifikasi Penyakit Liver Menggunakan RapidMiner	2020	*Compare* ROCs, *Confusion Matrix, Naïve Bayes*, KNN, *Decision Tree, Neural Network*	Seseorang sering tidak sadar atau terlambat mengetahui penyakit liver sehingga waktu diperiksa oleh dokter penyakit liver sudah parah.	Metode yang terbaik untuk mengklasifikan penyakit Liver ini adalah *Decision Tree* dengan hasil akurasi sebesar 72,89%.
5	Rony Asmara, Jefri Setiawan, Meilany Nonsi Tentua [15]	Komparasi Algoritma C4.5, *Naïve Bayes* dan K-*Nearest Neighbor* Pada Pasien yang Terkena Penyakit Diabetes	2020	Pengumpulan Data, *PreProcessing Data*, C4.5, *Naïve Bayes*, K-*Nearest Neighbor*	Sebanyak 197 pasien dalam dataset ini artinya 50% sampel acak dari pasien dengan penyakit retinopati diabetik yang mempunyai resiko tinggi seperti pendefinisian *Diabetic Retinopathy Study* (DRS).	Metode C4.5 adalah Metode yang tepat untuk mengatasi permasalahan pada data ini dikarenakan memiliki Tingkat Akurasi(67%), Presisi(66%) dan *Recall*(93%) yang paling tinggi diantara 2 metode yang lain

2.2 Dasar Teori

2.2.1 Data *Mining*

Data *Mining* adalah proses pengumpulan data pengolahan data untuk mendapatkan informasi yang penting berkaitan dengan data tersebut. Proses mengekstraksi serta mengidentifikasi informasi data tersebut dapat dilakukan menggunakan perangkat lunak dengan bantuan memakai matematika, statistik, *artificial Intelegent* dan teknik *machine learning* [17]. Data *mining* umumnya memakai proses eksplorasi dan analisis sejumlah besar data dalam menemukan bentuk atau pola serta aturan yang bermakna [18]. Data *mining* memperbaharui sejumlah besar data menjadi pengetahuan. Data *Mining* menganalisis kumpulan data yang diamati untuk mengidentifikasi data yang tidak diketahui dan memperpendek data menggunakan aturan baru yang bisa dipahami dan bermanfaat bagi pemilik [19].



Gambar 2.1 Proses Data *Mining* [20].

Penjelasan proses data *mining* pada Gambar 2.1 sebagai berikut:

1. Data *Cleaning*: untuk membersihkan *noise* dan data yang tidak cocok.
2. Data *Integration*: untuk memadukan serta mengintegrasikan sejumlah data referensi.

3. Data *Selection*: untuk memilih data yang sesuai dari basis data untuk dianalisa.
4. Data *Transformation*: untuk mentransformasikan data *ringkasan* ataupun fungsi agregasi.
5. Data *Mining*: Ini adalah proses penting menggunakan metode untuk mengekstrak bentuk data implisit.
6. *Pattern Evaluation*: mengenali bentuk-bentuk untuk merepresentasikan pengetahuan sesuai beberapa nilai menarik.
7. *Knowledge Presentation*: teknik representasi serta memvisualisasi data yang dipakai dalam presentasi pengetahuan yang dihasilkan untuk pengguna.

2.2.2 *Naïve Bayes*

Naïve Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (aturan Bayes) dengan asumsi independensi (ketidak tergantungan) yang kuat (naif). Dengan kata lain, dalam *Naïve Bayes* model yang digunakan adalah "model fitur independent" [21]. *Naive Bayes* didasarkan pada teorema Bayes dan memiliki kemampuan klasifikasi yang sama dengan pohon keputusan dan jaringan saraf. *Naïve Bayes* telah terbukti mencapai tingkat akurasi dan kecepatan yang tinggi saat diterapkan pada database dengan data besar.

Yang menguntungkan dari metode klasifikasi ini adalah naïve bayes membutuhkan sedikit data training dalam menentukan parameter prediktif. Alih-alih menentukan matriks kovarians penuh, maka hanya varians fitur yang tentukan karena fitur yang bebas. Untuk teks ulasan 'd' dan kategori 'c' (positif, negatif), kemungkinan bersyarat untuk tiap kategori yang ditambahkan adalah P (c | d) [22]. Berdasarkan teorema Bayes, angka ini bisa ditentukan dengan memakai persamaan [23]:

$$P(c|d) = \frac{P(d|c)\times P(c)}{P(c)} \qquad (2.1)$$

Persamaan tersebut dinyatakan menjadi [21] :

$$P(y) = \underset{y\in[1,,,K]}{argmax} \prod_{i=1}^{n} p(x_i|y_k) \quad (2.2)$$

Keterangan:

P = Probabilitas

y = Kategori atau Kelas

x = nilai total

c = Kelas

d = Data dari kelas yang tidak diketahui

Naïve Bayes tidak hanya menangani data menggunakan tipe *polynomial*, tetapi juga menangani data menggunakan tipe numerik atau kontinyu [24]. Saat diterapkan pada data kontinyu, diasumsikan bahwa nilai kontinu diasosiasikan dengan tiap-tiap kelas yang didistribusikan berdasarkan distribusi *Gaussian*. Misalnya data *training* berisi atribut kontinu. Segementasikan data berdasarkan kelas, setelah itu melakukan perhitungan *mean* dan varian dari tiap kelas yang dinyatakan dengan persamaan (2.3) berikut ini [25]:

$$P\ (x_i|y) = \frac{1}{\sqrt{2\pi\sigma y}} exp^{\frac{(x_i-u_{ij})^2}{2\sigma^2 y}} \quad (2.3)$$

Keterangan:

P = Probabilitas

x_i = data atribut ke-i

y = kelas (label)

π = nilai phi (3.14 atau 22/7)

σ = Standar Deviasi

e = Eksponensial

μ = nilai rata-rata

Tahapan metode *Naïve Bayes* yang digunakan di penelitian ini adalah dimulai pembacaan data latih *Bio signal* yang telah dibagi lewat *cross validation* serta tidak terdapat *missing value*. Setelah itu diteruskan dengan proses pemodelan yang menentukan nilai rata-rata serta standar deviasi tiap atribut yang terdapat pada data set

bio signal. Untuk menentukan nilai rata-rata dan standar deviasi dilaksanakan penentuan secara terpisah sesuai dengan kategori/kelas 25 (aktif atau pasif) yang terdapat di dataset. Setelah itu menentukan proses solusi memakai data uji.

2.2.3 C4.5

Algoritma C4.5 merupakan pengembangan dari algortima ID3 yang digunakan dalam membentuk suatu pohon keputusan [26]. Algoritma C.45 digunakan untuk 10 klasifikasi data. Proses pengklasifikasian dilaksanakan memakai metode pohon keputusan [27]. Pohon keputusan merupakan penggambaran permodelan berdasarkan persoalan yang terdiri dari berbagai macam keputusan menuju ke arah solusi. Langkah-langkah dalam pembuatan pohon keputusan memakai algoritme C4.5 adalah:

1. Mempersiapkan data training yaitu data yang telah dipisah-pisah dalam kategori/kelas tertentu.
2. Mencari akar dari pohon memakai cara penentuan nilai gain yang paling tinggi dari setiap atribut atau dengan mengambil nilai *index entropy* paling rendah. *Index entropy* ditentukan dengan rumus.

   $$Info\ (D) = -\sum_{i=1}^{m} Pi \log_2(pi) \qquad (2.4)$$

   Keterangan:

   Pi = atribut
3. Mencari nilai entropi tiap atribut dilaksanakan sebanyak jumlah data yang terpartisi pada atribut tersebut. Mencari nilai entropi dilaksanakan dengan menggunakan persamaan:

   $$Info\ A(D) = \sum_{J=1}^{V} \frac{|D_j|}{|D|} \times infoD_j \qquad (2.5)$$

   Keterangan:

   |D| = atribut ke-n

   $|D_j|$= nilai total kasus

   Info D_j = nilai entropy atribut ke-n

4. Menentukan nilai *gain* memakai rumus:

$$Gain\ (A) = \ info\ (D) - \ info_a(D) \qquad (2.6)$$

5. Melakukan kembali langkah nomer 2 sehingga semua data *record* terpatisi. Pematisian pohon keputusan berhenti pada saat:
   a. Semua tupel di dalam *record* dalam simpul N menghasilkan kelas yang sama.
   b. Tidak terdapat atribut di *record* yang terpatisi lagi.
   c. Tidak terdapat *record* pada cabang yang kosong.

2.2.4 K-NN (K-*Nearest Neighbor*)

K-NN merupakan algoritma klasifikasi yang memanfaatkan konsep jarak terpendek antara data yang dievaluasi dengan K tetangga terdekatnya, dan kelas dengan jarak terpendek dan jumlah kelas terbanyak akan menjadi kelas tempat data yang dievaluasi berada.

Langkah-langkah dalam penerapan metode K-NN adalah sebagai berikut:

1. Membuat dokumen X dari semua sampel pelatihan untuk membentuk vektor fitur yang sama (X_1, X_2, X_3, …, X_n).
2. Menentukan nilai kesamaan semua sampel latih dengan dokumen X. Ambil dokumen ke-I (d_{i1}, d_{i2}, d_{i3}, …., d_{im}).

$$sim\ (X, d_i) = \frac{\sum_{j=1}^{m} X_f \cdot D_y}{\sqrt{\left(\sum_{j=1}^{m} X_j\right)^2} \cdot \sqrt{\left(\sum_{j=1}^{m} d_y\right)^2}} \qquad (2.7)$$

3. Menentukan k sampel dengan nilai lebih besar dari kesamaan N pada SIM (X, d_i), (i=1, 2, …, N), dan menerapkannya sebagai kumpulan K-NN dari X. Setelah itu menentukan nilai probabilitas X ke tiap-tiap kategori. Berikut adalah rumusnya:

$$\sum_{d_i \in KNN} SIM(X, d_i), y(d_i, C_j) \qquad (2.8)$$

Dimana, $y(d_i,C_j)$ adalah fungsi atribut kategori yang memenuhi persamaan berikut:

$$y(d_i, C_j) = \begin{cases} 1, d_i \in C_j \\ 0, d_i \notin C_j \end{cases} \quad (2.9)$$

Uji dokumen X untuk mengetahui kategorinya dengan melihat P (X, C_j) terbesar [28].

2.2.5 *Confussion Matrix*

Confusion matrix juga disebut *error matrix*. C*onfusion matrix* akan menginformasikan perbandingan hasil klasifikasi yang dijalankan sistem dengan hasil klasifikasi sesungguhnya. *Confussion matrix* merupakan tabel matriks yang menggambarkan kinerja model klasifikasi melalui serangkaian data testing yang nilainya sudah diketahui [30]. Beberapa *performance matrix* dari *confussion matrix* yang terkenal adalah *accuracy, precision*, dan *recall*.

Accuracy menunjukkan keakuratan model bisa mengklasifikasikan dengan sesuai. *Accuracy* adalah rasio prediksi yang benar (positif dan negatif) untuk semua data. *Accuracy* adalah seberapa dekat nilai prediksi dengan nilai aktual. Nilai *accuracy* dalam persen dinyatakan dalam persamaan yaitu:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} x\ 100\% \quad (2.10)$$

Precision menjelaskan keakuratan model yang diinginkan dengan nilai prediksi yang ditunjukkan oleh model serta merupakan rasio prediksi benar positif dari semua hasil yang diprediksi secara positif. Nilai *precision* persen yang dinyatakan dalam persamaan yaitu:

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (2.11)$$

Recall menggambarkan keberhasilan model dalam mendapatkan informasi serta merupakan rasio prediksi benar positif yang dibandingkan dengan semua data yang positif. Nilai recall dapat didapatkan dalam persamaan:

$$\text{Recall} = \frac{TP}{TP+FN}\ x\ 100\% \quad (2.12)$$

Tabel 2.2 *Confussion Matrix*

		Actual Values	
		1 (*Positive*)	**2 (*Negative*)**
Predicted Values	**1 (*Positive*)**	TP	FP
	2 (*Negative*)	FN	TN

Keterangan dari Tabel 2.2 *Confussion matrix* yang menjelaskan nilai:

1. *True Positive* (TP), adalah data positif yang diprediksi positif (benar). Contohnya *sentiment analysis* dalam data berisi kata kunci *sentiment positive* serta setelah diprediksi memang benar mempunyai tujuan untuk melakukan *sentiment* berbentuk *positive*.
2. *True Negative* (TN), adalah data negatif yang diprediksi negatif (benar). Contohnya *sentiment analysis* dalam data tidak mempunyai kata kunci *negative* setelah diprediksi benar tidak mempunyai tujuan untuk melakukan sentiment berbentuk positif.
3. *False Positive* (FP), adalah data negatif namun diprediksi sebagai data positif (salah atau eror). Contohnya *sentiment analysis* pada data tidak mempunyai kata kunci negatif namun setelah diprediksi mempunyai tujuan untuk melakukan sentiment berbentuk negatif.
4. *False Negative* (FN), merupakan data positif namun diprediksi sebagai data negatif (salah atau error). Misalnya *sentiment analysis* pada data mengandung kata kunci negatif namun setelah di prediksi tidak mempunyai tujuan untuk melakukan *sentiment* berbentuk negatif.