

## BAB II

### STUDI PUSTAKA

#### 2.1. Penelitian Terkait

Penulisan ini tidak lepas dari hasil penelitian yang pernah dialami sebelumnya sebagai suatu bahan perbandingan kajian. Hasil dari penelitian yang dijadikan sebagai pembanding tidak dapat dipisahkan dengan masalah penelitian yaitu penemuan dan diagnosis *penyakit diabetes melitus*. Berikut merupakan beberapa penelitian yang digunakan sebagai bahan rujukan :

Penelitian yang dilakukan oleh Hadli Fadli Santoso, Eko Darwiyanto, dan Untari Novia Wisesty dengan judul “Analisis dan Implementasi *Metode Certainty Factor* Pada Sistem Pendukung Pengambilan Keputusan untuk Penyakit *Diabetes Melitus*”. *Diabetes melitus* adalah penyakit dimana tingkat gula darah seseorang meningkat karena masalah *insulin*. Ini adalah penyebab kematian keempat terbesar di dunia. Salah satu penyebab meningkatnya angka tahunan dari penyakit ini yaitu tidak sehatnya gaya hidup. Peningkatan angka penderita *diabetes melitus* tersebut tidak diimbangi dengan jumlah tenaga profesional. Sebagian besar penderita *diabetes melitus* tidak terdiagnosa karena minimnya tenaga profesional serta pengetahuan masyarakat terkait penyakit mematikan ini. Hasil penelitian ini digunakan untuk mendukung sistem pengambilan keputusan para spesialis yang mendukung diagnosis diabetes dini. Metode yang dipergunakan dalam sistem ini ialah metode *Certainty Factor*. Pasien dapat didiagnosis dengan SPPK *diabetes melitus* dengan akurasi 80%, sesuai dengan temuan implementasi dan pengujian, sistem pendukung keputusan ini dinilai sangat membantu dalam proses diagnosa *diabetes melitus*.(Santoso, Darwiyanto, & Wisesty, 2016)

Penelitian yang dilakukan oleh Risky Aswi Ramadhani dan Ratih Kumalasari Niswatin dengan judul “Sistem Diagnosa *Diabetes* Menggunakan Metode K-NN”. *Diabetes melitus* adalah faktor risiko penyakit jantung, *glaukoma*, dan kerusakan ginjal. Saat ini terdapat 10 juta penderita *diabetes melitus* di Indonesia, namun hanya sedikit orang yang mengetahui bahwa dirinya terjangkit penyakit *diabetes melitus*. Hal ini dikarenakan terbatasnya jumlah tenaga medis, utamanya kota-kota kecil. Sebuah teknik diagnosis yang bertujuan untuk

mengurangi jumlah *diabetes melitus* ditetapkan untuk meminimalkan jumlah penderita *diabetes melitus* dan meningkatkan kesadaran publik akan bahaya *diabetes melitus*. Pendekatan KNN digunakan dalam sistem ini, yang mencakup enam variabel: di malam hari sering buang air kecil, luka/goresan sulit sembuh, mudah lelah, gusi berdarah, penurunan cepat dan tiba-tiba dari berat badan, dan keinginan kuat untuk makan dan minum. Data yang dipergunakan pada saat penelitian ini sebanyak 24 data latih dan 7 data uji, serta menggunakan nilai  $k=3$ . Dari hasil implementasi dan pengujian, disimpulkan bahwa sistem diagnosa *diabetes melitus* berbasis KNN dapat membantu pemerintah mengurangi jumlah penderita diabetes. Hasil uji akurasi metode KNN mencapai 85,71%. Berdasarkan temuan ini, pendekatan KNN adalah pilihan terbaik untuk mengembangkan sistem diagnosis *diabetes melitus*.(Ramadhani & Niswatin, 2018).

Penelitian yang dilakukan oleh Noviandi dengan judul “Implementasi Algoritma Decision Tree C4.5 Untuk Prediksi Penyakit *Diabetes*”. *Diabetes melitus* adalah kondisi kronis yang mengancam jiwa. Kenaikan yang tidak terkontrol pada gula darah dapat mengakibatkan konsekuensi terjadinya komplikasi yang serius. Jumlah penderita penyakit ini setiap tahun selalu meningkat. *Diabetes melitus* mempengaruhi 382 juta orang yang ada didunia pada tahun 2013. Penelitian ini bertujuan untuk membangun model prediksi kemungkinan seseorang terjangkit penyakit *diabetes melitus* menggunakan algoritma C4.5. Database Pima Indians Diabetes (PPID) digunakan untuk membuat prediksi yang berjumlah 768 data dengan delapan variabel yaitu jumlah kehamilan, kadar *glukosa*, tekanan darah, ketebalan lipatan kulit *trisep*, kadar *insulin*, *index masa tubuh (IMT)*, riwayat *diabetes melitus* dalam keluarga, serta usia. Akurasi hasil prediksi sebesar 70.32% dengan 9 rule, yaitu 4 rule untuk class tidak dan 5 rule untuk class iya.(Noviandi, 2018)

Berdasarkan rujukan dari tiga jurnal penelitian yang telah dilakukan sebelumnya, dapat disimpulkan bahwa perbandingan yang dijadikan alasan rujukan penelitian terletak pada metode penelitian guna memperoleh hasil sesuai dengan yang diharapkan. Pada penelitian saat ini mengajukan topik yang serupa dengan metode penelitian yang berbeda. Penelitian ini akan menggunakan metode *K-Nearest Neighbor* (KNN). Metode klasifikasi *K-Nearest Neighbor* (KNN)



merupakan salah satu metode yang digunakan untuk mengklasifikasikan objek berdasarkan data latih dengan jarak terdekat. Teknik *K-Nearest Neighbor (KNN)* bertujuan untuk mengkategorikan objek baru menggunakan atribut dan sampel dari data latih. Dengan menerapkan metode klasifikasi *K-Nearest Neighbor (KNN)*, diharapkan dapat ditemukan aturan klasifikasi yang dapat dipergunakan membangun sistem pendukung keputusan agar dapat mengidentifikasi seseorang terserang *diabetes melitus* atau tidak sehingga kemungkinan terburuk akibat penyakit ini dapat dicegah.

## 2.2. Landasan Teori

### 2.2.1. *Diabetes Melitus*

*Diabetes melitus* merupakan penyakit metabolik yang ditandai dengan *hiperglikemia* akibat terganggunya kinerja hormon insulin. (American Diabetes Association, 2017). *Hiperglikemia* atau kadar gula darah tinggi terjadi karena tidak bisa menghasilkan cukup *insulin* pada *pankreas*, maupun akibat disfungsi *insulin*. *Diabetes melitus* dapat membahayakan organ dan pembuluh darah, seperti mata, ginjal, saraf, dan jantung. (American Diabetes Association, 2020).

*Diabetes melitus* atau kencing manis yaitu penyakit kronis di mana tubuh tidak dapat membuat cukup *insulin* atau tidak menggunakannya dengan benar. Penyakit ini diagnosa melalui pengamatan kadar gula dalam darah. *Insulin* bertanggung jawab untuk menggunakan *glukosa* dalam aliran darah sebagai sumber energi ke sel-sel tubuh. (IDF, 2019).

Berdasarkan definisi di atas, *diabetes melitus*, kadang-kadang dikenal sebagai "gula manis", adalah penyakit yang didefinisikan oleh ketidakmampuan tubuh untuk mencerna karbohidrat, lipid, dan protein, yang mengakibatkan kadar gula darah tinggi (*hiperglikemia*).

Penyebab seseorang menderita *diabetes melitus* bisa dikarenakan banyak faktor, faktor tersebut diantaranya tekanan darah tinggi, obesitas, riwayat *diabetes melitus* dalam keluarga, usia, serta pola hidup dan diet tidak sehat. Faktor-faktor ini dapat terbagi menjadi 2 kategori: faktor risiko yang tidak bisa diubah dan faktor risiko yang bisa diubah. Faktor risiko

yang tidak bisa diubah adalah faktor keluarga dengan riwayat *diabetes melitus.*, serta usia  $\geq 45$  tahun. Sedangkan faktor risiko yang bisa diubah adalah obesitas ( $IMT \geq 25 \text{kg/m}^2$ ), kurangnya aktivitas fisik, hipertensi atau tekanan darah tinggi, serta pola hidup dan diet tidak sehat. (American Diabetes Association, 2018)

### 2.2.2. Sistem Pendukung Keputusan

Kata "sistem" berasal dari kata Yunani "*systema*" yang berarti "bersatu" dan mengacu pada entitas yang terdiri dari komponen atau bagian yang saling berhubungan. Sistem Pendukung Keputusan (SPK) adalah konsep khusus dari sistem yang menghubungkan informasi terkomputerisasi dengan pengguna atau pengambil keputusan. Istilah "sistem pendukung keputusan" mengacu pada sistem yang menggabungkan komputer ke dalam semua proses pengambilan keputusan. (Sihotang & Siboro, 2016).

Pada awal 1970-an, Michael S. Scott Morton pertama kali menggunakan istilah *Management Decision System* untuk Sistem Pendukung Keputusan (SPK). Sistem berbasis komputer tersebut dirancang untuk membantu proses pengambilan keputusan dengan menggunakan data dan model tertentu untuk memecahkan berbagai masalah yang tidak terstruktur. Setelah itu SPK mulai dikembangkan hingga menghasilkan sistem berbasis komputer yang dapat memecahkan berbagai macam masalah, mulai dari masalah yang tidak terstruktur sampai masalah yang semi terstruktur (Nofriansyah & Defit, 2017).

Berdasarkan uraian diatas dapat ditarik kesimpulan bahwa Sistem Pendukung Keputusan (SPK) adalah sebuah sistem yang dirancang dengan tujuan untuk membantu perorangan, organisasi maupun perusahaan dalam proses pengambil keputusan.

### 2.2.3. *K-Nearest Neighbor* (KNN)

Algoritma *K-Nearest Neighbor* (KNN) adalah metode klasifikasi objek atau data baru berdasarkan data/tetangga (*neighbor*) terdekat dari data

tersebut. Metode KNN dibagi menjadi dua tahap, yaitu tahap pembelajaran (*training*) dan tahap klasifikasi atau pengujian (*testing*). Pada tahap pembelajaran, dilakukan penyimpanan fitur serta klasifikasi dari data pembelajaran (*training*), sedangkan pada tahap pengujian, proses perhitungan dilakukan terhadap data baru yang akan dilakukan tindakan uji coba (yang belum diketahui klasifikasinya). Data yang digunakan pada tahap pembelajaran disebut dengan data *training* atau data latih, data ini yang digunakan untuk melatih algoritma. Sedangkan data yang digunakan pada tahap pengujian disebut dengan data *testing* atau data uji, data ini dipergunakan agar dapat menguji dan mengetahui kinerja algoritma yang telah dilatih sebelumnya ketika data baru tersedia. (Baharuddin, Hasanuddin, & Azis, 2019).

Langkah-langkah yang dilakukan dalam penerapan algoritma KNN yaitu:

1. Menentukan nilai k.

Nilai k optimal dalam algoritma ini tergantung pada jumlah data. Nilai k yang besar belum tentu menghasilkan tingkat akurasi yang tinggi, begitupun pula sebaliknya. Nilai k yang digunakan biasanya merupakan angka ganjil hal itu untuk menghindari jumlah label classimbang, sehingga akan menyulitkan proses penarikan kesimpulan. Pencarian nilai k optimal merupakan kunci dari keberhasilan algoritma ini.

2. Menghitung jarak masing-masing objek terhadap data baru. Terdapat empat metode pengukuran jarak yaitu *euclidean*, *manhattan*, *minkowski* dan *chebyshev*.

a. Rumus *Euclidian Distance*.

$$d(x, y) = \left( \sum_{k=1}^m (x_{ik} - x_{jk})^2 \right)^{1/2} \quad (2.1)$$

- Dimana :
- $d(x,y)$  = jarak antara  $x_{ik}$  dan  $x_{jk}$
  - $x_{ik}$  = nilai X pada data latih
  - $x_{jk}$  = nilai X pada data uji
  - m = batas jumlah banyaknya data

Ide rumus ini berasal dari rumus *pythagoras*, yang banyak digunakan untuk mencari sisi miring dari sebuah segitiga.



b. Rumus *Manhattan Distance*.

$$d(x, y) = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (2.2)$$

Dimana :  $d(x,y)$  = jarak antara  $x_{ik}$  dan  $x_{jk}$   
 $x_{ik}$  = nilai X pada data latih  
 $x_{jk}$  = nilai X pada data uji  
 $m$  = batas jumlah banyaknya data

Rumus ini mencari jarak dengan menjumlahkan semua selisih dari jarak  $x_{ik}$  dan  $x_{jk}$

c. Rumus *Minkowski Distance*.

$$d(x, y) = \left( \sum_{k=1}^m |x_{ik} - x_{jk}|^r \right)^{1/2} \quad (2.3)$$

Dimana :  $d(x,y)$  = jarak antara  $x_{ik}$  dan  $x_{jk}$   
 $x_{ik}$  = nilai X pada data latih  
 $x_{jk}$  = nilai X pada data uji  
 $m$  = batas jumlah banyaknya data  
 $r$  = kuadrat jarak

Ide dari rumus ini diambil dari konsep aljabar dengan objek yang berupa vektor berdimensi n. Nilai r yang digunakan dalam rumus ini berupa nilai bebas yang bukan 1 dan 2, karena jika  $r = 1$  maka akan diperoleh rumus jarak *manhattan distance*, dan jika  $r = 2$  maka akan diperoleh rumus jarak *euclidean distance*.

d. Rumus *Chebyshev Distance*.

$$d(x, y) = \max_{k=1} |x_{ik} - x_{jk}| \quad (2.4)$$

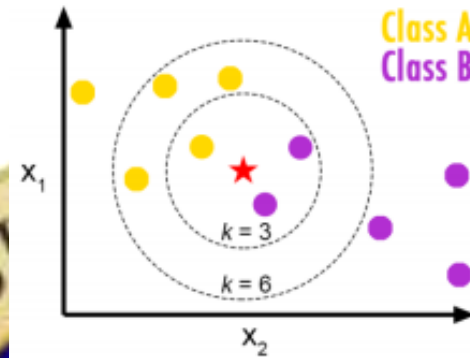
Dimana :  $d(x,y)$  = jarak antara  $x_{ik}$  dan  $x_{jk}$   
 $x_{ik}$  = nilai X pada data latih  
 $x_{jk}$  = nilai X pada data uji

Rumus ini mencari jarak yang terbesar antara  $x_{ik}$  dan  $x_{jk}$

3. Mengurutkan objek dari mulai jarak terdekat sampai jarak terjauh  
Semakin besar hasil perhitungan jarak *euclidean*, maka jarak antara kedua objek semakin jauh, dan semakin kecil hasil perhitungan jarak *euclidean*, maka jarak antara kedua objek semakin dekat.
4. Mengumpulkan label class Y (klasifikasi *nearest neighbor*).

5. Melakukan prediksi berdasarkan label class Y yang paling mayoritas.

Dengan ilustrasi berikut, data baru akan ditentukan berdasarkan klasifikasi tetangga/tetangga yang paling banyak di sekitarnya, berikut ilustrasinya :



**Gambar 2. 1** Ilustrasi penggunaan nilai k pada metode KNN

Berikut adalah kelebihan serta kekurangan metode KNN :

a. Kelebihan

- Pelatihan sangat cepat
- Sederhana dan mudah dimengerti
- Efektif untuk data besar
- Tahan terhadap data yang memiliki derau

b. Kekurangan

- Nilai k menjadi bias (perlu menentukan parameter k)
- Nilai komputasi yang kompleks
- Keterbatasan memori
- Mudah teripu dengan atribut yang tidak relevan

#### 2.2.4. Data Mining

Data mining atau disebut juga penambangan data adalah proses mengekstraksi serta mengidentifikasi informasi dengan menggunakan kecerdasan buatan, statistik, matematika, dan pembelajaran mesin (Karim, Salim, & Azis, 2018). Data mining, juga dikenal sebagai penambahan data, adalah proses yang meneliti nilai tambahan dari kumpulan data dalam bentuk pengetahuan yang sebelumnya tidak diketahui. Data mining adalah

proses pengolahan data secara otomatis, Data yang diproses biasanya adalah data dalam jumlah yang sangat besar. Tujuan dari data mining adalah untuk menemukan pola atau hubungan yang dapat memberikan informasi yang berguna.(Mardi, 2017)

Menurut (Cahyanti, Rahmayani, & Husniar) data mining diartikan sebagai proses penemuan pola pada suatu data. Adapun tugas dari data mining dibagi menjadi enam kelompok yaitu, deskripsi, estimasi, asosiasi, klasifikasi, prediksi, dan *clustering*.

Data mining merupakan suatu proses untuk menemukan informasi-informasi terbaru yang bermanfaat dan belum diketahui sebelumnya di dalam suatu *database* yang besar, di mana proses yang digunakan adalah teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin (Mulyati, Husein, & Ramdhan, 2020). Terdapat tiga karakteristik dari data mining yaitu:

1. Data mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. Data mining biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
3. Data mining berguna untuk membuat keputusan yang kritis, terutama dalam strategi

#### **2.2.5. Confusion Matrix**

*Confusion matrix* adalah suatu metode evaluasi yang digunakan untuk melakukan perhitungan akurasi pada konsep data mining. Akurasi dalam klasifikasi adalah persentase ketepatan rekam data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Sari, Astuti, Saluza, Faradillah, & Yunita, 2021). Untuk menghitung tingkat akurasi (*accuracy*), terdapat 4 variabel kemungkinan dari hasil deteksi yang harus dipahami yaitu nilai *True Positive*, *True Negative*, *False Positive*, dan *False Negative*. Penjelasan tentang variabel-variabel tersebut akan dirangkum pada **Tabel 2.1**.



**Tabel 2. 1** Confusion Matrix

	<i>Predicted Class</i>		
		<i>Class Yes</i>	<i>Class No</i>
<i>Actual Class</i>	<i>Class Yes</i>	<i>True Positive</i> (TP)	<i>False Negative</i> (FN)
	<i>Class No</i>	<i>False Positive</i> (FP)	<i>True Negative</i> (TN)

- True Positive* (TP) atau Benar Positif adalah kasus dimana data diprediksi positif oleh sistem dan prediksi sistem memang benar adanya.
- True Negative* (TN) atau Benar Negatif adalah kasus dimana data diprediksi negatif oleh sistem dan prediksi sistem memang benar adanya.
- False Positive* (FP) atau Salah Positif adalah kasus dimana data diprediksi positif oleh sistem tetapi kenyataannya adalah data bernilai negatif.
- False Negative* (FN) atau Salah Negatif adalah kasus dimana data diprediksi negatif oleh sistem tetapi kenyataannya adalah data bernilai positif.

*Accuracy* atau akurasi merupakan rasio prediksi benar positif dan benar negatif (TP + TN) dibandingkan dengan keseluruhan data (TP + FP + TN + FN). Rumus perhitungan akurasi dapat dilihat dibawah ini.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.5)$$