BAB II
LANDASAN TEORI

2.1 Penelitian Terkait

Beberapa penelitian terdahulu terkait *text mining* ini telah dilakukan dari tahun ke tahun dengan beberapa metode dan algoritma klasifikasi sebagai berikut:

1. Hendro Priyatman, Fahmi Sajid, dan Dannis Haldivany pada tahun 2019 dengan judul "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa". Parameter yang digunakan untuk membangun aplikasi ini adalah IPK dan kehadiran. (Priyatman, Sajid, dan Haldivany 2019)
2. Puspita Sari pada tahun 2017 dengan judul "*Improve K-Means* terhadap status gizi pada balita" dalam penelitian ini penulis mengubah sistem manual yang masih digunakan puskesmas Abeli dalam pemberkasan dan pengelompokkan status gizi balita(Sari, Pramono, dan Sagala 2017).
3. Yessica Putri Santoso, Marlina, dan Halim Agung pada tahun 2018 dengan judul " Implementasi Metode *K-Means Clustering* pada Sistem Rekomendasi Dosen Tetap Berdasarkan Penilaian Dosen" pada penelitian ini penulis mengimplementasikan metode clustering dengan algoritma K-means dalam pengelompokan data dosen. (Santoso, Marlina, dan Agung 2018)
4. Rozzi Kesuma Dinata, Safwandi, Novia Hasdyna, Nur Azizah pada tahun 2022 dengan judul " Analisis K-Means Clustering pada Data Sepeda Motor". Pada penelitian ini penulis menerapkan *K-Means* dalam pengelompokan sepeda motor dan menghasilkan 3 cluster yaitu murah, standard, dan mahal. (Dinata, dkk 2020)

Tabel 2. 1 Penelitian Terdahulu

No	Peneliti	Judul	Tujuan	Hasil Penelitian	Keterbatasan
1	Hendro Priyatman, Fahmi Sajid, dan Dannis Haldivany	Klasterisasi Menggunakan Algoritma *K-Means Clustering* untuk Memprediksi Waktu Kelulusan Mahasiswa	Untuk memprediksi waktu kelulusan mahasiswa serta memberikan sarana untuk mengetahui perkiraan waktu lulus	Implementasi algoritma *K-Means* dalam data mining sudah berhasil, dan bisa menampilkan informasi prediksi kelulusan mahasiswa, namun untuk melihat kemampuan *real K-Means Clustering* dalam memprediksi waktu kelulusan tergantung pada mahasiswa itu sendiri.	Parameter yang terbatas, dan tidak menggunakan algoritma lain untuk memperoleh keputusan yang beragam.
2	Puspita Sari	Improve K-Means terhadap Status Nilai Gizi pada Balita	Untuk mengubah sistem manual yang masih digunakan puskesmas Abeli dalam pemberkasan dan pengelompokan status gizi balita	Aplikasi penentuan status gizi pada balita dengan nilai akurasi 95%, sensitivity 93,5%, dan specificity 100% pada puskesmas Abeli.	Dibutuhkan pemeliharaan dan pengawasan untuk sistem
3	Yessica Putri Santoso, Marlina, dan Halim Agung	Implementasi Metode *K-Means Clustering* pada Sistem Rekomendasi Dosen Tetap Berdasarkan Penilaian Dosen	Untuk membantu Kaprodi dalam menentukan dosen tetap berdasarkan hasil penilaian yang didapatkan dosen melalui sistem	Hasil rekomendasi dosen tetap dengan kategori layak pada proses *clustering* sebesar 55% tingkat keberhasilan, dengan jumlah 39 data dosen tetap yang layak dari 70 data dosen yang di *cluster* sehingga algoritma *K-Means Clustering*	*K-Means Clustering* tidak cocok digunakan pada kasus ini.

No	Peneliti	Judul	Tujuan	Hasil Penelitian	Keterbatasan
				tidak cocok digunakan pada studi kasus ini.	
4.	Rozzi Kesuma Dinata, Safwandi, Novia Hasdyna, Nur Azizah	Analisis *K-Means Clustering* pada Data Sepeda Motor	Untuk merekomendasikan bagi penguna dalam menentukan pemilihan sepeda motor yang diinginkan.	Hasil analisis performasi *K-Means* dari 15 pengujian dari setiap uji coba yang dilakukan, diperoleh rata-rata persicion sebesar 76%, nilai recall sebesar 76%, dan Accurasi sebesar 81%.	Penerapan *K-Means* untuk *clustering* masih dapat dikembangkan dengan data yang lain.

2.2 Status Gizi

Menurut Suharjo (1983), status gizi adalah keadaan tubuh sebagai akibat dari pemakaian, penyerapan, dan penggunaan makanan. Makanan yang memenuhi gizi tubuh, umumnya membawa ke status gizi memuaskan. Jika kekurangan atau kelebihan zat gizi esensial dalam makanan untuk jangka waktu yang lama disebut gizi salah. Manifestasi gizi salah dapat berupa gizi kurang dan gizi lebih.

Balita adalah anak usia dibawah lima tahun yang berumur 0-4 tahun 11 bulan. Penilaian Status Gizi (PSG) adalah sebuah metode mendeskripsikan kondisi tubuh sebagai akibat keseimbangan makanan yang dikonsumsi dengan penggunaannya oleh tubuh, yang biasanya dibandingkan dengan suatu nilai normatif yang ditetapkan (Kementerian Kesehatan RI 2011).

Status gizi anak ditentukan dengan beberapa kriteria, berdasarkan berat badan dan panjang/tinggi badan meliputi Berat Badan menurut Umur (BB/U), Panjang/Tinggi Badan menurut Umur (PB/U atau TB/U) dan Berat Badan menurut Panjang/Tinggi Badan (BB/PB atau BB/TB) (PMK No.2 Tahun 2020).

2.3 Data Mining

Menurut Fayyad dalam buku (Kusrini, 2009) Istilah *data mining* dan *knowledge discovery in database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi

berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining (Putri, Izman, and Dian 2014) .

Pengelompokkan data mining ada 5 yaitu:

a. Estimasi

   Estimasi hampir sama dengan klasifikasi, perbedaannya variable target estimasi lebih kearah numeric dari pada kearah kategori.

b. Forcasting

   Prediksi hampir sama dengan klasifikasi dan estimasi, perbedannya dalam prediksi nilai dari hasil aka nada dimasa mendatang.

c. Klasifikasi

   Dalam klasifikasi, terdapat target variable kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori yaitu: pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

d. Clustering

   Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan tidak memiliki kemiripan dengan record- record dalam kluster lain.

e. Asosiasi

   Tugas asosiasi dalam data mining adalah menentukan atribut yang muncul dalam satu waktu. Biasanya asosiasi diterapkan di supermarket dengan tatacara penyusunan barang.

2.3.1 Clustering

Pengelompokkan data-data ke dalam sejumlah kelompok (*cluster*) berdasarkan kesamaan karakteristik masing-masing data pada kelompok-kelompok yang ada. Pengelompokkan data dibedakan menurut struktur kelompok, keanggotaan data dalam kelompok, dan kekompakan data dalam kelompok. Menurut struktur, pengelompokkan dibagi dua, yaitu *hierarki* dan *partitioning*. Dalam *hierarki*, satu data tunggal bisa dianggap sebuah kelompok, dua atau lebih. Pengelompokkan *partitioning* membagi setiap data hanya menjadi anggota satu kelompok (Asroni, Fitri, and Prasetyo 2018).

2.3.2 K-Means

K-means merupakan salah satu metode pengelompokkan data non-hierarki

yang mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain. Adapun tujuan pengelompokkan data ini adalah untuk meminimalkan fungsi objektif yang di set dalam suatu kelompok dan memaksimalkan variasi antar kelompok

Metode K-Means berusaha mengelompokkan data yang ada ke dalam beberapa kelompok, dimana data dalam suatu kelompok mempunyai karakteristik yang berbeda dengan data yang ada di dalam kelompok yang lain. Dasar algoritma k-means adalah sebagai berikut:

1. Tentukan nilai K sebagai cluster yang ingin dibentuk
2. Inisialisasi K sebagai centeroid yang dapat dibangkitkan sevara random.
3. Hitung jarak setiap data ke masing-masing centeroid menggunakan persamaan Euclidean Distance yaitu sebagai berikut:

   $$\mathrm{d}(P,Q) = \sqrt{\sum_{j=1}^{p}(x_j(P) - x_j(Q))^2} \qquad (2.1)$$

   Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centeroidnya.

4. Tentukan posisi centeroid baru (*k*)
5. Kembali ke langkah 3 jika posisi centeroid baru dengan centeroid lama tidak sama.

2.4 Z-Score

Angka baku atau Z-Score adalah bilangan yang menunjukkan tingkat penyimpangan data dari rata-rata dalam satuan standar deviasi atau seberapa jauh suatu nilai tersebut menyimpang dari rata-rata dengan satuan simpang baku (*s*)(Kementerian Kesehatan RI 2011). Angka baku digunakan untuk mencari normalitas data. Rumus angka baku adalah sebagai berikut:

$$\mathbf{Z = (xi-\mu) / s} \qquad (2.2)$$

Dengan:

Z = Nilai Z-Score

xi = Nilai data ke i

i = 1,2,3,…,n

μ = nilai rata-rata

s = nilai standart deviasi

Tabel 2. 2 Kategori Ambang Batas Status Gizi Anak Berdasarkan Indeks

Indeks	Kategori Status Gizi	Ambang Batas (*Z-Score*)
Berat Badan Menurut Umur (BB/U)	Gizi Sangat Kurang	<-3 SD
	Gizi Kurang	-3 SD sd <-2 SD
	Gizi Normal	-2 SD sd +1 SD
	Gizi Lebih	>+1 SD
Panjang Badan Menurut Umur (PB/U) atau Tinggi Badan Menurut Umur (TB/U)	Sangat Pendek	<-3 SD
	Pendek	-3 SD sd <-2 SD
	Normal	-2 SD sd +3 SD
	Tinggi	>+3 SD
Berat Badan Menurut Tinggi Badan (BB/TB) atau Berat Badan Menurut Panjang Badan (BB/PB)	Gizi Buruk	<-3 SD
	Gizi Kurang	-3 SD sd <-2 SD
	Gizi Baik	-2 SD sd +1 SD
	Berisiko Gizi Lebih	>+1 SD sd +2 SD
	Gizi Lebih	>+2 SD sd +3 SD
	Obesitas	>+3 SD
Indeks Massa Tubuh menurut Umur (IMT/U)	Gizi Buruk	<-3 SD
	Gizi Kurang	-3 SD sd <-2 SD
	Gizi Baik	-2 SD sd +1 SD
	Beresiko Gizi Lebih	>+1 SD sd +2 SD
	Gizi Lebih	>+2 SD sd +3 SD
	Obesitas	>+3 SD
Indeks Massa Tubuh Menurut Umur	Gizi Buruk	<-3 SD
	Gizi Kurang	-3 SD sd <-2 SD
	Gizi Baik	-2 SD sd +1 SD
	Gizi Lebih	+1 SD sd +2 SD
	Obesitas	>+2 SD

(Sumber : PMK No.2 Tahun 2020)

2.4.1 Nilai Rata-Rata

Rata-rata merupakan nilai yang dapat mewakili sekelompok data. Agar rata-rata dapat mewakili sekelompok data dengan baik, syarat yang harus dimiliki data adalah data tersebut tidak boleh memiliki nilai exstem baik di ujung data maupun di awal data. Nilai ekstrem adalah nilai yang terlalu kecil atau nilai yang terlalu

besar, karena jika nilai ini dimiliki oleh data, maka akan mempengaruhi rata-rata sehingga rata-rata tidak menggambarkan keberadaan data keseluruhan. Dalam penelitian ini rata-rata yang digunakan adalah rata-rata aritmatika. (Kesehatan, 2017)

$$\mu = \frac{x_{1+}x_{2+\ldots.+}x_n}{n} \quad (2.\ 3)$$

Dengan:

μ = nilai rata-rata

x = nilai data ke-i

n = jumlah data yang diamati

2.4.2 Standar Deviasi

Varian dan standart deviasi (simpangan baku) adalah ukuran-ukuran keragaman (variasi) data statistic yang paling sering digunakan. Standar deviasi (simpangan baku) merupakan akar kuadrat dari varian $s = \sqrt{s^2}$. Salah satu cara untuk mengetahui keragaman dari suatu kelompok data adalah dengan mengurangi setiap nilai data dengan rata-rata kelompok data tersebut, selanjutnya semua hasilnya dijumlahkan. Namun cara seperti itu tidak bisa digunakan karena hasilnya akan selalu menjadi 0.

Oleh karena itu, solusi agar nilainya tidak menjadi 0 adalah dengan mengkuadratkan setiap pengurangan nilai data dan rata-rata kelompok data tersebut, selanjutnya dilakukan penjumlahan. Hasil penjumlahan kuadrat (sum of squares) tersebut akan selalu bernilai positif. Dalam penerapannya, nilai varian tersebut bisa untuk menduga varian populasi. Nilai varian populasi lebih besar dari varian sampel. Oleh karena itu, agar tidak bisa dalam menduga varian populasi, maka n sebagai pembagi penjumlahan kuadrat *(sum of squares)* diganti dengan n-1 (derajat bebas) agar nilai varian sampel mendekati varian populasi. Untuk menyeragamkan nilai satuannya, maka varian diakar kuadratkan sehingga hasilnya adalah standart deviasi (simpangan baku).

$$s^2 = \sum(xi - x)^2 / n - 1 \quad (2.\ 4)$$

Dengan:

s = standar deviasi

xi = nilai data ke-i

i = 1,2,3…,n

n = jumlah data yang diamati

2.5 Perancangan Sistem

Perancangan sistem adalah suatu fase di mana diperlukan suatu keahlian perancangan untuk elemen-elemen komputer yang akan mengunakan sistem yaitu pemilihan peralatan dan program komputer untuk sistem yang baru.(Suryadi 2015).

2.5.1 Hypertext Prepocessor (PHP)

Menurut Supono & Putratama (2018: 1) mengemukakan bahwa PHP (hypertext preprocessor) adalah suatu bahasa pemrograman yang digunakan untuk menterjemahkan basis kode program menjadi kode mesin yang dapat dimengerti oleh komputer yang bersifat server-side yang ditambahkan ke HTML.

2.5.2 MySQL

Menurut Risnandar (2013:92) mendefinisikan bahwa"MySQL merupakan basis data yang bersifat open source sehingga banyak digunakan di dunia.

2.5.3 Data Flow Diagram

Menurut Kristanto (2008 : 61), "*Data Flow Diagram* (DFD) adalah suatu model logika data atau proses yang dibuat untuk menggambarkan darimana asal data dan kemana tuuan data yang keluar dari sistem, di mana data disimpan, proses apa yang menghasilkan data tersebut dan interaksi antara data yang tersimpan dan proses yang dikenakan pada data tersebut".

Adapun simbol-simbol *Data Flow Diagram* adalah sebagai berikut:

Tabel 2. 3 Simbol - Simbol DFD

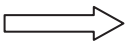
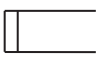
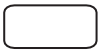
Nama	Simbol	Keterangan
Proses		Proses atau fungsi atau prosedur; pada pemodelan perangkat lunak yang akan diimplementasikan dengan pemograman terstruktur, maka pemodelan notasi inilah yang harusnya di dalam kode program
Berkas		File atau basisdata atau penyimpanan; pada pemodelan

Nama	Simbol	Keterangan
		perangkat lunak yang akan diimplementasikan dengan pemograman terstruktur, maka pemodelan notasi inilah yang harusnya dibuat menjadi tabel-tabel basis data yang dibutuhkan.
Entitas Luar		Entitas luar (external entity) orang yang berinteraksi dengan perangkat lunak yang dimodelkan atau sistem lain yang terkait dengan aliran data dari sistem yang dimodelkan
Aliran data		Aliran data merupakan data yang dikirim antar proses, dari penyimpanan ke proses, atau dari proses ke masukkan.

2.5.4 Flowchart

Flowchart merupakan penggambaran secara grafik dari langkah-langkah dan urutan prosedur suatu program,. Biasanya mempengaruhi penyelesaian masalah yang khusunya perlu dipelajari dan dievaluasi lebih lanjut. (Budiman et al. 2021)

Tabel 2. 4 Simbol - Simbol Flowchart

Simbol	Nama	Deskripsi
	Sistem	Menunjukan sistem
	Eksternal *entity*	Menunjukan bagian luar sistem atau sumber *input* dan *output* data
	Garis aliran	Menunjukan arus data antar simbol/proses
	Garis aliran	Aliran material
	Data Storage	Digunakan untuk menyimpan arus data atau arsip seperti file transaksi, file induk atau file referensi dan lain – lain.
	Proses	Suatu proses yang dipicu atau didukung oleh data.

	Conector (On-page connector)	Digunakan untuk penghubung dalam satu halaman
	Conector (Off-page connector)	Digunakan untuk penghubung berbeda halaman

2.5.5 Entity Relationship Diagram (ERD)

Menurut (Sari, Pramono, and Sagala 2017), "*Entity Relationship Diagram (ERD)* adalah suatu rancangan atau bentuk hubungan suatu kegiatan di dalam sistem yang berkaitan langsung dan mempunyai fungsi di dalam proses tersebut". Adapun simbol-simbol yang digunakan dalam *Entity Relationship Diagram (ERD)* dengan notasi Chen, yaitu:

Tabel 2. 5 Simbol – Simbol ERD

Simbol	Nama	Keterangan
	Entity	Entitas merupakan data inti yang akan tersimpan; bakal tabel pada basis data; benda yang memiliki data dan harus disimpan datanya agar dapat diakses oleh sistem komputer; penamaan entitas biasanya lebih kekata benda dan belum merupakan nama tabel
	Atribut	*Field* atau kolom data yang butuh disimpan dalam suatu entitas
	Multivalue	*Field* atau kolom data yang butuh disimpan dalam suatu entitas yang dapat memiliki nilai lebih dari satu.
	Relasi	Relasi yang menghubungkan antar entitas; biasanya diawali dengan kata kerja.
	Association	Penghubung antara relasi dan entitas di mana dikedua ujungnya punya *multiplicity* ke-mungkinan jumlah pemakaian. Kemungkinan

		jumlah maksimum keterhubungan antara entitas yang lain disebut dengan *one to many* menghubungkan entitas A dan entitas B.