BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1 Tinjauan Pustaka

Penelitian ini menggunakan beberapa sumber pustaka yang berhubungan dengan kasus atau metode dalam penelitian ini sebagai acuan. Acuan yang digunakan pada penelitian ini berupa karya ilmiah, jurnal, hingga buku, berikut tinjauan pustaka yang digunakan pada penelitian ini:

Tabel 2.1 Tinjauan Pustaka

No	Penulis	Metode	Hasil
1.	Asih Asmarani	K-Nearest Neighbor	Prediksi benar sebanyak 20 dan salah 10 dengan tingkat akurasi sebesar 66,6667 % dengan jarak k = 5
2.	Fitrokh Nur Ikhromr, Ipin Sugiyarto, Umi Faddillah, Bibit Sudarsono	Naves Bayes Dan K-Nearest Neighbor	Evaluasi menggunakan 2000 set data pasien diabetes K-Nearest Neighbor menghasilkan akurasi sebesar 99% sedangkan Naives Bayes mengahsilkan akurasi sebesar 75% serta membuktikan bahwa model evaluasi menggunakan algoritma K-Nearest Neighbor mendapatkan hasil terbaik.
3.	Feri Irawan, Tati Suprapti, dan Agus Bahtiar	Naves Bayes Dan K-Nearest Neighbor	Melakukan perbandingan perhitungan secara manual Hasil yang diperoleh dari klasifikasi Naive Bayes *accuracy* sebesar 76.6 *precision* sebesar 76.8, *recall* sebesar 76.7 sedangkan K-Nearest Neighbor mendapatkan *accuracy* sebesar 92.6, *precision* sebesar 92.6, *recall* sebesar 92.6.
4.	Happy Andrian Dwi Fasnuari, Haris Yuana, M. Taofik Chulkamdi	K-Nearest Neighbor	Diagnosa dapat terjamin keasliannya dan proses uji klinis tersebut tentunya memakan waktu yang lama. Jurnal ini menggunakan sebanyak 108 *data training* dan 27 *data testing* dan menggunakan algoritma K-Nearest Neighbor (KNN) menghasilkan akurasi 93% pada K=9, presisi 100%,

Tabel 2.1 Tinjauan Pustaka (Lanjutan)

No	Penulis	Metode	Hasil
4.			recall 60% dan F1-Score 75%. Dengan tingkat akurasi sebesar 93% maka penelitian ini dinilai telah berhasil.
5.	Erfan Karyadiputra dan Agus Setiawan	Decision Tree C4.5, Naive Bayes dan K-Nearest Neighbors.	Dari penelitian ini menunjukkan bahwa algoritma *Decision Tree* C4.5 menjadi algoritma terbaik berdasarkan hasil performance akurasi prediksi sebesar 96,35% dengan nilai AUC sebesar 0,949 sehingga termasuk ke dalam kategori *excellent classification.*
6	Candrani Sri Murtono	K-Nearest Neighbors	Klasifikasi k-NN dilakukan dengan cara menentukan ketetanggaan terdekat.Penelitian ini menggunakan 9 data faktor resiko dengan Hasil yang diperoleh dari klasifikasi pada data ini adalah *accuracy* sebesar 98,08%, *precision* sebesar 96,30%, dan *recall* sebesar 100.00%.

Berdasarkan tabel 2.1 dapat diketahui bahwa referensi yang dipakai yaitu referensi pertama adalah Jurnal Informatika dan Rekayasa Komputer (JAKAKOM) Volume 2, Nomor 2 yang ditulis oleh Asih Asmarani, dkk (2022). Penelitian ini bertujuan mengimplementasikan algoritma untuk memprediksi Penyakit Diabetes agar dapat memantau, memberi infromasi, dan menghimbau masyarakat untuk menjaga pola hidup sehat, serta memperoleh hasil prediksi benar sebanyak 20 dan salah 10 dengan tingkat akurasi sebesar 66,667% dengan jarak k = 5. Pada penelitian ini memiliki kesamaan pada variable yang penulis gunakan yaitu Penyakit Diabetes Mellitus dan kesamaan pada metode algoritma yang digunakan yaitu K-Nearest Neighbor.

Referensi kedua adalah *Journal of Information Technology and Computer Science* (INTECOMS) Volume 6 Nomor 1 yang dilakukan oleh Fitrokh Nur Ikhromr, Ipin Sugiyarto, Umi Faddillah, Bibit Sudarsono (2023). Penelitian ini dilakukan untuk menentukan seorang pasien memiliki risiko diabetes dengan

lebih cepat dan akurat. Pengujian model pada penelitian ini menggunakan algoritma Navies Bayes dan K-Nearest Neighbor. Hasil evaluasi menggunakan 2000 set data pasien diabetes K-Nearest Neighbor menghasilkan akurasi sebesar 99% sedangkan Naives Bayes mengahsilkan akurasi sebesar 75% serta membuktikan bahwa model evaluasi menggunakan algoritma K-Nearest Neighbor mendapatkan hasil terbaik. Pada penelitian ini memiliki kesamaan pada variable dan algoritme yang penulis gunakan.

Referensi ketiga adalah Jurnal Teknik Elektro dan Informatika Volume 18 Nomor 1 yang disusun oleh Feri Irawan, Tati Suprapti, dan Agus Bahtiar (2023) dengan judul. Penelitian ini bertujuan memprediksi menggunakan metode klasifikasi untuk menentukan algoritma apa yang cocok dalam mendiagnosa Penyakit Diabetes Mellitus Tipe 2. Penelitian ini menggunakan dua algoritma Naive Bayes dan K-Nearest Neighbor. Jurnal ini memiliki hasil dalam melakukan perbandingan perhitungan secara manual hasil yang diperoleh dari klasifikasi Naive Bayes mendapatkan *accuracy* sebesar 76.6, *precision* sebesar 76.8, *recall* sebesar 76.7 sedangkan K-Nearest Neighbor mendapatkan *accuracy* sebesar 92.6, *precision* sebesar 92.6, *recall* sebesar 92.6. Pada penelitian ini memiliki kesamaan pada variable yang diteliti yaitu Diabetes Mellitus Tipe 2 dan Algoritma K-Nearest Neighbor

Referensi keempat adalah Jurnal Ilmiah Teknik Informatika : ANTIVIRUS Vol. 16 No.2 yang dilakukan oleh Happy Andrian Dwi Fasnuari, Haris Yuana, M. Taofik Chulkamdi (2022). Penelitian ini dilakukan karena Penyakit Diabetes Mellitus diperlukan beberapa uji kesehatan agar hasil diagnosa dapat terjamin keasliannya dan proses uji klinis tersebut tentunya memakan waktu yang lama. Jurnal ini menggunakan sebanyak 108 *data training* dan 27 *data testing* yang diolah menggunakan algoritma K-Nearest Neighbor (K-NN). Menghasilkan akurasi 93% pada K=9, presisi 100%, recall 60% dan *F1-Score* 75%. Dengan tingkat akurasi sebesar 93% maka penelitian ini dinilai telah berhasil menerapkan metode KNN untuk melakukan klasifikasi terhadap penyakit

diabetes melitus.. Pada penelitian ini memiliki kesamaan pada variable Diabetes Mellitus Tipe 2 dan algoritma *K-Nearest Neighbors* yang penulis gunakan.

Referensi kelima adalah Jurnal Media Informasi Sains dan Teknologi : Teknosains Volume 16, Nomor 2 yang dilakukan oleh Erfan Karyadiputra dan Agus Setiawan (2020). Penelitian bertujuan untuk memprediksi kemungkinan awal seseorang terindikasi penyakit diabetes dengan menggunakan teknik data mining yaitu metode Algoritma *Decision Tree* C4.5, *Naive Bayes* dan *K-Nearest Neighbors*. Penelitian ini memiliki hasil dari penelitian ini menunjukkan bahwa algoritma *Decision Tree* C4.5 menjadi algoritma terbaik berdasarkan hasil performance akurasi prediksi sebesar 96,35% dengan nilai AUC sebesar 0,949 sehingga termasuk ke dalam kategori *excellent classification.* Pada penelitian ini memiliki kesamaan pada variable Diabetes Mellitus Tipe 2 dan algoritma *K-Nearest Neighbors* yang penulis gunakan.

Referensi keenam adalah Karya Tulis Ilmiah yang dilakukan oleh Candrani Sri Murtono Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim pada Tahun 2022. Penelitian ini memiliki tujuan untuk mengetahui model dan hasil klasifikasi potensi Penyakit Diabetes Mellitus agar masyarakat dapat melakukan tindakan preventif terhadap Penyakit Diabetes Mellitus berdasarkan hasil yang didapatkan. Data yang digunakan merupakan Data Diabetes Mellitus dari RST tk.II dr.Soepraoen Malang. Klasifikasi K-NN dilakukan dengan cara menentukan ketetanggaan terdekat. Penelitian ini menggunakan 9 data faktor resiko dengan. Hasil yang diperoleh dari klasifikasi pada data ini adalah *accuracy* sebesar 98,08%, *precision* sebesar 96,30%, dan *recall* sebesar 100.00%.

2.2 Landasan Teori

2.2.1 Penyakit Diabetes

Diabetes Mellitus adalah penyakit kronis yang terjadi ketika tubuh tidak secara efektif menggunakan insulin yang dihasilan *hiperglikemia* atau gula darah tinggi merupakan efek dari diabetes yang tidak terkontrolnya yang menyebabkan

kerusakan pada sistem tubuh, khususnya saraf dan pembuluh darah (Arisman, 2015).

Salah satu penyakit yang bisa menyebabkan komplikasi penyakit berat lainnya seperti darah tinggi, jantung dll adalah Penyakit Diabetes. Hal ini dikarenakan adanya gangguan yang terjadi pada metabolisme dimana tidak terproduksinya insulin oleh pankreas. Penyakit Diabetes ada di seluruh dunia tetapi peningkatan jumlah penderitanya berbeda di setiap negara (Fiarni et al., 2019). Untuk insulin itu sendiri adalah hormon yang sangat penting dalam pengaturan kadar glukosa darah. Dalam konteks Penyakit Diabetes, insulin menjadi fokus utama karena gangguan pada produNksi atau penggunaan insulin dapat menyebabkan peningkatan kadar glukosa dalam darah, yang merupakan ciri khas diabetes. Di Indonesia prevalence akan Penyakit Diabetes mencapai 10,9%, dan mengalami peningkatan (Fiarni et al., 2019).

Klasifikasi Penyakit Diabetes Mellitus berdasarkan etiologi menurut Perkeni (2015) adalah sebagai berikut :

1. Diabetes Mellitus (DM) tipe 1

Diabetes Mellitus yang terjadi karena kerusakan atau destruksi sel beta di pankreas. kerusakan ini berakibat pada keadaan defisiensi insulin yang terjadi secara absolut. Penyebab dari kerusakan sel beta antara lain autoimun dan idiopatik.

2. Diabetes Melitus (DM) tipe 2

Penyebab Diabetes Mellitus Tipe 2 seperti yang diketahui adalah resistensi insulin. Insulin dalam jumlah yang cukup tetapi tidak dapat bekerja secara optimal sehingga menyebabkan kadar gula darah tinggi di dalam tubuh. Defisiensi insulin juga dapat terjadi secara relatif pada penderita Diabetes Mellitus Tipe 2 dan sangat mungkin untuk menjadi defisiensi insulin absolut.

3. Diabetes Mellitus (DM) tipe lain

Penyebab Diabetes Mellitus tipe lain sangat bervariasi. Diabetes Mellitus tipe ini dapat disebabkan oleh defek genetik fungsi sel beta, defek genetik kerja 8 insulin, penyakit eksokrin pankreas, endokrinopati pankreas, obat, zat kimia, infeksi, kelainan imunologi dan sindrom genetik lain yang berkaitan dengan Diabetes Mellitus.

Gejala Diabetes Melitus Menurut (Misnadiarly, 2006) Gejala Diabetes Mellitus dapat digolongkan menjadi gejala akut dan gejala kronik. Gejala akut Penyakit Diabetes Mellitus ini dari satu penderita ke penderita lainnya tidaklah selalu sama dan gejala yang disebutkan disini adalah gejala yang umum timbul dengan tidak mengurangi kemungkinan adanya variasi gejala lain, bahkan ada Penderita Penyakit Diabetes yang tidak menunjukkan gejala apa pun sampai pada saat tertentu. Berikut gejala akut Penyakit Diabetes Mellitus:

1. Pada permulaan gejala ditunjukkan meliputi tiga serba banyak yaitu banyak makan (*polifagia*), banyak minum (*polidipsia*), dan banyak kencing (*poliura*) atau disingkat "3P". Pada fase ini biasanya penderita menunjukkan berat badan yang terus naik dan bertambah gemuk, karena pada saat ini jumlah insulin masih mencukupi.

2. Bila keadaan tersebut tidak cepat diobati, lama kelamaan mulai timbul gejala yang disebabkan oleh kurangnya insulin. Jadi bukan 3P lagi melainkan hanya 2P (*polidipsiadan poliura*) dan 9 beberapa keluhan lain seperti :

1) Nafsu makan mulai berkurang.

2) Timbul rasa mual jika kadar glukosa darah melebihi 500mg/dl.

3) Banyak minum.

4) Banyak kencing.

5) Berat badan turun 5-10 kg dengan cepat dalam waktu 2-4 minggu.

6) Mudah letih

7) Bahkan penderita dapat mengalami *koma diabetik* (tidak sadarkan diri) jika tidak segera diobati. Dimana *koma diabetik* merupakan koma yang diakibatkan kadar glukosa darah terlalu tinggi (melebihi 600mg/dl)

Gejala kronik (menahun) merupakan gejala yang ditunjukkan sesudah beberapan tahun mengidap penyakit Diabetes Mellitus. Ciri-ciri gejala kronik yaitu:

1) Kulit terasa panas seperti ditusuk jarum

2) Kulit menebal

3) Mudah lelah

4) Mata minus

5) Untuk wanita akan ada gejala gatal di sekitar kemaluan

6) Gigi mudah goyah

7) Gairah seksual lemah

8) Terjadi keguguran pada ibu hamil

Fator Risiko Diabetes Mellitus yaitu faktor risiko bisa diubah dan tidak bisa diubah (Kemenkes, 2013). Faktor risiko yang tidak bisa diubah antara lain :

1. Gender

Gender merupakan sebuah alat vitual yang dimiliki oleh seseorang sejak lahir yang dibedakan antara lelaki dan wanita. Saat usia dewasa awal lelaki maupun wanita mengidap risiko yang sama untuk Penyakit Diabetes Mellitus, tetapi setelah usia 30 tahun, wanita memiliki resiko yang lebih tinggi dibandingkan lelaki (Riskesdas, 2017). Secara fisik wanita memiliki peluang mengidap Penyakit Diabetes Mellitus dengan meningkatnya indeks massa tubuh yang akan membuat penyaluran lemak tubuh lebih mudah terkumpul disebabkan proses hormon yang akan berisiko Penyakit Diabetes Mellitus Tipe 2 (Wahyuni, 2010).

2. Usia

Penyakit Diabetes Mellitus dapat terjadi pada berbagai usia, tetapi ada beberapa kisaran usia yang lebih rentan terhadap risiko pengembangan penyakit ini. Untuk usia muda (10-20 Tahun), terutama Diabetes Mellitus tipe 1, sering kali didiagnosis pada anak-anak dan remaja. Pada usia ini, sistem kekebalan tubuh dapat menyerang sel-sel penghasil insulin di pankreas, menyebabkan kadar insulin menjadi sangat rendah atau bahkan nol. Selain itu, Diabetes Mellitus tipe 2 juga semakin umum terjadi pada remaja, terutama dengan meningkatnya prevalensi obesitas dan gaya hidup tidak sehat, setelah itu ada usia Dewasa Muda (20-40 Tahun) pada usia dewasa muda, risiko Penyakit Diabetes Mellitus Tipe 2 mulai meningkat, terutama bagi mereka yang memiliki faktor risiko seperti riwayat keluarga diabetes, obesitas, atau gaya hidup sedentari. Penelitian menunjukkan bahwa pola makan yang buruk dan kurangnya aktivitas fisik di usia ini dapat berkontribusi pada pengembangan Penyakit Diabetes Mellitus dan yang terkahir terdapat pada usia Pertengahan (40 Tahun ke atas) setelah usia 40 tahun, risiko Diabetes Mellitus Tipe 2 meningkat secara signifikan. Ini disebabkan oleh penurunan sensitivitas insulin dan perubahan metabolisme seiring bertambahnya usia. Oleh karena itu, pemeriksaan rutin untuk kadar glukosa darah sangat dianjurkan bagi individu dalam kelompok usia ini. Pada usia 40 tahun fisiologis manusia mengalami perubahan lebih cepat, setelah lanjut usia Diabetes Mellitus Tipe 2 sering muncul terutama pada mereka yang obesitas, sehingga tubuhnya tidak peka terhadap insulin (Garnita, 2017).

3. Riwayat keluarga

Dengan Diabetes Mellitus Faktor genetik sangat mempengaruhi terjadinya Penyakit Diabetes Mellitus Tipe 2, jika salah satu orang tuanya menderita Penyakit Diabetes Mellitus maka resikonya 15% dan jika kedua orang tuanya menderita Penyakit Diabetes Mellitus maka resikonya adalah 75% pada anak, jika seorang ibu penderita Penyakit Diabetes Mellitus memiliki risiko 10-30% lebih besar dari pada seorang ayah, hal ini disebabkan lebih besar penurunan gen

sewaktu dalam kandungan ibu, jika saudara kembar identik menderita Diabetes Mellitus maka risiko adalah 10%-90% (Garnita, 2017).

Riwayat lahir dengan Berat Bayi Lahir Rendah (BBLR) Seseorang yang mengalami bayi berat lahir rendah memiliki kerusakan pankreas sehingga kemampuan pankreas untuk memproduksi insulin akan terganggu, hal ini kemungkinan dimasa dewasanya menderita Diabetes Mellitus tipe 2 (Kemenkes, 2010). Faktor yang dapat diubah Faktor resiko diabetes melitus yang dapat diubah meliputi, antara lain:

1. Gaya hidup yang tidak sehat. Kemunculan Penyakit Diabetes Mellitus di pengaruhi dari gaya hidup yang kurang sehat seperti pola makan yang tidak seimbang 12 dengan kadar kolesterol yang tinggi, rokok, alkohol, asupan gula yang berlebihan, minimnya olah raga dan porsi istirahat sampai stres (Khasanah, 2015).

2. Berat badan lebih (Obesitas) Derajat kegemukan dengan IMT > 23. IMT menjadi penting dalam konteks diabetes karena ada hubungan yang signifikan antara kegemukan atau obesitas dan risiko pengembangan diabetes, terutama Penyakit Diabetes Mellitus Tipe 2.

3. Kadar Glukosa Kadar glukosa yang tinggi dapat mempengaruhi kemampuan pembuluh darah untuk berfungsi dengan baik, yang dapat menyebabkan peningkatan tekanan darah. Ini terjadi karena pembuluh darah kehilangan kemampuan untuk melebar, dan jumlah cairan dalam tubuh dapat meningkat, yang semuanya berkontribusi pada hipertensi

4. Tekanan darah menjadi salah satu parameter penting dalam pengelolaan dan diagnosis penyakit diabetes karena adanya hubungan yang erat antara diabetes dan hipertensi (tekanan darah tinggi). Berikut adalah beberapa alasan mengapa tekanan darah dimasukkan ke dalam kriteria penyakit diabetes

Terdapat beberapa cara diagnosa Penyakit Diabetes Mellitus yang bisa dilakukan yaitu dengan mengukur kadar glukosa yang terkandung dalam darah, apabila kadar glukosa dalam darah konsentrasinya melebihi batas normal maka

orang tersebut dikategorikan penderita Penyakit Diabetes Mellitus (Detik health, 2019). Namun deteksi dini Penyakit Diabetes Mellitus diperlukan karena adanya fase asimtomatik yang cukup lama, fase asimtomatik adalah kondisi penyakit yang sudah positif diderita tetapi tidak menimbulkan gejala klinis pada penderita (Eldridge, 2022). Diagnosis dini Penyakit Diabetes Mellitus hanya mungkin dengan penilaian yang tepat dari kedua gejala tanda umum dan kurang umum, yang dapat ditemukan dalam fase yang berbeda dari inisiasi penyakit hingga diagnosis (Islam et al., 2020). Dalam konteks Penyakit Diabetes Mellitus, pengumpulan data dapat digunakan untuk mengidentifikasi faktor risiko, memprediksi perkembangan penyakit, dan mendukung pengambilan keputusan klinis (Meilani, 2014)

2.2.2 Data Mining

Data mining adalah metode yang dipakai untuk menggali informasi yang belum ditemukan dengan cara manual dari suatu kumpulan data disebut dengan data mining. Dengan data mining, informasi-informasi implisit dan berharga dari sebuah data dapat diekstrak. Data mining merupakan serangkaian proses untuk menggali 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. Data mining merupakan bagian dari proses penemuan pengetahuan dari *Basis Data Knowledge Discovery In*.

Ilmu data mining adalah perpaduan ilmu dari *Artificial Intelligence*, statistik, dan penelitian basis data yang selalu meningkat. Menurut artikel metode data mining merupakan sebuah proses menentukan ikatan yang mengandung arti, pola, dan keterkaitan dengan mengolah kelompok data. Dalam data mining terdapat 6 metode yang biasa di jalankan yaitu ramalan atau prediksi, penggambaran atau deskripsi, klasifikasi, estimasi, asosiasi dan clustering.

Berdasarkan Mardi (2017) dalam amna 2023, data mining dapat dikelompokkan menjadi beberapa kelompok, sesuai tugas yang dapat dilakukan yaitu:

1. Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecendrungan yang terdapat dalam data.

2. Estimasi

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 dan variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

3. Prediksi

Prediksi mempunyai kemiripan dengan teknik klasifikasi, tapi di sini data bagi kelas sesuai dengan perilaku atau nilai yang diprediksi pada waktu mendatang. Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang dan pendapatan rendah.

Data mining diterapkan untuk mendapatkan pola informasi baru berupa model dari suatu database (Anggraeni & Ramadhani, 2018). Peranan data mining dalam bidang kesehatan memiliki potensi besar terutama untuk menemukan pola-pola yang tersembunyi dalam suatu dataset rekam medis yang kemudian pola-pola tersebut dimanfaatkan untuk mendiagnosa awal penyakit secara klinis (Edy, 2012).

Beberapa metode algoritma dalam data mining yang sering digunakan untuk memprediksi seperti *Decision Tree C4.5, Naive Bayes* dan *K-Nearest Neighbors.* Akurasi model *Decision Tree* C4.5 sangat tergantung pada atribut pembentuknya, oleh karena itu pengolahan awal data sangat penting untuk menghasilkan model sederhana yang tingkat akurasinya tinggi (Lasut, 2012).

2.2.3 Klasifikasi

Klasfikasi artinya proses pola atau fungsi yang menyebutkan atau membedakan suatu konsep atau kelas data (Fadlan, Ningsih & Windarto, 2018). Teknik ini dilakukan pada data baru dengan memanipulasi data yang ada serta hasil yang diklasifikasikan untuk menyampaikan beberapa aturan (Asroni, Respati & Riyadi, 2018) Pada akhir pembelajaran, sebuah algoritma klasifikasi akan membentuk sebuah contoh klasifikasi dengan cara menganalisis data pelatihan (Ardiansyah & Walim, 2018)

Klasifikasi data merupakan suatu proses yang menemukan properti-properti yang sama pada sebuah himpunan obyek 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 data latih yang akan membedakan atribut ke dalam kategori atau kelas yang sesuai model. Metode Algoritma *K-Nearest Neighbor* melakukan klasifikasi berdasarkan pencocokan dari nilai sejumlah fitur tetangga terdekatnya dengan menghitung kedekatan kasus lama dengan kasus baru (Mustafa & Simpen, 2019)

2.2.4 Algoritma K-Nearest Neighbor

Metode *K-Nearest Neighbor* (K-NN) adalah teknik klasifikasi pembelajaran mesin non-parametrik yaitu, tidak ada asumsi untuk distribusi data yang mendasarinya. Dengan kata lain, struktur model ditentukan dari dataset dengan menyimpan semua kasus yang tersedia dan memprediksi kasus baru berdasarkan ukuran kesamaan (Sarker dkk, 2018)

K-Nearest Neighbor (KNN) dikerjakan dengan menentukan kelompok K objek pada data latih yang terdekat dengan tujuan pada data baru atau data pengujian. Dibutuhkan metode klasifikasi sebagai metode yang ahli menemukan informasi. Algoritma ini bertindak berdasarkan jarak terdekat dari data latih ke data pengujian untuk menetapkan *K-Nearest Neighbor*. sesudah mengumpulkan *K-Nearest Neighbors*, lalu diambil sebagian besar *K-Nearest Neighbors* untuk dibuat prediksi dari sampel uji (Argina A. , 2020).

Algoritma *K-Nearest Neighbor* menggunakan klasifikasi ketetanggaan sebagai nilai prediksi untuk *data query* yang baru atau *data testing*. Untuk menghitung jauh atau dekatnya tetangga dapat dihitung menggunakan rumus *Euclidean Distance* (Fitri Yunita, (2016) Menurut (Karegowda dkk, 2012) langkah yang dapat dilakukan untuk mengklasifikasikan data dengan menggunakan metode *k-Nearest Neighbor* adalah sebagai berikut :

1. Menyiapkan data latih dan data uji.
2. Mendefinisikan nilai *k*.
3. Melakukan perhitungan nilai jarak antara data latih dengan data uji, dengan rumus penghitung jarak *euclidean* sebagai berikut

$$d(x_i, x_j) = \sqrt{\sum_{i,j=1}^{n} (x_i - x_j)^2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.1)$$

Keterangan:

d (x_1, x_2) = Jarak

n = Variabel Data

N = Dimensi Data

Xi = Data latih (*data training*)

Xj = Data uji (*data testing*)

4. Pengelompokkan data berdasarkan perhitungan jarak.
5. Mengelompokkan data berdasarkan nilai tetangga terdekat.
6. Memilih nilai yang sering muncul dari tetangga terdekat sebagai acuan prediksi data selanjutnya.

Berdasarkan Kuhn dan Johnson (2013) dikatakan bahwa untuk dataset ukuran sedang direkomendasikan menggunakan nilai K=5-20, selain itu

berdasarkan Buku "*Concepts and Techniques*" yang ditulis oleh Han, dkk (2011) mengatakan bahwa pemilihan nilai K yang ganjil seringkali digunakan untuk menghindari bias dan resiko *overfitting*.

2.2.5 *Split Validation*

Metode *split validation* merupakan sebuah metode yang seluruh *data record* dan atribut digunakan dengan tujuan akan memiliki data set yang sesuai dengan asumsi-asumsi yang telah ditentukan. Data set yang akan ditentukan dibagi menjadi dua yaitu *data training* dan *data testing* dengan perbandingan dapat menggunakan 60:40, 70:30, 75:25, 80:20, 90:10, dsb. (Rahman et al., 2018).

Data training berfungsi untuk menciptakan model. Sedangkan, *data testing* merupakan sebuah pengujian dalam bentuk yang telah dibentuk dengan data lainnya dengan tujuan untuk mengetahui tingkat akurasi dari model tersebut (Nasution et al., 2019) Berdasarkan prinsip pareto yang terdapat pada buku saku analisis pareto yang ditulis oleh heru dan sunarto pada tahun 2020 dikatakan bahwa disarankan untuk menggunakan sekitar 70-80% data untuk pelatihan, sisanya 20-30% harus diberikan untuk pengujian. Setiap pengolahan data mining memiliki *split validation* yang berbeda-beda untuk memiliki hasil yang akurat (Witten et al., 2008).

2.2.6 *Confusion Matrix*

Confusion Matrix merupakan metode perhitungan guna melakukan analisis kualitas model klasifikasi dalam mengenali *tuple* dari kelas yang ada (Suyanto, 2018) *Confusion Matrix* merupakan pengukuran performa buat permasalahan klasifikasi *machine learning* dimana keluaran bisa berbentuk 2 kelas ataupun lebih. *Confusion Matrix* merupakan tabel dengan 4 campuran berbeda dari nilai prediksi serta nilai aktual (Hozairi Hozairi, 2021). *Confusion Matrix* merupakan tabel yang berisikan hasil dari proses klasifikasi. Sebagai contoh pada tabel 2.2 dibawah ini.

Tabel 2.2. Keterangan Posisi Tabel *Confusion Matrix*

	True 1	*True* 0
Prediction 1	TP (*True Positif*)	FP (*False Positif*)
Prediction 0	FN (*False Negatif*)	TN (*True Negatif*)

Berdasarkan tabel 2.2 dapat diketahui tabel confusion matrix akan menghasilkan nilai *True Positive, True Negative, False Positive*, dan *False Negative.* Ketika pengklasifikasi melakukan klasifikasi dan memiliki data yang benar, nilai *True Positive* dan *True Negative* akan berperan dalam memberikan informasi ini. Pada saat yang sama, jika pengklasifikasi membuat kesalahan dalam mengklasifikasikan data, nilai *False Positive* dan *False Negative* akan memberikan informasi ini (Han, Kamber, dan Pei 2011).

Setelah didapatkan nilai *Confusion Matrix* maka nilai yang telah didapatkan akan digunakan untuk mendapatkan nilai *accuracy, precission, recall* dan *Receiver Operation Characteristic* (ROC). Persamaan menurut (Mangande, 2020) untuk menghitung Nilai Akurasi:

$$Accuracy = \frac{TP + TN}{TP + TB + FP + FN} \times 100\% \quad . . . . . \quad (2.2)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad . . . . . \quad (2.3)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad . . . . . \quad (2.4)$$

Keterangan:

TP (*True Positive*) : Jumlah positif yang diklasifikasikan sebagai positif

TN (*True Negative*) : Jumlah negatif yang diklasifikasikan sebagai negatif

FP (*False Positive*) : Jumlah negatif yang diklasifikasikan sebagai positif

FN (*False Negative*) : Jumlah positif yang diklasifikasikan sebagai negatif

2.2.7 *F1 Score*

F1 Score merupakan ukuran performa model *machine learning* yang dikembangkan oleh Rijsbergen tahun 1979 untuk menilai keseimbangan antara presisi dan recall. *F1 Score* didefinisikan sebagai ukuran keselarasan antara presisi dan recall, *F1 Score* juga dikenal sebagai *F-Measure* atau *F-Score*. (Rijsbergen, 1979 dalam khun, dkk 2013). Rumus *F1 Score* adalah sebagai berikut:

$$F1\text{-}Score = \frac{2 \text{ x (Presisi x Recall)}}{\text{(Presisi + Recall)}} \quad . . . . . (2.5)$$

F1 Score merupakan ukuran yang lebih baik daripada presisi atau recall saja, karena mempertimbangkan keseimbangan antara keduanya (Powers, 2011). *F1 Score* memiliki beberapa kelebihan, seperti:

1. Mempertimbangkan keseimbangan antara presisi dan recall.

2. Lebih akurat daripada presisi atau recall saja.

3. Dapat digunakan untuk menilai performa model pada data tidak seimbang. *F1 Score* juga memiliki beberapa kekurangan, *F1 Score* sensitif terhadap perubahan data dan tidak mempertimbangkan biaya kesalahan, sehingga perlu dipertimbangkan dalam konteks aplikasi (Provost & Fawcett, 2013).