BAB II
LANDASAN TEORI

2.1. Penelitian Terkait

Dalam melakukan penelitian ini, peneliti mencari referensi dari beberapa sumber, yang berkaitan dengan judul yang diambil. Berikut beberapa referensi yang berkaitan dengan judul penelitian: Seleksi Fitur Pada Algoritme C4.5 dan Algoritme *Random Forest* dalam Mendiagnosis Penyakit Ginjal Kronis. Secara detail penelitian terdahulu dapat dilihat pada tabel 2.1.

Tabel 2 1 Penelitian-Penelitian Terdahulu

No	Peneliti	Judul Penelitian	Tahun	Metode dan Hasil Penelitian	Kelebihan	Kelemahan
1	Hilda Amalia	Perbandingan Metode Data Mining SVM dan *Neural Network* untuk Klasifikasi Penyakit Ginjal Kronis	2018	Metode Data Mining *Support Vector Machine* (SVM) dan *Neural Network*	Ditampilkannya nilai *Accuracy* sebesar 95,16% (SVM)dan *table Confusion Matrix*	Tidak ditampilkan-nya data (atribut data)
2	Warid Yunus	Algoritma K-*Nearest* Neighbor Berbasis *Particle Swarm Optimization*(PSO)Untuk Prediksi Penyakit Ginjal Kronik	2018	Algoritma K-*Nearest* Neighbor Berbasis *Particle Swarm Optimization*(PSO)	Ditulisnya nilai akurasi, presisi, recall dan UAC pada Algoritma K-*Nearest* Neighbor	Tidak ditampilkan-nya data (atribut data)
3	Toni Arifin dan Daniel Ariesta	Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma *Naïve Bayes Classifier* Berbasis *Particle Swarm Optimization*(PSO)	2019	Algoritma *Naïve Bayes Classifier* Berbasis *Particle Swarm Optimization*(PSO)	Ditampilkannya peningkatan nilai *Accuracy* sebesar 1,75% (dari 97,00% ke 98,75%) dan *table Confusion Matrix*	-
4	I Gede Aditya Mahardika Pratama, dkk	Diagnosis Penyakit Ginjal Kronis dengan Algoritma C4.5, K-Means dan BPSO	2022	Algoritma C4.5, K-Means dan BPSO	accuracy sebesar 96,875%, precision sebesar 97%, recall sebesar 96,869% dan f-measure sebesar 97,487%.	Tidak ditampilkan-nya data (atribut data)
5	Eky Cahya Putra Witjaksana, dkk	Perbandingan akurasi algoritma *Random Forest* dan algoritma *Artficial Neural Network untuk klasifikasi penyakit diabetes*	2021	Algoritma *Random Forest* dan Algoritma *Artificial Neural Network* dalam melakukan klasifikasi penyakit diabetes berdasarkan dataset Pima Indians Diabetes	Dalam penelitian ini Algoritma *Random Forest* akurasi sebesar 90,62%,Algoritma *Artificial Neural Network* akurasi sebesar 82,29%	

No	Peneliti	Judul Penelitian	Tahun	Metode dan Hasil Penelitian	Kelebihan	Kelemahan
6	Julia Triani, dkk	Komparasi Dalam Prediksi Gagal Jantung Dengan Menggunakan Metode C4.5 dan Naïve Bayes	2023	Algoritma C4.5 dan Algoritma *Naïve Bayes*	Ditulisnya nilai akurasi, presisi, recall dan f1-score pada Algoritma C4.5 dan Algoritma *Naïve Bayes* dengan outlier dan tanpa outlier	-
7	I Made Agus Oka Gunawan, dkk	Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4.5 Dengan Rapidminer	2023	Algoritma Decision Tree Series C4.5	Dengan menggunakan Algoritma Decision Tree Series C4.5 tingkat Accuracy sebesar 80,43%	-
8	Umri Erdiansyah, dkk	Komparasi Metode K-Nearest Neighbor dan Random Forest Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil	2022	Algoritma K-Nearest Neighbor dan Algoritma Random Forest	Dari hasil Pengujian tingkat akurasi metode KNN sebesar 90.00 % dan metode Random Forest dengan tingkat akurasi 85,50 %.	-
9	Sabrina Adnin Kamila, dkk	Klasifikasi Penyakit Jantung Menggunakan *Decision Tree* dan *Random Forest*	2023	Algoritma Decisiion Tree dan Algoritma *Random Forest*	*Random forest* menghasilkan nilai akurasi sebesar 81.82% dan *decision tree* menghasilkan nilai akurasi sebesar 77.44%.	-
10	Sunanto dan Ghazi Falah	Penerapan Algoritma C4.5 untuk membuat model prediksi pasien yang mengidap penyakit diabetes	2022	Metode yang digunakan dalam penelitian ini menggunakan Algoritma C4.5	Dalam penelitian ini diperoleh akurasi prediksi sebesar 95,51%.	-

No	Peneliti	Judul Penelitian	Tahun	Metode dan Hasil Penelitian	Kelebihan	Kelemahan
11	Laura Sari, dkk	Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma *Random Forest*	2023	Algoritma Random Forest dan Algoritma *Naïve Bayes*	Algoritma *Random Forest* menghasilkan nilai Accuracy yang lebih tinggi yaitu sebesar 98,4%, naïve bayes sebesar 95,1%	
12	Lia andiani, dkk	Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest	2019	Algoritma K-Nearest Neighbor dan Algoritma Random Forest	Dalam penelitian ini akurasi metode KNN 93% dan metode Random Forest 72%.	-

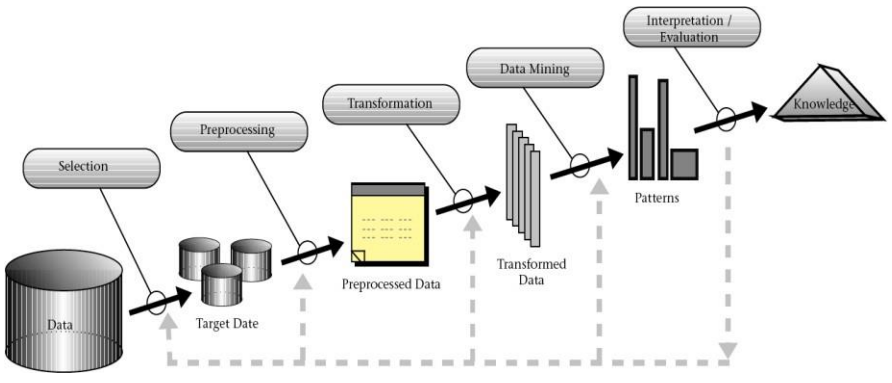
Dari beberapa penelitian terkait dapat disimpulkan bahwa penerapan metode atau algoritme C4.5 dan algoritme *Random Forest* dalam mengklasifikasi penyakit dapat dimanfaatkan dalam mendiagnosis penyakit yang diderita oleh pasien. Selain itu hasil penelitian yang diperoleh dengan menggunakan algoritme yang sama bisa berbeda, Hal ini disebabkan karena terjadinya beberapa masalah seperti ketidakseimbangan data akibat terlalu banyaknya atribut sehingga tingkat akurasi yang dihasilkanpun berbeda-beda.[12]

Oleh sebab itu dipenelitian ini saya menitik beratkan pada tingkat akurasi yang maksimal dan menggunakan algoritme C4.5 dan algoritme *Random Forest* dalam mengklasifikasi penyakit Penyakit Ginjal Kronis.

2.2. Teori Dasar

2.2.1. Knowledge Discovery in Database (KDD)

Pada Pada metode analisis akan dimining dengan melalui tahapan knowledge discovery in database (KDD). Proses KDD adalah proses menggunakandata mining untuk mengekstrak pengetahuan apa yang dianggap sesuai dengan spesifikasi ukuran dan batas, menggunakan database bersama dengan preprocessing yang diperlukan, pengambilan sampel dan transformasi dari database. Adapun untuk menganalisis data dalam penerapan data mining ini menggunakan tahapan Knowledge Discovert in Database (KDD).istilah data mining dan knowledge discovery in database (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi yang tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan dengan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining (Nofriansyah, 2014). Secara detail proses KDD dapat dilihat pada gambar 2.1.



Gambar 2 1 Proses *Knowledge Discovery in Database* (KDD)

Berdasarkan gambar di atas, proses KDD secara garis besar dijelaskan sebagai berikut:

a. Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, kata memilih data-data seperti apa saja yang kita butuhkan untuk ptoses lebih lanjut dan kemudian data disimpan dalam suatu berkas, terpisah dari basis data operasional sehingga memberikan kemudahan untuk penggunaan berikutnya.

b. Pre-processing (Cleaning)

Kondisi data yang diperoleh, baik dataset dari prifat maupun data sekunder baik dari suatu perusahaan maupun eksperimen pribadi, terdapat beberapa dataisian yang kosong atau data tidak sama atau tidak sejenis atau data yang tidak valid atau juga hanya sekedar salah ketik. Selain itu, ada juga atribut data yangtidak relevan akan dilakukan pre-processing data dengan tujuan keberadaan data tersebut bermanfaat tidak mengurangi mutu hasil data yang di olahnantinya.

c. Data Transformation

Beberapa teknik data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa teknik standar seperti analisis asosiasi dan klastering hanya bisa menerima input katagorikal. Karenanya data berupa angka numerikyang berlanjut bisa dibagi-bagi menjadi beberapa interval.Proses ini sering disebut binning. Disini juga dilakukan pemilihan data yang diperlukan oleh teknik data mining yang dipakai. Transformasi dan pemilihan data ini juga menentukan kualitas dari hasil data mining nantinya.

d. Data Mining

Data mining merupakan pola atau informasi yang diambil dari data terpilih dengan menggunakan teknik atau metode tertentu, baik klasifikasi, ertimasi atau lainnya. Banyak algoritme atau metode yang digunakan dalam data mining, tergantung dataset yg diperoleh dan tujuannya. Data yang digunakan dalam penelitian untuk bisa menjadi sebuah model yang baik idealnya mencukupi sebagai data riset. Baik tidaknya hasil data, dipengaruhi besar kecilnya data, semakin banyak data memungkinkan semakin sedikit kesalahan sehingga semakin bagus model yang dihasilkan.

e. Interpretation (Evaluation)

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interprestation. Tahap ini mencakup pemerikasaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta pola atau informasi yang ada sebelumnya.[13]

2.2.2. Data Mining

Data mining merupakan serangkaian proses untuk memperoleh nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data. Data mining mulai ada sejak 1990-an sebagai cara yang benar dan tepat untuk mengambil pola dan informasi yang digunakan untuk menemukan hubungan antara data untuk melakukan pengelompokkan ke dalam satu atau lebih cluster sehingga objek - objek yang berada dalam satu cluster akan mempunyai kesamaan yang tinggi antara satu dengan lainnya. Ilmu Data mining yang dipelajari saat ini merupakan perpaduan ilmu dari artificial intelligence, statistik, dan penelitian basis

data yang selalu meningkat sesuai dengan perkembangan zaman. data Data mining merupakan sebuah proses menentukan ikatan yang mengandung arti, pola, dan keterkaitan dengan mengeksplorasi data yang ada [14]. Data mining merupakan bagian dari proses penemuan pengetahuan dari basis data *Knowledge Discovery in Database*[15]. Salah satu teknik yang ada di dalam data mining adalah supervised learning. Teknik Supervised Learning sendiri membutuhkan training dataset yang nanti nya dengan menentukan nilai input akan menghasilkan output sebagai target [16].

Data Mining dibagi menjadi beberapa metode berdasarkan tugas yang dapat dilakukan diantaranya Asosiasi (Association), Pengklasteran (Clustering), Prediksi (Prediction), Estimasi (Estimation) dan Klasifikasi (Classification) [11].

1. Asosiasi (Association) Assosiasi merupakan proses pengelompokan didalam data mining untuk menemukan suatu hubungan yang terdapat pada nilai atribut dari sekumpulan data yang dimiki baik dalam skala besar maupun kecil, sedangkan Klasifikasi adalah teknik yang dilakukan untuk memprediksi class atau properti dari setiap instance data, dan Clustering sendiri memiliki makna mengelompokkan data tanpa berdasarkan kelas data tertentu ke dalam kelas objek yang sama sesuai tujuan penelitian yang di fokuskan.
2. Pengklasteran (Clustering) Pengklusteran merupakan proses membagi data dalam suatu himpunan ke dalam beberapa kelompok yang kesamaan datanya dalam suatu kelompok lebih besar daripada kesamaan data tersebut dengan data dalam kelompok lain.
3. Prediksi (Prediction) Prediksi merupakan suatu proses memperkirakan secara sistematis berdasarkan data yang telah ada tentang sesuatu yang paling mungkin

terjadi di masa depan, sehingga selisih antara sesuatu yang terjadi dengan hasil perkiraan dapat diperkecil. Perlu disampaikan dalam memprediksi tidak harus memberikan jawaban secara cepat dan tepat sesuai dengan kejadian yang akan terjadi, melainkan mencari jawaban sedekat mungkin yang akan terjadi kedapan

4. Estimasi (Estimation) Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

5. Klasifikasi (Classification) Klasifikasi merupakan pengolahan data yang menghasilkan pola dalam untuk memprediksi label kelas sampel harus diklasifikasikan. Beberapa teknik klasifikasi dalam data mining telah diusulkan dalam bidang-bidang seperti pembelajaran mesin, sistem pakar dan statistik [17].

2.2.3. Algoritma C4.5

Algoritma C4.5 adalah salah satu metode untuk membuat pohon keputusan berdasarkan training data yang telah disediakan. Beberapa pengembangan yang dilakukan pada C45 antara lain bisa mengatasi missing value, bisa mengatasi data kontinu, dan prunning. Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5 (Larose, 2005), yaitu:

1) Mempersiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya atau disebut data masa lalu dan sudah dikelompokkan dalam kelas-kelas tertentu.
2) Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai

gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai entropy. Untuk menghitung nilai entropy digunakan rumus persamaan 1.

$$Entropy\ (\mathrm{S}) = \sum_{i=1}^{n} -\mathrm{pi} * \log 2(\mathrm{pi}) \quad \text{.....(1)[18]}$$

Keterangan:

S = Himpunan kasus
n = Jumlah partisi S
pi = Proporsi Si terhadap S

3) Menghitung nilai Gain menggunakan Persamaan 2.[18]

$$Gain(\mathrm{S},\mathrm{A}) = Entropy(\mathrm{S}) = \sum_{i=1}^{n} \frac{|\mathrm{Si}|}{|\mathrm{S}|}\ Entropy\ (\mathrm{Si}) \quad \text{.....(2)}$$

Keterangan:

S = Himpunan kasus
A = Fitur
n = Jumlah partisi atribut A
$|Si|$ = Proporsi Si terhadap S
$|S|$ = jumlah kasus dalam S [19]

4) Sebelum mencari *Gain Ratio*, di cari S*plit Info* dengan rumus persamaan 3.

$$\mathrm{SplitInfo}\ (\mathrm{S},\mathrm{A}) = \sum_{i=1}^{c} \frac{|\mathrm{Si}|}{|\mathrm{S}|}\ log2\ \frac{|\mathrm{Si}|}{|\mathrm{S}|} \text{.....(3)[20]}$$

Keterangan:

S = Jumlah data sampel

Si = Jumlah masing-masing pada setiap atribut.

5) *Gain Ratio*, dengan rumus persamaan 4.

$$Gain\ Ratio\ (\mathrm{S},\mathrm{A}) = \frac{\mathrm{Gain\ (S,A)}}{\mathrm{SplitInfo(S,A)}} \text{.....(4)[21]}$$

6) Ulangi langkah ke 2 dan langkah ke 5 hingga semua record terpartisi

Proses partisi pohon keputusan akan berhenti saat:

a. Semua record dalam simpul N mendapat kelas yang sama.
b. Tidak ada atribut di dalam record yang dipartisi lagi.
c. Tidak ada record di dalam cabang yang kosong(Pambudi & Setiawan, n.d.2018).

7) Ulangi langkah ke 2 dan langkah ke 3 hingga semua record terpartisi
8) Proses partisi pohon keputusan akan berhenti saat:
   a. Semua record dalam simpul N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam record yang dipartisi lagi.
   c. Tidak ada record di dalam cabang yang kosong.[23]

2.2.4. Algoritme *Random Forest*

Random forest merupakan salah satu metode yang digunakan untuk klasifikasi dan regresi[24]. Random Forest merupakan algoritma berbasis ensemble yang dibangun berdasarkan Decision Tree. Algoritma berbasis ensemble adalah gabungan dari beberapa teknik pembelajaran mesin (machine learning) yang digabungkan menjadi satu model prediktif [25]. Metode random forest (RF) merupakan metode yang dapat meningkatkan hasil akurasi, karena dalam membangkitkan simpul anak untuk setiap node dilakukan secara acak [26]. Algoritma atau prosedur dalam membangun Random Forest pada gugus data yang terdiri dari n amatan dan terdiri atas q variabel independen [27]. Algoritma random forest cocok untuk diterapkan pada data dengan jumlah yang besar, termasuk dalam masalah diagnosis penyakit yang dalam penelitian ini adalah penyakit ginjal kronis[28].

Algoritme *Random forest* (RF) adalah suatu algoritma yang digunakan pada klasifikasi data dalam jumlah yang besar. Klasifikasi random forest dilakukan melalui penggabungan pohon (tree) dengan melakukan training pada sampel data yang dimiliki. Penggunaan pohon (tree) yang semakin banyak akan mempengaruhi

akurasi yang akan didapatkan menjadi lebih baik. Penentuan klasifikasi dengan random forest diambil berdasarkan hasil voting dari tree yang terbentuk. Pemenang dari tree yang terbentuk ditentukan dengan vote terbanyak. [6]. Pohon keputusan dimulai dengan cara menghitung nilai entropy sebagai penentu tingkat ketidakmurnian atribut. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan 1.

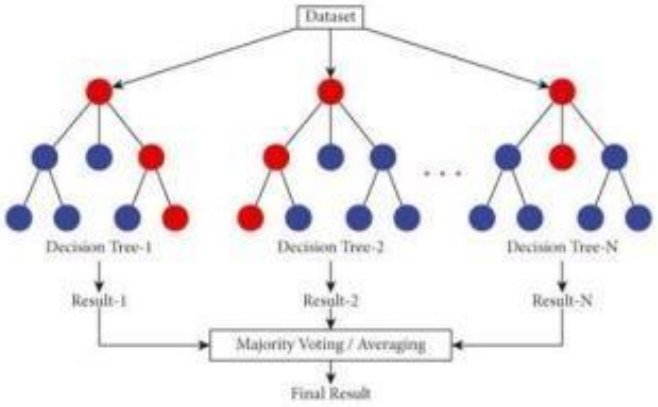
Didilanjutkan dengan menghitung gain digunakan rumus seperti pada persamaan 2 dan menentukan gain tertinggi. Selanjutnya mencari Gain Ratio menggunakan persamaan 4, sebelum mencari Gain Ratio, di cari S*plit Info* menggunakan persamaan 3.

Tahapan penyusunan dan pendugaan menggunakan Random Forest adalah (Breiman, 2001; Breiman & Cutler, 2003):

a. (tahapan bootstrap) melakukan penarikan contoh acak dengan pemulihan berukuran n dari gugus data training.
b. (tahapan random sub-setting) dengan menggunakan gugus data bootstrap, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m < p, lalu pemilah terbaik dipilih berdasarkan m peubah penjelas tersebut.

c. Ulangi langkah a-b sebanyak k kali sehingga diperoleh sebuah hutan yang terdiri atas k buah pohon acak [29].

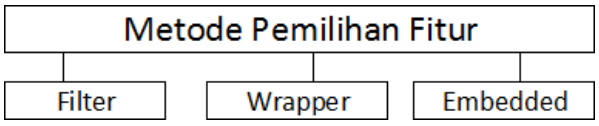
Secara detail struktur *random forest* dapat dilihat pada gambar 2.2.



Gambar 2 2 Struktur *Random Forest* [5]

2.2.5. Seleksi Fitur (*Feature Selection*)

Memilih fitur-fitur penting dikenal sebagai seleksi fitur atau seleksi atribut atau seleksi variabel. Keuntungan terpenting dari pemilihan fitur ini adalah tidak hanya meningkatkan akurasi, namun juga meningkatkan efisiensi klasifikasi. Hal ini dapat dilakukan secara manual atau otomatis, namun memilih secara manual dengan kumpulan data yang sangat besar memakan waktu dan rumit, oleh karena itu berbagai algoritma pemilihan fitur telah disajikan dalam pembelajaran mesin. Pemilihan fitur dalam pembelajaran mesin memiliki tipe yang berbeda seperti *filter*, *wrapper*, dan metode *embedded*. [30] Secara detail metode pemilihan fitur dapat dilihat pada gambar 2.3.



Gambar 2 3 Metode Pemilihan Fitur

a) Metode Penyaringan (*Filter method*)

Metode ini menggunakan pendekatan statistik untuk mencari korelasi atau hubungan antar atribut; sehingga atribut dengan korelasi paling kecil dapat dihilangkan dengan mudah yang pada gilirannya menghasilkan subset fitur. Itu tidak bergantung pada algoritma pembelajaran mesin tertentu. Ini tidak hanya mengurangi konsumsi waktu tetapi juga meningkatkan akurasi model klasifikasi. Metode filter lebih cepat dibandingkan dengan metode wrapper.

b) Metode Pembungkus (*Wrapper method*)

Tidak seperti filter, filter ini bergantung pada algoritma pembelajaran mesin tertentu. Ini menciptakan beberapa model yang berbeda subset fitur. Subset yang memberikan performa lebih baik untuk model tertentu akan dipilih. Waktu komputasi akan lama ketika model menangani banyak fitur. Karena ini menciptakan beberapa model dengan subset fitur yang berbeda, hal ini dapat menyebabkan pemasangan berlebih.

c) Metode Tertanam (*Embedded method*)

Ini berisi kualitas metode filter dan pembungkus. Fitur dipilih selama proses pembuatan model, yaitu fitur yang dipilih selama iterasi pelatihan model. Oleh karena itu, pemilihan fitur yang relevan pasti akan mengurangi waktu komputasi dan meningkatkan akurasi model klasifikasi.

2.2.6. Metode Evaluasi

2.2.6.1. Confusion Matrix (Accuracy)

Confusion matrix merupakan alat yang digunakan untuk menggambarkan kinerja model klasifikasi pada data uji yang sudah diketahui hasil sebenarnya. *Confusion matrix* berbentuk tabel matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui. Perlu diketahui pada teknik *confussion matrix* dapat dilakukan beberapa perhitungan mulai dari nilai akurasi, precission dan recall maupun F1 score. Secara detail *confussion matrix* dapat dilihat pada tabel 2.2.

Tabel 2 2 Confussion matrix[31]

Aktual	Prediksi	
	+	-
+	TP	FN
-	FP	TN

• True Positif (TP) = jumlah tebakan yang benar, di mana model kita menebak seseorang terinfeksi PKG, dan kenyataannya memang benar.

• False positif (FP) = jumlah tebakan di mana model kita menebak seseorang terinfeksi PGK, tapi kenyataannya itu salah

• True Negatif (TN): jumlah tebakan yang benar, di mana model kita menebak seseorang tidak terinfeksi PKG, dan kenyataannya memang benar.

• False Negatif (FN): merupakan jumlah tebakan di mana model kita menebak seseorang tidak terinfeksi PGK, tapi kenyataannya itu salah..[32]

Rumus yang digunakan dalam confussion matrix untuk memperoleh *accuracy* pada persaman 5 sebagai berikut:

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad \text{.....(5) [33]}$$

Rumus yang digunakan dalam confussion matrix untuk memperoleh *precission* pada persaman 6 sebagai berikut:

$$Precission = \frac{\text{TP}}{\text{TP+FP}} \quad \text{.....(6)}$$

Rumus yang digunakan dalam confussion matrix untuk memperoleh *recall* pada persaman 7 sebagai berikut:

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \quad \text{.....(7)}$$

Rumus yang digunakan dalam confussion matrix untuk memperoleh *F1 Score* pada persaman 8 sebagai berikut:

$$F1\ Score = 2 * \frac{Precission*Recall}{Precission+Recall} \quad \text{.....(8) [34]}$$

2.2.6.2. ROC Curve

Kurva ROC (*Receiver Operating Characteristic*) menunjukan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan confusion matrix. ROC adalah grafik dua dimensi dengan false positives sebagai garis horizontal dan true positives untuk mengukur perbedaaan performasi metode yang digunakan.[35] Visualisasi hasil perhitungan digambarkan dengan Area Under ROC Curve (AUC). Tingkatan dalam pengukuran kualitas *classifier* menggunakan ROC dilihat berdasarkan nilai akurasi dengan rentang yang diperlihatkan pada Tabel 2.3.

Tabel 2 3 Nilai Kualitas *Classifier*.[36]

Rentang Akurasi	**Kualitas *Classifier***
0.90 - 1.00	*Sangat Baik*
0.80 - 0.90	*Baik*
0.70 - 0.80	*Cukup*
0.60 - 0.70	*Rendah*
0.50 - 0.60	*Sangat Rendah*