BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Penelitian terkait dengan penelitian ini adalah sebagai berikut :

Didik Srianto dan Edy Mulyanto (2016) melakukan perbandingan algoritma K-Nearest Neighbor dan Naive Bayes untuk klasifikasi tanah layak tanam pohon jati. Setiap algoritma dilakukan pengujian yang berbeda, pada algoritma K-Nearest Neighbor dilakukan pengujian nilai dari parameter k (jarak antar data) dengan nilai 1, 3, 5, 7, dan 9. Untuk algoritma Naive Bayes dilakukan pengujian sebanyak 5 kali dengan precentage split yang dibagi menjadi data training dan data testing yaitu 95% dan 5%, 85% dan 15%, 75% dan 25%, 65% dan 35%, 55% dan 45%. Hasil yang didapatkan pada penelitian ini adalah bahwa algoritma K-Nearest Neighbor medapatkan nilai akurasi yang lebih tinggi dengan nilai rata - rata akurasi sebesar 96.66% dan dibandingkan dengan Algoritma Naive Bayes dengan nilai rata - rata akurasi sebesar 82.63%.

Muqorobin, Kusrini, dan Emha Taufiq Luthfi (2019) melakukan penelitian untuk memprediksi keterlambatan pembayaran sumbangan pembinaan pendidikan sekolah dengan menggunakan algoritma Naive Bayes dengan menambahkan fitur seleksi *Informaton Gain*. Penelitian ini melakukan pengujian dengan menggunakan 30 dataset yang dibagi menjadi 2 yaitu 20 data training dan 10 data testing dengan menggunakan 2 skenario pengujian. Hasil yang didapatkan bahwa dengan menambahkan fitur seleksi information gain meningkatkan tingkat akurasi pada algoritma Naive Bayes dengan tingkat akurasi sebesar 90% dan tanpa menambahkan *feature selection* pada algoritma Naive Bayes mendapatkan akurasi sebesar 80%.

Ivandari, Tria Titiani Chasanah, dkk (2017) melakukan seleksi atribut dengan menggunakan Information Gain untuk meningkatkan performa klasifikasi dalam persetujuan kredit dengan menggunakan algoritma K-Nearest Neighbor. Pengujian algoritma K-Nearest Neighbor dilakukan dengan menguji nilai dari parameter k (jarak

antar data) dengan nilai 1, 3, 5, 7, 9, 11, 13, 15, dan 17 dengan menggunakan Rapid Miner serta melakukan pengujian atribut yang telah dilakukan *Feature Selection Information Gain*. Hasil yang didapatkan bahwa dengan menambahkan *Information Gain*, nilai akurasi algoritma K-Nearest Neighbor menjadi meningkat. pada *dataset UCI credit approval* sebelum menambahkan *Feature Selection* nilai akurasi yang didapat sebesar 74.93% dan pada *dataset credit card customer* tanpa menambahkan *Feature Selection* mendapatkan nilai akurasi sebesar 91.52%. setelah menambahkan *Feature Selection Information Gain* nilai akurasi pada *dataset UCI credit approval* mendapatkan akurasi sebesar 82.46% dan pada *dataset credit card customer* mendapatkan nilai akurasi sebesar 94.78%.

Tabel 2.1 Kajian Terkait

No	Penulis	Judul	Keterangan
1	Didik Srianto dan Edy Mulyanto (2016)	Perbandingan K-Nearest Neighbor dan Naive Bayes untuk Klasifikasi Tanah Layak Tanam Pohon Jati	Pada penelitian ini algoritma K-Nearest Neighbor mendapatkan nilai akurasi yang lebih tinggi dengan nilai k=3 dengan nilai akurasi 97.14% dan algoritma Naive Bayes dengan precentage split 75% dan 25 % mendapatkan nilai akurasi sebesar 88.46%.

No	Penulis	Judul	Keterangan
2	Muqorobin, Kusrini, Emha Taufiq Luthfi (2019)	Optimasi Metode Naive Bayes dengan *Feature Selection Information Gain* Untuk Prediksi Keterlambatan Pembayaran Sumbangan Pembinaan Pendidikan Sekolah	Pada penelitian ini bahwa akurasi yang didapatkan untuk prediksi keterlambatan Pembayaran sumbangan pembinaan pendidikan sekolah dengan menggunakan algoitma naive bayes dengan menambahkan *Feature Selection* mendapatkan nilai akurasi yang lebih tinggi sebesar 90% dan tingkat akurasi tanpa menambahkan *Feature Selection* sebesar 80%.
3	Ivandari, Tria Titiani Chasanah, Sattriedi Wahyu Binabar dan M. Adib AL Karomi (2017)	*Data Attribute Selection With Information Gain to Improve Credit Approval Classification Performance using K-Nearest Neighbor Algorithm*	Penelitian ini menunjukkan bahwa penggunaan *Feature Selection Information Gain* dalam klasifikasi persetujuan kredit dapat meningkatkan performa algoritma K-Nearest Neighbor. Pada *dataset UCI credit approval* akurasi meningkat sebesar 7.53% dan pada *dataset credit card customers* meningkat sebesar 3.26%.

2.2 Masa Studi Mahasiswa

Masa studi mahasiswa merupakan waktu mahasiswa menyelesaikan studinya di suatu perguruan tinggi. Seperti yang diketahui bahwa setiap perguruan tinggi memiliki batas waktu dalam menyelesaikan studinya, jika telah melewati batas waktu yang telah ditentukan oleh perguruan tinggi maka mahasiswa tersebut tidak bisa melanjutkan studinya lagi. Di Universitas Tanjungpura terutama di jurusan Informatika batas waktu mahasiswa menyelesaikan studinya adalah selama 7 tahun sebanyak 14 semester dan harus menyelesaikan 144 sks. Jika telah melewati batas waktu tersebut dan tidak menyelesaikan 144 sks maka mahasiswa tersebut akan mengalami drop out. Permasalahan tesebut dapat memberikan dampak yang buruk untuk suatu universitas maupun mahasiswa itu sendiri. Salah satu aspek penilaian untuk menentukan akreditas universitas maupun jurusan adalah jumlah mahasiswa menyelesaikan studinya tepat waktu. Dengan banyaknya jumlah mahasiswa yang lulus tepat waktu dan menurunnya jumlah mahasiswa yang lulus tidak tepat waktu maka akan memberikan dampak yang baik untuk universitas maupun jurusan untuk kedepannya. Maka dari itu, perlu diketahui mahasiswa mana saja yang berpotensi mengalami lulus tidak tepat waktu sehingga kedepannya dapat mengambil langkah dan diberikan pendampingan khusus untuk menghindari hal tersebut terjadi.

2.3 Jurusan Informatika Universitas Tanjungpura

Jurusan Informatika Universitas Tanjungpura merupakan salah satu jurusan yang ada di Fakultas Teknik Universitas Tanjungpura. Program Studi Teknik Informatika merupakan salah satu program studi jenjang sarjana (S1) di lingkungan Fakultas Teknik Universitas Tanjungpura, berdiri pada tanggal 18 mei 2004 dengan SK Dirjen DIKTI Nomor 1664/D/T/2004. Saat ini penyelenggaraan Program Studi Teknik Informatika berdasarkan SK Penyelenggaraan Nomor 7325/D/T/K-N/2011 tanggal 6 Juni 2011 dan telah terakreditasi B sesuai SK BAN-PT Nomor 0770/SK/BAN-PT/Akred/S/III/2017 tanggal 21 Maret 2017 (Teknik Informatika Untan, 2016).

Setiap tahunnya Jurusan Informatika Universitas Tanjungpura merupakan salah satu jurusan yang banyak diminati oleh calon mahasiswa baru yang ingin melanjutkan studi ke jenjang yang lebih tinggi dalam bidang teknologi informasi. Jurusan Informatika Universitas Tanjungpura memiliki visi yaitu pada tahun 2020 Program Studi Teknik Informatika menjadi pusat pengembangan ilmu pengetahuan dan SDM dibidang teknologi informasi di Kalimantan Barat (Teknik Informatika Untan, 2016). Selain itu Jurusan Informatika Universitas Tanjungpura memiliki tujuan yaitu sebagai berikut (Teknik Informatika Untan, 2016).:

1. Menyelenggarakan program studi yang terakreditasi sangat baik, yang memiliki sistem pendidikan dan pembelajaran yang didukung oleh SDM, fasilitas pendidikan dan sistem manajemen akademik yang baik.
2. Menjadi pusat kajian teknologi informasi di Kalimantan Barat yang mampu meningkatkan suasana akademis dan iklim kerja.
3. Menghasilkan lulusan yang bermoral, berbudi pekerti, berwawasan, serta mempunyai pengetahuan dasar keahlian yang luas agar mampu menggunakan komputer dalam berbagai bidang aplikasi, dan menguasai berbagai metode dan teknik pemecahan masalah dengan menggunakan komputer.

Jurusan Informatika Universitas Tanjungpura terus menjaga dan meningkatkan kualitas mahasiswa dengan melakukan monitoring terhadap nilai akademik mahasiswanya agar dapat menyelesaikan studinya tepat waktu dan menghasilkan mahasiswa yang berprestasi yang dapat bersaing satu sama lain dalam dunia kerja. Dengan meningkatnya jumlah mahasiswa lulus tepat waktu dan mahasiswa yang berprestasi maka nantinya akan memberikan dampak yang positif untuk meningkatkan akreditas Universitas, Fakultas, maupun Jurusan.

2.4 *Data Mining*

Data Mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari machine learning, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang

besar (Larose, 2005). *Data Mining* diibaratkan seperti menggali emas yang ada dibawah tanah, akan tetapi di data mining dilakukan dari menggali data yang ada di database untuk dijadikan suatu informasi yang bermanfaat. Penggunaan *data mining* dapat membantu dalam memecahkan permasalahan yang ada dengan melakukan analisa terhadap data yang telah didapatkan dari proses penggalian data di database. Seperti dalam pemanfaatan dari *data mining* dalam bidang bisnis yang dapat digunakan untuk mendapatkan informasi dalam meminimalkan pengeluaran keuangan perusahaan dan meningkatkan keuntungan, dalam bidang medis dapat memberikan informasi berupa penyakit yang di derita pasien, dan dalam bidang akademik dapat mengevaluasi nilai akademik mahasiswa. Terdapat beberapa tugas utama yang diselesaikan dengan *data mining* antara lain (Nofriansyah & Nurchayo, 2015) :

1). Estimasi

Digunakan untuk melakukan estimasi terhadap sebuah data baru yang tidak memiliki keputusan berdasarkan histori data yang telah ada. Contohnya ketika melakukan estimasi pembiayaan pada saat pembangunan sebuah Hotel baru pada kota yang berbeda.

2). Prediksi

Algoritma prediksi biasanya digunakan untuk memperkirakan atau forecasting suatu kejadian sebelum kejadian atau peristiwa tertentu terjadi. Contohnya pada bidang Klimatologi dan Geofisika, yaitu bagaimana Badan Meterologi Dan Geofisika (BMKG) memperkirakan tanggal tertentu bagaimana Cuacanya, apakah Hujan, Panas dan lain sebagainya. Ada beberapa metode yang sering digunakan salah satunya adalah Metode Rough Set.

3). Klasifikasi

Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Contoh pemanfaatannya adalah pada bidang Akademik yaitu Klasifikasi siswa yang layak masuk kedalam kelas unggulan atau akselerasi di sekolah tertentu.

4). Clustering

Digunakan untuk menganalisis pengelompokkan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokkan belum didefinisikan sebelum dijalankannya *tool data mining*. Biasanya menggunakan metode *neural network* atau statistik, analitikal hierarki cluster. Clustering membagi item menjadi kelompok - kelompok berdasarkan yang ditemukan *tool data mining*.

5). Asosiasi

Digunakan untuk mengenali kelakuan dari kejadian - kejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian. Adapun metode pemecahan masalah yang sering digunakan seperti Algoritma Apriori. Contoh pemanfaatan Algoritma Asosiasi yaitu pada Bidang Marketing ketika sebuah Minimarket melakukan Tata letak produk yang dijual berdasarkan Produk-produk mana yang paling sering dibeli konsumen, selain itu seperti tata letak buku yang dilakukan pustakawan di perpustakaan.

Metode pelatihan data mining memiliki 2 pendekatan, yaitu : supervised learning dan unsupervised learning. 2 pendekatan tersebut memiliki pengertian masing – masing, yaitu sebagai berikut (Ridwan et al., 2013):

1. Supervised Learning

   Supervised Learning merupakan metode belajar dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi, digunakan beberapa contoh data yang mempunyai output atau label selama proses training.

2. Unsupervised Learning

   Unsupervised learning, metode ini diterapkan tanpa adanya Latihan (training) dan tanpa ada guru (teacher). Guru di sini adalah label dari data.

2.5 Klasifikasi

Pengertian klasifikasi secara luas dapat diartikan sebagai cara dalam mengelompokkan data sesuai dengan kelasnya masing – masing dengan melakukan perhitungan terlebih dahulu. Menurut (Han et al., 2012) Klasifikasi adalah proses

penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Dalam klasifikasi terdapat 2 tipe data yang dapat diolah yaitu tipe data nominal dan tipe data numeric. Tipe data nominal merupakan tipe data yang berbentuk suatu kata sedangkan tipe data numeric merupakan tipe data yang berbentuk angka. Dengan adanya teknik klasifikasi, maka dapat mempermudah dalam mengelompokkan atau memprediksi suatu data yang belum diketahui kelasnya. Proses klasifikasi didasarkan pada empat komponen (Gorunescu, 2011) :

1. Kelas

   Variabel dependen yang berupa kategorikal yang merepresentasikan ‘label’ yang terdapat pada objek. Contohnya: resiko penyakit jantung, resiko kredit, customer loyalty, jenis gempa.

2. Predictor

   Variabel independen yang direpresentasikan oleh karakteristik (atribut) data. Contoh dari predictor adalah : merokok, mengkonsumsi alkohol, tekanan darah, frekuensi pembelian barang, status perkawinan, arah angin dan kecepatan, musim, dan lain - lain.

3. Training Dataset

   Satu set data yang berisi nilai dari kedua komponen di atas yang digunakan untuk ‘Melatih’ model menentukan kelas yang sesuai, berdasarkan predictor yang digunakan. Sebagai contoh dari training dataset adalah : kelompok pasien yang diuji pada serangan jantung

4. Testing Dataset

   Testing Dataset berisikan data baru yang akan diklasifikasikan oleh model yang telah dibuat dan akurasi klasifikasi dievaluasi.

Klasifikasi sendiri termasuk ke dalam kategori supervised learning yang dimana data yang telah diketahui kelasnya dijadikan sebagai data latih agar nantinya dapat diketahui kelas pada data testing. Terdapat berbagai algoritma yang dapat digunakan untuk proses klasifikasi antara lain : K-Nearest Neighbor, Naive Bayes, C 4.5, dan Support Vector Machine**.**

2.6 *Information Gain*

Information Gain (IG) merupakan suatu pengukuran yang dilakukan untuk melakukan seleksi terhadap atribut-atribut sehingga dapat disimpulkan atribut apa saja yang akan digunakan (Muqorobin, Kusrini, 2019). Setiap atribut yang digunakan untuk proses klasifikasi tidak semuanya mempunyai pengaruh. Atribut yang tidak berpengaruh dapat menurukan kinerja algoritma klasifikasi, maka sebab itu diperlukan proses mengseleksi atribut yang akan digunakan untuk proses klasifikasi nantinya. *Information Gain* merupakan salah satu *feature selection* yang dapat membantu dalam proses klasifikasi dalam mendapatkan atribut yang paling pengaruh untuk menentukan kelas pada data. Dalam proses *Information Gain* dilakukan perankingan pada setiap atribut kemudian disusun dari nilai tertinggi ke nilai terendah. Dalam pemilihan atribut teknik seleksi fitur Information gain dibagi dengan dua macam yaitu menggunakan atribut dengan batas threshold (*threshold* > 0.01*)* dan rangking dalam jumlah fitur tertentu (n = 5,10,15) (Nurina Sari, 2016). Pada penelitian ini, pemilihan atribut dilakukan dengan n=5 untuk evaluasi nilai akademik 4 semester dan n=10 untuk evaluasi nilai akademik 7 semester. Tahap dalam perhitungan *Information Gain* pertama – tama melakukan perhitungan entropi pada setiap atribut dengan menggunakan rumus persamaan berikut (Han et al., 2012):

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \quad \textbf{(1)}$$

Keterangan:

D : Himpunan Kasus

m = Jumlah Partisi D

p_i = Proporsi dari Di Terhadap D

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x\ Info(D_j) \quad \textbf{(2)}$$

Keterangan :

D = Himpunan Kasus

A = Atribut

V = Jumlah Partisi Atribut A

$|D_j|$ = Jumlah Kasus Pada Partisi ke j

$|D|$ = Jumlah Kasus Dalam D

$Info(D_j)$ = Total Entropi Dalam Partisi

Setelah mendapatkan nilai entropi pada setiap atribut, kemudian dilakukan perhitungan untuk mendapatkan nilai gain pada setiap atribut dengan menggunakan rumus berikut :

$$Gain(A) = Info(D) - Info_A(D) \quad \textbf{(3)}$$

Keterangan :

$Gain\ (A)$ = Nilai Gain Atribut A

$Info(D)$ = Total Entropi

$Info_A(D)$ = Entropi A

Nilai gain setiap atribut selanjutnya diurutkan dari nilai gain yang tertinggi ke nilai terendah. Atribut dengan nilai gain yang tinggi merupakan atribut yang paling pengaruh sehingga nantinya akan dilakukan proses klasifikasi dengan atribut yang paling pengaruh dan proses klasifikasi dengan semua atribut. hasil dari setiap proses akan dibandingkan untuk melihat apakah dengan penambahan *Feature Selection Information Gain* dapat meningkatkan akurasi pada algoritma atau menurunkan akurasi algoritma.

2.7 Algoritma K-Nearest Neighbor

Algoritma K-Nearest Neighbor merupakan salah satu algoritma yang dapat digunakan untuk mengklasifikasi data dan termasuk ke dalam *supervised learning* dan juga merupakan contoh dari *instance-based learning*. *Instance-based learning* merupakan data training di simpan untuk digunakan sebagai klasifikasi pada data baru yang belum diketahui kelasnya, sehingga lebih mudah ditemukan dalam membandingkan kemiripan dengan data training (Larose, 2005). Selain itu menurut (Ndaumanu et al., 2014) Algoritma K-Nearest Neighbor adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Algoritma K-Neaerest Neighbor telah digunakan oleh para peneliti untuk menyelesaikan beberapa masalah dalam proses klasifikasi dan telah memberikan hasil yang memuaskan. Nilai k dalam K-Nearest Neighbor sangatlah berpengaruh dalam proses klasifikasi. Dikarenakan nilai k merupakan jumlah data terdekat antara data training dan data testing yang akan mempengaruhi keakuratan proses klasifikasi nantinya. Untuk menentukan jarak antara data training dan data testing maka digunakan rumus *euclidean distance* yang ditunjukkan pada persamaan berikut (Yustanti, 2012) :

$$\boldsymbol{d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}} \qquad \textbf{(4)}$$

Keterangan :

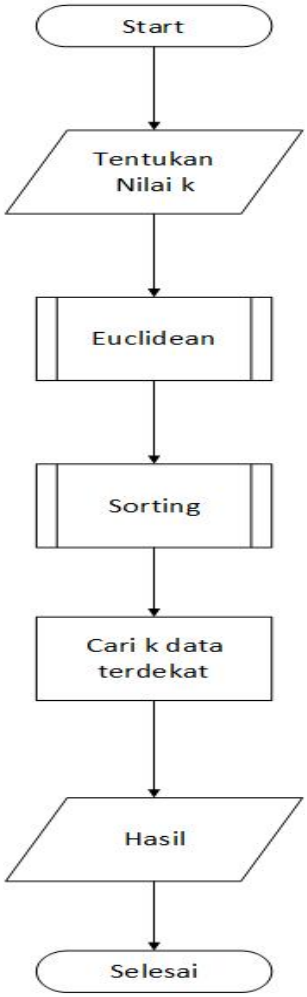
d = Jarak kedekatan antara data training dan data testing

i = Record (baris) ke-i dari tabel

x = Data Training

y = Data Testing

Selain itu, berikut merupakan *flowchart* dari algoritma K-Nearest Neighbor (Qodrat, 2017) :



Gambar 2.1 Flowchart K-Nearest Neighbor

Langkah-langkah algoritma KNN

1. Menentukan parameter K (jumlah tetangga paling dekat),
2. Menghitung kuadrat jarak euclid *(euclidean distance)* masing-masing objek terhadap data sampel yang diberikan
3. Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak euclidean terkecil
4. Mengumpulkan kategori Y (klasifikasi nearest neighbor)
5. Dengan menggunakan kategori mayoritas maka dapat hasil klasifikasi

2.8 Algoritma Naive Bayes

Naive Bayes merupakan algoritma klasifikasi yang termasuk juga kedalam *supervised learning*. Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan (Saleh, 2015). Selain itu juga ada pengertian lain dari Naive Bayes yaitu pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Bustami, 2014). Pada algortima Naive Bayes dalam penggunaan pada algoritma ini hanya membutuhkan jumlah data pelatihan (Data Training) yang sedikit dalam proses pengklasifikasian, akan tetapi dengan jumlah data training yang banyak maka akan menambah keakuratan dalam proses pengklasifikasiannya. Berikut merupakan rumus perhitungan dari algoritma naive bayes (Jananto, 2013):

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \quad \textbf{(5)}$$

Keterangan :

X = Data dengan class yang belum diketahui

H = Hipotesis data X merupakan suatu class spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (posteriori Probabilitas)

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

P(X|H) = Probabilitas X berdasarkan kondisi tersebut

P(X) = Probabilitas dari X

jika ada data yang bersifat numerik maka akan dilakukan proses pencarian rata - rata dan standard deviasi dari data numerik tersebut. rumus untuk mencari rata – rata dan standar deviasi yaitu sebagai berikut (Putri et al., 2014):

$$\mu = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} \quad \textbf{(6)}$$

Keterangan :

μ = hasil rata – rata (*mean*)

x_n = nilai sampel ke - n

n = jumlah sampel

Rumus standar deviasi :

$$б = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}{n-1}} \quad \textbf{(7)}$$

Setelah pencarian nilai rata – rata dan standard deviasi selesai maka dilakukan perhitungan data yang bersifat kontinu dengan rumus densitas Gauss, Berikut merupakan rumusnya :

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2(\sigma)^2}} \quad \textbf{(8)}$$

Keterangan :

б = standar deviasi

μ = rata – rata (mean)

π = 3,14

x = nilai atribut i

e = 2,7183

P = Peluang

y = Kelas yang dicari

2.9 *Confusion Matrix*

Confusion Matrix adalah sebuah tabel yang menyatakan jumlah data uji yang benar diklasifikasikan dan jumlah data uji yang salah diklasifikasikan (Indriani, 2014). Fungsi dari *confusion matrix* adalah untuk menentukan seberapa akurat algoritma klasifikasi yang digunakan untuk melakukan klasifikasi pada data. Suntoro (2019) mengatakan bahwa pengukuran evaluasi pada peranan *data mining* klasifikasi adalah mengukur akurasi dan perhitungan akurasi berdasarkan pada *confusion matrix* yang terbentuk. Contoh *confusion matix* untuk klasifikasi ditunjukkan pada table 2.2.

Tabel 2.2 Confusion Matrix

	Prediksi Ya	Prediksi Tidak
Aktual Ya	TP	FN
Aktual Tidak	FP	TN

Keterangan Tabel 2 Confusion Matrix sebagai berikut :

TP (True Positive) : Jumlah data Positif yang diklasifikasi sebagai data Positif

FP (False Positive) : Jumlah data Negatif yang diklasifikasikan sebagai data Positif

TN (True Negative) : Jumlah data Negatif yang diklasifikasikan sebagai data Negatif

FN (False Negative) : Jumlah data Positif yang diklasifikasikan sebagai data Negatif

Hasil dari *confusion matrix* nantinya dapat menentukan nilai dari akurasi, presisi, recall, dan f1-score. Presisi adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai actual (Tanjung et al., 2016). F1-Score merupakan rata-rata harmonik dari Precision dan Recall. F1-Score memberikan gambaran seberapa presisi dan seberapa baik model

dalam memberikan label klasifikasi (Santoso et al., 2018). Rumus dalam menghitung presisi, recall dan akurasi sebagai berikut :

$$\text{Precision} = \frac{TP}{TP + FP} \quad \textbf{(9)}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \textbf{(10)}$$

$$\text{Akurasi} = \frac{TP + TN}{TP + FP + TN + FN} \quad \textbf{(11)}$$

$$\text{F1-Score} = 2\frac{Presisi\ x\ Recall}{Presisi + Recall} \quad \textbf{(12)}$$

2.10 *PHP (Hypertext Preprocessor)*

PHP (Hypertext Preprocessor) adalah bahasa pemrograman *server side scripting* yang dapat di intergrasikan dengan *HTML (HyperText Markup Language)* dan javascript agar terbentuknya sebuah halaman web yang dinamis. Menurut Welling & Thomson (2017) Php adalah Bahasa *server-side scripting* yang dibuat khusus untuk web. Karena php merupakan *server side scripting* maka semua *script* yang dibuat di proses di sisi server dan hasil dari proses tersebut akan di tampilkan ke browser client. Php diciptakan oleh Rasmus Lerdorf, seorang *software engineer* asal greenland pada tahun 1995. Php telah menjadi salah satu bahasa pemrograman yang sering dipakai oleh web developer, dikarenakan php bersifat open source dan dapat digunakan secara gratis. Selain itu php memiliki banyak kelebihan yaitu sebagai berikut (Welling & Thomson, 2016) :

1. Performance

   Php mempunyai performasi yang tinggi yang dapat bekerja dengan cepat.

2. Integrasi ke Database

   Php memiliki native connection yang tersedia untuk berbagai macam database sistem. Selain Mysql, Php bisa langsung dikoneksikan ke PostgreSQL, Oracel, MongoDB, MSSQL, dan lainnya.

3. *Built-in Libraries*

   Karena Php di design untuk digunakan di web, sehingga memiliki banyak fungsi *built-in* untuk melakukan tugas yang berhubungan dengan web yang sangat bermanfaat.

4. Biaya

   PHP merupakan bahasa *scripting open source*, sehingga siapapun dapat menggunakannya secara gratis.

5. Mudah Dipelajari

   Php mudah untuk dipelajari. Bahasa pemrograman Php didasari pada bahasa pemrograman pada umumnya, seperti C dan Perl.

Pada penelitian ini nantinya menggunakan bahasa pemrograman php untuk membuat sistem klasifikasi.

2.11 Mysql

Mysql (My Structured Query Language) adalah aplikasi atau sistem untuk mengelola database atau manajemen data (Sugiyanto, 2013). Mysql bertugas untuk mengatur atau mengelola data yang tersimpan dalam database. Mysql pertama di kali direlease pada tahun 1995 yang di kembangkan oleh perusahaan yang bernama Mysql AB. Mysql bersifat *open source* dan dapat digunakan tanpa harus bayar. Mysql mempunyai beberapa kelebihan yaitu sebagai berikut (Welling & Thomson, 2016):

1. Performance

   Mysql mempunyai performasi yang tinggi.

2. Biaya Rendah

   Mysql tersedia secara gratis dibawah lisensi open source atau dengan biaya rendah dibawah lisensi komersial.

3. Mudah Digunakan

   Kebanyakan Database modern telah menggunakan SQL (Structure Query Language). Jika telah menggunakan RDBMS lain maka tidak ada masalah untuk beradaptasi menggunakan Mysql. Mysql juga lebih mudah untuk di setting daripada produk yang sama.

4. Portability

   Mysql dapat digunakan di berbagai sistem operasi.

Dengan kelebihan yang ada pada mysql membuat banyak programmer menjadikan Mysql sebagai tempat penyimpanan data terutama para web developer. Mysql menggunakan Bahasa *SQL* dalam pengelolaan data yang ada di database. Menurut Yakub (2008) *SQL* merupakan bahasa komputer yang mengikuti standar ANSI (American National Standard Institute), yaitu sebuah bahasa standar yang digunakan untuk mengakses dan melakukan manipulasi sistem basis data. *SQL* merupakan kependekan dari *Structured Query Languange*, yang dimana untuk menampilkan, menambahkan, mengedit dan menghapus data harus menggunakan query agar dapat melakukan pengelolaan data tersebut. Yakub (2008) mengemukakan bahwa secara umum *SQL* dibagi menjadi tiga, yaitu:

1) *Data Definition Language (DDL)*

   DDL merupakan suatu perintah yang berfungsi untuk mendefinisikan atribut-atribut basis data, tabel, atribut serta hubungan. Beberapa sintaks yang sering dijumpai adalah sebagai berikut:

   A) CREATE TABLE, bertugas untuk membuat tabel.

   B) CREATE INDEX, bertugas untuk membuat suatu indeks dalam tabel.

   C) DROP TABLE, bertugas untuk menghapus suatu tabel.

   D) DROP INDEX, bertugas untuk menghapus suatu indeks tabel.

   E) ALTER TABLE, bertugas untuk mengubah struktur suatu tabel.

2) *Data Manipulation Language (DML)*

   DML adalah sekumpulan perintah yang digunakan untuk melakukan proses manipulasi data pada suatu basis data. Berikut penjelasan sintaks-sintaks yang digunakan pada *DML*:

   A) INSERT, bertugas untuk menambahkan data ke dalam suatu tabel dalam suatu basis data.

   B) UPDATE, bertugas untuk merubah (update) data dalam suatu tabel dalam suatu basis data.

C) SELECT, bertugas untuk mengakses data dari suatu tabel dalam basis data.

D) DELETE, bertugas untuk menghapus data dari suatu tabel dalam basis data.

3) *Data Control Language (DCL)*

DCL merupakan kelompok perintah yang berfungsi untuk mengendalikan pengaksesan data, *DCL* digunakan untuk menangani masalah keamanan dalam database server. Sintaks yang digunakan dalam *DCL* adalah sebagai berikut:

A) GRANT, digunakan untuk memberikan atau mengijinkan seorang *user* untuk mengakses tabel dalam basis data tertentu.

B) REVOKE, digunakan untuk mencabut suatu hak akses dalam basis data tertentu.