BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Hasran(2020) dalam "Klasifikasi Penyakit Jantung menggunakan metode K-nn (Studi Kasus: Lab.Riset Fakultas Ilmu Komputer Universitas Muslim Indonesia)". Penelitian ini mencakup pengukuran performa (akurasi, presisi, *recall* dan *f-measure*) metode KNN dengan nilai K 3 hingga 9 pada objek 1000 data pasien penyakit jantung yang diperoleh dari pusat dataset *UCI Machine Learning Repository.* Hasil dari pengukuran performa diperoleh nilai K terbaik adalah 6 dimana nilai akurasi 85%, presisi 78%, *recall* 93% dan *f-measure* sebesar 85%.

Ikhsan Nuh Atthalla, Adithia Jovandy, hanif Habibie, 2020 dalam "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode *K-Nearest Neighbor* Studi Kasus(Universitas Sriwijaya palembang)". Pada penelitian ini penghitungan jarak kemiripan menggunakan jarak *minkowski*, Jarak *minkowski* adalah jarak di dalam ruang vektor yang telah ditentukan yang bisa dianggap sebagai generalisasi dari kedua jarak *Euclidean* dan jarak *Manhattan*. Dengan metode KNN diperoleh hasil paling akurat sebesar 80%.

Zuriati Z, Qomariyah N, 2023 dalam "Klasifikasi penyakit *stroke* menggunakan metode KNN Studi kasus (Politeknik Negeri Lampung)". Tujuan penelitian adalah untuk menerapkan algoritma KNN untuk klasifikasi data penyakit *stroke*. Percobaan dilakukan dengan melakukan 3 skenario *split* data *training* dan *testing* dengan perbandingan: 90%:10%, 80%:20% , 70%:30%. Selain itu juga melakukan percobaan untuk nilai k=3 dan k=5 untuk menemukan akurasi terbaik.

Nilai k menyatakan berapa banyak jumlah *neighbo*r atau data yang terdekat dengan suatu objek bahwa dengan uji coba nilai k yang berbeda-beda, menghasilkan tingkat akurasi yang berbeda pula. Performa algoritma KNN terbaik didapatkan pada percobaan dengan pembagian data *training* dan data *testing* 90% : 10% dengan nilai k= 5 dan akurasi 85%.

Syahrani Lonang, Dwi, Normawati, 2021 dalam "Klasifikasi *Stunting* Pada Balita menggunakan metode KNN Studi Kasus (Puskesmas Ubung, Kecamatan Jonggat, Kabupaten Lombok Tengah)". Pada penelitian ini disimpulkan bahwa peneliti dapat menerapkan dan membuat sistem klasifikasi status *stunting* pada balita dengan menerapkan metode *k-nearest neighbor* dengan seleksi fitur *backward elimination*. Tingkat akurasi yang didapatkan *k-nearest neighbor* dengan fitur seleksi *backward elimination* mencapai 70% meningkat 0,30%.

Zulfachmi, Andi Firman Syahputra, Bayu Indra Prasetyo, Aurora Elsa Shafira, 2023 dalam "Klasifikasi Tingkat Dehidrasi Berdasarkan Warna *Urin* menggunakan metode KNN Studi Kasus (kampus STT Indonesia Tanjung Pinang)". Pada penelitian ini Dari 50 sampel yang diambil terdapat urin yang diklasifikasi sebagai dehidrasi berjumlah 30, dan yang tidak dehidrasi sebanyak 20 melalui proses RGB. Tingkat akurasi dari metode *K-Nearest Neighbor* adalah 83,7%, 26 sampel dengan klasifikasi akurat, dan 4 sampel yang tidak akurat.

Dari penelitian yang telah dilakukan oleh peneliti-peneliti sebelumnya, terdapat perbedaan yang bisa dilihat pada tabel 2.1.

Tabel 2. 1 Perbandingan Penelitian

Penulis, Tahun	Objek	Masalah	Metode	Hasil
Hasran,2020	Jantung	Penyakit Jantung	*K Nearest Neighbor*	Hasil dari pengukuran performa diperoleh nilai K terbaik adalah 6 dimana nilai akurasi 85%
Ikhsan Nuh Atthalla, Adithia Jovandy, hanif Habibie,2018	Kanker Payudara	Payudara dalam tingkat ganas	*K Nearest Neighbor*	Klasifikasi penyakit kanker payudara memperoleh hasil paling akurat 80%
Zuriati Z, Qomariyah N, 2023	*Stroke*	Penyakit *Stroke*	*K Nearest Neighbor*	Klasifikasi menggunakan algoritma KNN untuk dataset penyakit *stroke* telah berhasil dilakukan,nilai akurasi tertinggi didapatkan pada nilai k = 5 dengan komposisi perbandingan data *training* dan *testing* 90% : 10% dengan nilai akurasi 80%
Syahrani Lonang, Dwi Normawati, 2021	*Stunting*	Klasifikasi *Stunting* Pada Balita	*K Nearest Neighbor*	Tingkat akurasi yang didapatkan *k-nearest neighbor* dengan *fitur seleksi backward elimination* mencapai 92,2% Meningkat 0,30%
Zulfachmi, Andi Firman Syahputra, Bayu Indra Prasetyo, Aurora Elsa Shafira, 2023	*Urin*	Klasifikasi Tingkat Dehidrasi Berdasarkan warna *Urin*	*K Nearest Neighbor*	Tingkat akurasi dari metode *K-Nearest Neighbor* Adalah 83,7% dengan 26 sampel dengan klasifikasi akurat,dan 4 sampel yang tidak akurat
Maria Hawila Katharina Runesi, 2023	Anemia	Klasifikasi Penyakit Anemia	*K Nearest Neighbo*r	Memprediksi apakah pasien mengidap penyakit anemia atau tidak anemia

2.2 Dasar Teori

2.2.1 Anemia

Menurut *World Health Organization* (WHO) Anemia (dalam bahasa Yunani: ἀναιμία anaimia, artinya kekurangan darah, dari ἀν- an-, "tidak ada" + αἷμα haima, "darah", disebut juga kurang darah) adalah keadaan saat jumlah sel darah merah atau jumlah hemoglobin (protein pembawa oksigen) dalam sel darah merah berada di bawah normal. Sel darah merah mengandung hemoglobin yang berperan mengangkut oksigen dari jantung yang diperoleh dari paru-paru, dan kemudian mengantarkannya ke seluruh bagian tubuh. Menurut dokter Rumah Sakit Umum Daerah Naibonat kab.kupang, Yulia Masneno mengatakan bahwa kekurangan darah merah disebut anemia dan yang normal atau tidak kekurangan darah merah disebut tidak anemia. Seseorang dikatakan anemia bila konsentrasi hemoglobin (Hb) nya kurang dari 13,5 g/dL atau hematokrit (Hct) kurang dari 41% pada laki-laki, dan konsentrasi Hb kurang dari 11,5 g/dL atau Hct kurang dari 36% pada perempuan (WHO, 2024).

Data WHO menunjukan total penduduk dunia yang mengalami anemia adalah 1,62 miliyar orang, dengan prevalensi data 305 juta diantaranya adalah anak sekolah. Penderita anemia di Indonesia tergolong tinggi lantaran tidak memenuhi standar Organisasi Kesehatan Dunia (WHO) sebesar 20%, terutama anak-anak hingga remaja (Yanti, 2017). Berdasarkan data Kementerian Kesehatan (Kemenkes) 2018, penderita anemia di Indonesia kurang lebih 7.5 juta orang. Penderita kekurangan kadar hemoglobin pada balita di tanah air mencapai 38,5%, usia sekolah 26,5%, dan remaja (15-24 tahun) 32%. Prevalensi anemia berdasarkan

daerah menunjukkan bahwa penderita yang tinggal di pedesaan memiliki angka lebih tinggi (22,8%) dibandingkan yang tinggal di perkotaan (20,6%). Sementara itu, prevalensi anemia pada remaja berusia lebih dari 15 tahun sebesar 22,7%.

Hemoglobin merupakan zat warna yang terdapat dalam darah merah yang berguna untuk mengangkut oksigen ($O_{2)}$ dan karbondioksida CO_2 dalam tubuh. Hemoglobin merupakan parameter luas untuk menentukan status anemia pada skala luas, (Adriani & Wirjatmandi, 2012). Indeks Eritrosit atau *Mean Cospuscular Value* adalah nilai rata-rata yang dapat memberi keterangan mengenai rata-rata eritrosit dan mengenai banyak hemoglobin per-eritrosit. Pemeriksaan MCV digunakan sebagai pemeriksaan penyaring untuk mendiagnosis terjadinya anemia dan mengetahui anemia berdasarkan morfologinya (Gandasoebrata R,2013).

MCV (Mean Corpuscular Volume) atau VER (Volume Rata-rata) adalah volume rata-rata sebuah eritrosit yang dinyatakan dengan satuan *femtoliter*(fl), Nilai normal MCV = 82-92 fl. MCH (*Mean Corspuscular Hemoglobin*) atau HER (*Hemoglobin Eritrosit Rata-rata*) adalah jumlah hemoglobin per-eritrosit yang dinyatakan dengan satuan pikogram(pg), Nilai normal MCH = 27-31 pg. MCHC (*Mean Corspuscular Hemoglobin Concentration*) atau KHER (Konsentrasi Hemoglobin Eritrosit Rata-rata) adalah konsentrasi hemoglobin yang didapat per-eritrosit yang dinyatakan dengan satuan gram per *desiliter* (gr/dl), nilai normal MCHC = 30-55 gram (Adriani & Wirjatmandi, 2012).

2.2.2 Data Mining

Secara sederhana data *mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies, 2004). Data *mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pramudiono, 2007). Data mining, sering juga disebut sebagai *knowledge discovery in database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, *history* untuk menemukan keteraturan, pola atau hubungan dalam *set* data berukuran besar(Santoso, 2007).

Karakteristik data mining sebagai berikut:

a. Data *mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
b. Data *mining* biasa menggunakan data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil lebih percaya.
c. Data *mining* berguna untuk membuat keputusan yang kritis, terutama dalam strategi (Davies, 2004).

2.2.3 Klasifikasi Data Mining

Klasifikasi data mining adalah sebuah proses menemukan definisi kesamaan karakteristik dalam suatu kelompok atau kelas (*class*). Metode klasifikasi merupakan teknik yang didasarkan pada atribut dari kelompok yang sudah didefenisikan, sehingga didapatkan suatu aturan yang digunakan untuk melakukan klasifikasi pada data dengan cara memanipulasi data yang sudah ada dan sudah diklasifikasi. (Novriansyah & Nurcahyo, 2015).

Metode ini termasuk ke dalam kelompok *supervised learning* yang setiap *item* datanya memiliki label atau kelas yang pengaruhi atribut. Tipe data yang cocok digunakan pada metode klasifikasi yaitu biner atau nominal sedangkan untuk tipe data ordinal kurang cocok sebab pada metode ini menggunakan pendekatan secara implisit. (Novriansyah & Nurcahyo, 2015).

2.2.4 K-Nearest Neighbor

Nearest Neighbor merupakan salah satu metode yang digunakan dalam menyelesaikan masalah pengklasifikasian. Algoritma K-nn didasarkan pada pembelajaran dengan analogi yaitu membandingkan data uji yang diberikan dengan data latih yang serupa. Dimana data latih diekspresikan oleh n-atribut yang kemudian setiap *record* pada data latih disimpan dalam n-dimensi. Sehingga, ketika memberikan suatu *record* data yang belum diketahui maka K-nn akan mencari pola untuk K data latih yang paling dekat dengan *record* yang belum diketahui. (Han,kamber, & Pei,2012).

Menurut (Suyanto,2017) ada beberapa hal menarik pada algoritma K-nn yaitu mudah diimplementasikan hanya dengan menggunakan cara sederhana dengan menentukan satu parameter K dan algoritma K-nn bekerja secara lokal dengan hanya memperhitungkan sejauh K data. Namun, disisi lain KNN juga memiliki kelemahan yaitu sangat sensitif terhadap *noise* ataupun *outlier* pada data. Selain itu, Algoritma ini kesulitan menentukan parameter K dalam proses pelatihan.

Langkah-langkah algoritma K-nn yaitu:

1. Memasukkan data latih dan data uji
2. Menghitung nilai K

3. Menghitung jarak *Eucledian* setiap data latih terhadap data uji menggunakan rumus

$$(x, y) = \sqrt{\sum_{I=1}^{n}(xi_{training} - yi_{testing})^2} \quad \ldots\ldots (1)$$

Keterangan :

D(x,y): jarak antara data latih dengan data uji,

n : jumlah data latih

x : data latih

y : data uji

4. Mengurutkan hasil perhitungan jarak mulai dari yang terkecil ke yang terbesar.
5. Mengumpulkan atau mengambil sejumlah data sesuai nilai K yang telah ditentukan pada langkah 2.
6. Menentukan hasil dari pengambilan data berdasarkan tetangga terdekat pada langkah 4 dapat diklasifikasikan berdasarkan kategori yang ditentukan.

2.2.5 Flowchart KNN

Gambar 2. 1 Flowchart K-nn

Berdasarkan gambar 2.1 proses pertama yang dilakukan adalah *input*data. Setelah *input* data dilanjutkan dengan menghitung jarak *euclidean* antara data yang akan diuji. Selanjutnya mengurutkan hasil jarak *euclidean* dari yang terkecil hingga yang terbesar kemudian menentukan jarak terdekat dari hasil pengurutan, dan langkah terakhir adalah menentukan hasil jarak terdekat yang dipilih. Berikut adalah langkah-langkah menghitung *jarak euclidean distance*:

1. Memasukkan data latih dan data uji

   Data latih dan data uji yang digunakan untuk menghitung manual jarak *euclidean distance* menggunakan data *real* yang diambil secara *random*. Untuk data pengujian dan pelatihan dapat dilihat pada tabel 2.2

Tabel 2. 2 Hitung Euclidean Distance

x1(Hb)	x2(MCH)	y1(MCHC)	y2(MCV)	Kelas
14.9	22.7	29.1	83.7	1
15.9	25.4	28.3	72	1
9	21.5	29.6	71.2	0
14.7	16	31.4	87.5	1
11.6	22.3	30.9	74.5	0
14.1	29.7	30.5	30.5	0
12.7	19.5	28.9	82.9	1

2. Menghitung nilai K = 5
3. Menghitung jarak *Euclidean* setiap data latih dan data uji menggunakan rumus :

$$(x, y) = \sqrt{\sum_{I=1}^{n}(xi_{training} - yi_{testing})^2} \qquad \text{.....(2)}$$

Keterangan :

D(x,y) : Jarak antara data latih dengan data uji

n : Jumlah data latih

x : Data latih

y : Data uji

Berdasarkan *dataset* diatas dimana Hemoglobin = x1, MCH =x2, MCHC = y1, dan MCV = y2.

a) Menghitung jarak dataset 1

$$D = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

$$D1 = \sqrt{(22.7 - 14.9)^2 + (83.7 - 29.1)^2}$$

$$D1 = \sqrt{60.84 + 605.15}$$

$$D1 = \sqrt{665.99}$$

$$D1 = \sqrt{25.80} = \sqrt{5.07} = 2.25$$

b) Menghitung jarak dataset 2

$$D = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

$$D2 = \sqrt{(25.4 - 15.9)^2 + (72 - 28.3)^2}$$

$$D2 = \sqrt{90.25 + 1.909.69}$$

$$D2 = \sqrt{92.159} = \sqrt{9.59} = 2.36$$

c) Menghitung jarak dataset 3

$$D = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

$$D3 = \sqrt{(21.5 - 9)^2 + (71.2 - 29.6)^2}$$

$$D3 = \sqrt{15.625 + 1.730.56}$$

$$D3 = \sqrt{157.98} = \sqrt{12.56} = 3.54$$

d) Menghitung jarak dataset 4

$D=\sqrt{(x2-x1)^2+(y2-y1)^2}$

$D4=\sqrt{(16-14.7)^2+(87.5-31.4)^2}$

$D4=\sqrt{1.69+3.147.21}$

$D4=\sqrt{4.837}$

D4= 2.19

e) Menghitung jarak dataset 5

$D=\sqrt{(x2-x1)^2+(y2-y1)^2}$

$D5=\sqrt{(22.3-11.6)^2+(74.5-30.9)^2}$

$D5=\sqrt{114.49+1.900.96}$

$D5=\sqrt{116,39}$

$D5=\sqrt{10.78}$ = 3.28

f) Menghitung jarak dataset 6

$D=\sqrt{(x2-x1)^2+(y2-y1)^2}$

$D6=\sqrt{(29.7-14.1)^2+(30.5-30.5)^2}$

$D6=\sqrt{2.38.68+0}$

$D6=\sqrt{238.68}$

$D6=\sqrt{15,44}=3,92$

g) Menghitung jarak dataset 7

$$D=\sqrt{(x2-x1)^2+(y2-y1)^2}$$

$$D7=\sqrt{(19.5-12.7)^2+(82.9-28.9)^2}$$

$$D7=\sqrt{46.24+2.916}$$

$$D7=\sqrt{49.156}$$

$$D7=\sqrt{7.01}=2.64$$

4. Perhitungan jarak dari yang terkecil

   Dataset 4 = 2.19

   Dataset 1 = 2.25

   Dataset 2 = 2.36

   Dataset 7 = 2.64

   Dataset 5 = 3.28

   Dataset 3 = 3.54

   Dataset 6 = 3.92

5. Jadi, untuk nilai k = 5. Lima tetetangga terdekat adalah dataset 4,Dataset 1, Dataset 2, Dataset 7, dan dataset 5.

6. Menghitung nilai mayoritas lanjut rata-rata

Dalam menghitung nilai rata-rata dari mayoritas yang sudah ditentukan terlebih dahulu dimana hanya akan menggunakan data dari nilai-nilai yang terkait dengan kelas “Tidak Anemia (1) seperti pada tabel 2.3

Tabel 2. 3 Mayoritas Data

KELAS	NILAI	KESIMPULAN
Dataset 4	2.19	Tidak Anemia(1)
Dataset 1	2.25	Tidak Anemia(1)
Dataset 2	2.36	Tidak Anemia(1)
Dataset 7	2.64	Tidak Anemia(1)

Rumus Rata-rata

$$\bar{X} = \frac{\sum_{i=1}^{k} x_{Ai}}{k} \qquad \text{.....(3)}$$

Keterangan :

$\bar{X}$ adalah nilai rata-rata

x_{Ai} adalah nilai individual yang termasuk dalam kelas

K adalah jumlah tertangga terdekat

Jadi, dari data diatas data dijabarkan sebagai berikut :

Rata – Rata = $\frac{2.19 + 2.25 + 2.36 + 2.64}{5}$

Rata -rata ≈ $\frac{9.44}{5}$

Rata – Rata ≈ 1.8

Jadi nilai rata-rata dengan data mayoritas dan tetangga terdekat 5 adalah 1.8.

2.2.6 Confusion Matrix

Confusion matrix digunakan untuk melakukan evaluasi kinerja pada metode klasifikasi dengan menganalisis tingkat akurasi dari *classifier* dalam mengenali *tuple* dari kelas yang berbeda. Ada beberapa istilah yang digunakan dalam *confusion matrix* yaitu TP (*True Possitive*) dan TN (*True Negative*) memberikan informasi jika *classifier* benar sedangkan FP (*False Possitive*) dan FN (*False Negative*) memberikan informasi ketika *classifier* salah. (Han, Kamber, & Pei, 2012).

Confusion matrix merupakan salah satu metode yang digunakan untuk mengukur performa dari suatu model klasifikasi yang telah dibuat, dimana *output* dapat berupa dua kelas atau banyak kelas. Pengukuran kinerja *confusion matrix* menggunakan tabel yang dapat dilihat pada gambar 2.4

Tabel 2. 4 Confusion Matrix

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	TP (True Positive)	FN (False Negative)
Negatif	FP (False Positive)	TN (True Negative)

Keterangan :

a. TP yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.

b. TN yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.

c. FN yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.

d. FP yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

$$\text{Akurasi} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100\% \quad \text{……...…..(4)}$$

$$\text{Presisi} = \frac{TP}{FP+TP} * 100\% \quad \text{…….(5)}$$

$$\text{Recall} = \frac{TP}{FN+TP} * 100\% \quad \text{…….(6)}$$

Keterangan :

a. ***Precision*** merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif..

b. ***Recall*** adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

c. ***Accuracy*** merupakan rasio prediksi Benar (positif dan negatif) dengan keseluruhan data.

2.2.7 Python

Bahasa pemrograman *python* merupakan salah satu bahasa pemrograman populer dan banyak digunakan dalam berbagai bidang, termasuk analisis data. *Python* menawarkan kemudahan pengguna dan fleksibilitas dalam penelitian atau pengembangan. Pada analisis data, *python* berproses sebagai *mining* data dalam berbagai hal berdasarkan suatu metode yang diimplementasikan pada *dataset* sehingga mampu menghasilkan suatu informasi, semakin kompleks metode yang digunakan semakin mendalam wawasan yang diperoleh dari data, namun dibutuhkan juga sumber daya manusia dan teknologi yang canggih. Dengan demikian, *Python* memberikan alat dan teknik yang memudahkan proses data

mining secara lebih efektif dan efisien. (Lo, et al., 2023).

2.2.8 Flask

Flask adalah sebuah *framework* (kerangka kerja) *web* yang dituliskan dalam bahasa pemrograman *python*. Umumnya *framework* ini dirancang untuk membangun aplikasi *web* dengan cara yang sederhana agar *developer* mudah memahaminya, juga menawarkan seperangkat alat dan juga fitur yang dapat membantu pengembang dalam mengembangkan aplikasi dan tidak memiliki ketergantungan dengan alat atau pustaka tertentu.

Flask Python adalah *framework* yang mampu memudahkan *developer* dalam mengimplementasikan fitur-fitur pemrograman. Contohnya seperti *routingURL*, pengolahan formulir, pengelolaan sesi pengguna, dan integrasi basis data (Flask Python, 2024).