BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Beberapa Penelitian yang terkait dengan penerapan metode Klasifikasi menggunakan Algortima C.45 diantaranya :

No	Judul, Penulis, Tahun	Dataset	Metode	Hasil	Kekurangan	Kelebihan
1	Evaluasi model decision tree c4.5 guna prediksi Posibilitas resiko obesitas Mochammad Yusa 1 , Wahyu Sindu 2015	Data yang digunakan adalah data simulasi	Decision Tree C.45	Dari hasil pengujian yang dilakukan, Model ini mempunyai nilai akurasi yang berada pada angka 80%.	-	Metode Model Klasifikasi Decision Tree menggunakan Algoritma C4.5 mempunyai performa yang baik dalam memberikan solusi prediktif terhadap kemungkinan

						terjadinya obesitas.
2	Penerapan Naïve Bayes Classification untuk Klasifikasi Tingkat Kemungkinan Obesitas Mahasiswa Sistem Informasi UIN Suska Riau Wiwik Muslehatin1 , Muhammad Ibnu2 ,	pada penelitian ini data diperoleh dari survei melalui kuisioner dan mengambil sampel secara acak	metode algoritma Naïve Bayes Classification	hasil pengujian menunjukkan dengan akurasi sebesar 66,67%	pada saat melakukan proses training dan testing klasifikasi tidak optimal jika tidak disimpan pada memori (database), untuk itu perlu disimpan hasil klasifikasi pada memori (database) untuk mengoptimalkan	kelebihan dengan kinerja yang baik

	Mustakim 2017				klasifikasi.	
3	Klasifikasi status gizi orang dewasa menggunakan algoritma naïve bayes (studi kasus klinik bhakti mulia cikarang) Wahyu Hadikristanto 1) , Tiara Deswara Pungkas 2) 2019	dataset yang diambil dari beberapa data pasien. Data yang diambil berupa dataset status gizi orang dewasa yang akan diakurasi kebenarannya	metode algoritma Naïve Bayes Classification	Uji coba ini dilakukan dengan 150 data, dan hasil akurasi yang didapat sebesar 88,67%.	status gizi orang dewasa menghasilkan nilai akurasi dengan mengguna kan algoritma Naïve Bayes dimana pada tahapan tersebut perlu fokus dalam menyiapkan sebuah data agar bersih dari segala noise data yang ada sehingga proses	algoritma Naïve Bayes akurat dalam perhitungan status gizi orang dewasa.

					klasifikasi dapat berjalan dengan baik dan akurat.	
4	Obesitas dan Obesitas Sentral pada Masyarakat Usia Dewasa di Daerah Perkotaan Indonesia Septiyanti1*, Seniwati2 2020	Desain penelitian ini adalah studi cross sectional	n data hasil riset kesehatan dasar (Riskesdas) tahun 2007, rumah tangga yang terpilih sebagai sampel khususnya kelompok biomedis di seluruh Indonesia	Subjek usia 40-59 tahun (dewasa akhir) lebih banyak yang mengalami obesitas, dengan proporsi sebesar 51,0%. Sementara berdasarkan jenis kelamin, subjek perempuan lebih banyak mengalami obesitas, dengan proporsi sebesar 66,9% dibandingka	Metode yang digunakan seharusnya bisa mengggunakan metode lain sebagai cara perhitungannya yang lebih optimal	Penelitian ini membuktikan bahwa ada perbedaan bermakna pada pemeriksaan biomedis pada mereka yang obesitas dengan tidak obesitas.

				n dengan subjek laki-laki (33,1%). Berdasarkan tingkat pendidikan, subjek tamat SLTA lebih banyak mengalami obesitas, disusul dengan subjek tamat SD, dengan proporsi masing-masing 32,6% dan 24,1%. Adapun berdasarkan pekerjaan, subjek yang bekerja sebagai ibu rumah tangga dan wiraswasta/pedagang memiliki proporsi		

				penderita obesitas yang paling besar, yaitu masing-masing 37,4% dan 21,8%. Total penderita obesitas adalah sebesar 3738 orang (28,5%).		
5.	Algoritma K-Means Clustering untuk Menentukan Nilai Gizi Balita Eni Irfiani#1, Siti Sulistia Rani*2 2018	Data yang digunakan adalah data balita dari Posyandu, Kartu Menuju Sehat (KMS), Tinggi Balita (TB) dan Berat Badan (BB) dengan rentan usia	Algoritma K-Means Clustering	Dari hasil tersebut diketahui masih terdapat 30% balita obesitas serta 11% balita kekurangan gizi	perlu adanya pendampingan dari Posyandu serta puskesmas terkait kepada orang tua balita sehingga jumlah balita yang kekurangan gizi	pembagian menjadi 5 cluster yaitu obesitas, gizi lebih, gizi baik, gizi kurang dan gizi buruk guna membantu kinerja para kader Posyandu dan Orang Tua balita dalam

		balita 0 sampai 36 bulan			dapat menurun di tahun berikutnya	penanganan dini kondisi nilai gizi balita
6	Implementasi K-Nearest Neighbor (KNN) Algoritma Untuk Klasifikasi Tingkat Obesitas Ayu Made Surya Indra Dewia1, Ida Bagus Gede Dwidasmaraa2 2020	Data yang digunakan dalam penelitian ini adalah data sekunder yaitu data tingkat obesitas berdasarkan pola makan dan kondisi fisik diperoleh dari www.kaggle.com.	Algoritma K-Nearest Neighbor (KNN)	Dari hasil yang telah dicapai, maka dapat disimpulkan bahwa algoritma KNN dapat mengklasifikasikan tingkat obesitas berdasarkan kebiasaan makan dan kondisi fisik cukup baik. Hal ini dibuktikan dengan pencapaian akurasi 78,98% dengan	Hasil akurasi yang didapatkan seharusnya bisa lebih besar menggunakan metode lain	Sudah dapat menentukan level obesitas meskipun tingkat akurasi yang tidak terlalu akurat

				parameter k = 2 pada simulasi menggunakan Rapid Buruh tambang.		

Dari hasil review beberapa jurnal diatas baik jurnal nasional dan internasional maka dapat disimpulkan untuk mendapatkan akurasi yang baik maka data yang dikumpulkan harus sangat banyak dan tingkat kesalahan data sedikit, kemudian dari setiap algoritma akan menghasilkan akurasi yang berbeda tergantung pada kualitas data tersebut. hasil perbandingan akurasi dapat dilihat dibawah ini:

Tabel 2.2 Hasil Akurasi Review Jurnal

No	Jurnal	Algoritma	Akurasi
1	Mochammad Yusa	DECISION TREE C4.5	80%
2	Wiwik Muslehatin	Naïve Bayes	66,67%
3	Wahyu Hadikristanto	NAÏVE BAYES	88,67%
4	Ayu Made Surya Indra Dewi	K-Nearest Neighbor (KNN)	78,98%

2.2 Landasan Teori

2.2.1 Obesitas

Kegemukan atau obesitas adalah suatu kondisi medis berupa kelebihan lemak tubuh yang terakumulasi sedemikian rupa sehingga menimbulkan dampak

merugikan bagi kesehatan, yang kemudian menurunkan harapan hidup dan atau meningkatkan masalah kesehatan. depresi merupakan salah satu gangguan mood yang memiliki dampak berupa pandangan tentang masa depan yang suram serta pesimistis, gagasan atau perbuatan membahayakan diri atau bunuh diri, tidur terganggu, dan nafsu makan terganggu, baik depresi maupun obesitas adalah masalah kesehatan masyarakat yang saling berhubungan. Depresi dan obesitas sama-sama terkait dengan stigma sosial, perasaan harga diri rendah, dan kondisi kesehatan kronis. Ketika depresi dan obesitas terjadi bersamaan, konsekuensi kesehatan dan sosial yang merugikan menjadi signifikan.(Murtane, 2021)

Indonesia sedang menjalani transisi nutrisi karena sepertiga orang dewasa sekarang kelebihan berat badan atau obesitas. Teori transisi nutrisi menunjukkan bahwa perkembangan ekonomi, urbanisasi dan globalisasi menghasilkan peningkatan konsumsi makanan olahan ultra dan penurunan aktivitas fisik, yang kemudian mengarah pada prevalensi penyakit kelebihan berat badan dan penyakit tidak menular yang lebih tinggi (Oddo et al., 2019)

Obesitas dan obesitas sentral telah menjadi masalah kesehatan masyarakat yang serius di negara berkembang seperti Indonesia. Meskipun 10 tahun telah berlalu sejak survei kesehatan nasional terbesar dilakukan pada tahun 2007, namun tidak ada analisis dan publikasi lebih lanjut mengenai obesitas dan obesitas sentral di Indonesia berdasarkan survei tersebut (Harbuwono et al., 2018)

2.2.2 Data Mining

Data mining adalah suatu metode pengolahan data yang memungkinkan untuk menggunakan koneksi data sebagai dasar pengambilan keputusan dengan mencari koneksi dari data yang tidak diketahui pengguna dan menyajikannya dalam format yang mudah dipahami. (Ridwan et al., 2013). Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu :

Deskripsi, Estimasi, Prediksi, Klasifikasi, Clustering, dan Asosiasi. (Muslehatin et al., 2017).

Definisi umum dari data mining itu sendiri adalah menggambar pola tersembunyi (hidden pattern) berupa pengetahuan yang sebelumnya tidak diketahui dari kumpulan data yang mungkin ada di database, gudang data, atau media penyimpanan informasi lainnya.

Penambangan data dilakukan dengan menggunakan alat khusus yang melakukan operasi penambangan data yang ditentukan berdasarkan model analitik. Data mining adalah proses analisis data yang berfokus pada pencarian informasi tersembunyi dari sejumlah besar data yang disimpan selama operasi perusahaan. Kemajuan luar biasa yang terus berlanjut dalam bidang data mining didorong oleh beberapa faktor antara lain: 1). Pertumbuhan yang cepat dalam kumpulan data. 2). Penyimpanan data dalam data warehouse, sehingga seluruh perusahaan memiliki akses ke dalam database yang andal. 3). Adanya peningkataan akses data melalui navigasi web dan internet. 4). Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi. 5). Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi. 6). Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan (Rahmawati & Merlina, 2018)

Secara umum, metode data mining dapat dibagi menjadi dua :

deskriptif dan prediktif. Deskriptif berarti data mining digunakan untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data.

Sedangkan prediktif berarti data mining digunakanuntuk membentuk sebuah model pengetahuan yang akan digunakan untuk melakukan prediksi (Suyanto, 2017)

Metode yang ada dalam data mining adalah sebagai berikut :

1. *Classification*

   Klasifikasi adalah proses menemukan sekumpulan model yang dijelaskan oleh kelas data, dan model tersebut dapat digunakan untuk memprediksi nilai untuk kelas objek yang tidak diketahui. Untuk mendapatkan model tersebut, kita perlu melakukan analisis terhadap data latih. Data uji digunakan untuk mengetahui tingkat akurasi dan model yang dihasilkan. Klasifikasi dapat digunakan untuk memprediksi nama atau nilai suatu objek data.

2. *Clustering*

   Data yang penunjukan kelasnya tidak diketahui dikelompokkan ke dalam sejumlah kelompok yang ditentukan menurut derajat kemiripannya. Metode ini akan digunakan dalam tugas akhir ini.

3. *Association*

   Tujuan dari metode ini adalah untuk menghasilkan satu set aturan yang menggambarkan satu set data yang sangat terkait..

4. *Regression*

   Regresi mirip dengan klasifikasi. Perbedaan utama terletak pada atribut yang menghasilkan nilai kontinu.

5. *Forecasting*

Prediksi (forecasting) berfungsi untuk melakukan prediksi kejadian yang akan diproses berdasarkan data sejarah yang ada.

6. *Sequence Analiysis*

Tujuan dari metode ini adalah untuk mengenali pola dari data diskrit sebagai contoh adalah menemukan kelompok gen dengan tingkat ekspresi yang mirip.

7. *Deviation Analiysis*

Tujuan dari metode ini adalah untuk menemukan penyebab perbedaan antara data yang satu dengan data yang lain dan biasa disebut sebagai outlier detection. Sebagai contoh adalah apakah sudah terjadi penipuan terhadap pengguna kartu kredit dengan melihat catatan transaksi yang tersimpan dalam basis data perusahaan tersebut.

2.2.3 Klasifikasi

Klasifikasi data adalah proses menemukan properti yang sama dalam satu set objek dalam database dan mengklasifikasikannya ke dalam kelas yang berbeda sesuai dengan model klasifikasi yang ditentukan. Tujuan klasifikasi adalah untuk menemukan model dalam data latih dan mengklasifikasikan atribut ke dalam kategori atau kelas yang sesuai dengan model tersebut.(Ente et al., 2020).

Untuk menggunakan metode klasifikasi tentunya harus menerapkan Algortima dalam Impelementasinya. Algortima yang akan digunakan adalah Decision Tree. Algortima C4.5 adalah ekstensi Quinlan untuk algoritma ID3 untuk menghasilkan

pohon keputusan (Decision Tree), algoritma C4.5 rekursif mengunjungi setiap node keputusan, memilih split optimal sampai tidak ada perpecahan lanjut yang memungkinkan (Larose, 2005 dalam Novandya, 2017).

Klasifikasi adalah teknik penambangan data prediktif yang menggunakan hasil yang diketahui dari kumpulan data yang berbeda untuk membuat prediksi tentang data nilai. Masalah dengan akurasi banyak algoritma klasifikasi adalah bahwa informasi diketahui hilang ketika berhadapan dengan data yang tidak seimbang, seperti ketika distribusi sampel antar kelas sangat miring. (Misdram, 2021). Dalam taksonomi, Anda memiliki variabel target kategoris, seperti strata pendapatan, yang dapat, misalnya, membagi Anda menjadi tiga kelas atau kategori: pendapatan tinggi, pendapatan menengah, dan pendapatan rendah. Model data mining memeriksa satu set besar catatan, masing-masing catatan yang berisi informasi tentang variabel target serta satu set input atau prediktor variable. Contoh tugas klasifikasi dalam bisnisdan penelitian meliputi: (Larose & Larose, 2014).

a. Menentukan apakah transaksi kartu kredit tertentu adalah penipuan
b. Menempatkan siswa baru pada jalur tertentu yang berkaitan dengan kebutuhan khusus
c. Menilai apakah aplikasi hipotek adalah risiko kredit yang baik atau buruk
d. Mendiagnosis apakah ada penyakit tertentu
e. Menentukan apakah surat wasiat ditulis oleh almarhum yang sebenarnya, atau curangoleh orang lain
f. Mengidentifikasi apakah perilaku keuangan atau pribadi tertentu menunjukkan kemungkinan ancaman teroris

Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya Naïve Bayes, SupportVector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan (Wibawa, 2018).

2.2.4 Decision Tree C.45

Pada dasarnya konsep dari algoritma C4.5 adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan (rule). C4.5 adalah algoritma yang cocok untuk masalah klasifikasi dan data mining. C4.5 memetakan nilai atribut menjadi kelas yang dapat diterapkan untuk klasifikasi baru (Xindong, 2009 dalam Novandya, 2017).

Berikut adalah rumus perhitungan entropy :

Menghitung Algoritma C4.5

Entrophy (S)= $\sum_{i=1}^{n} -pi\ log_2\ pi$

Keterangan :

S = Himpunan Kasus
n = Jumlah partisi S
p_i = probabilitas yang didapat dari jumlah kelas dibagi total kasus

Setelah menghitung nilai entropy dalam algoritma C4.5 pemilihan atribut dilakukan dengan menggunakan Information Gain. Untuk menghitung gain, yang bisa dihitung dengan formula sebagai berikut :

$$Gain\ (S,A) = Entrophy\ (S) = \sum_{i=1}^{n} * Entrophy\ (Si)$$

Keterangan :
S = Himpunan kasus

A = Atribut
n = Jumlah atribut
|Si| = Jumlah partisi ke -i
|S| = jumlah kasus dalam S

Apabila ada atribut yang mempunyai banyak nilai atribut perlu untuk menghitung gain ratio, sebelumnya perlu kita ketahui suatu istilah baru yang disebut split information, yang bisa dihitung dengan formula sebagai berikut :

$$\textbf{\textit{Split Info (S,A)}} = \sum_{i=1}^{c} \frac{Si}{S} log_2 \frac{Si}{S}$$

Keterangan :
S = ruang (data) sampel yang digunakan untuk training
A = atribut
S_i = jumlah sampel untuk atibut i

Dimana S_i sampai S_c adalah subset c yang dihasilkan dari pemecahan S dengan menggunakan atribut A yang mempunyai sebanyak c nilai. Selanjutnya gain ratio dihitung dengan cara :

$$\textbf{\textit{Gain Ratio (S,A)}} = \frac{Gain(S,A)}{SplitInfo(S,A)}$$

2.2.5 *Ada Boost*

AdaBoost adalah salah satu algoritma pengawasan penambangan data yang banyak digunakan untuk membangun model klasifikasi. AdaBoost sendiri pertama kali diperkenalkan oleh Yoav Freund dan Robert Schapire. (1995). (Zulhanif, 2015)

2.2.6 Akurasi

Akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal,akurasi adalah sebagian kecil dari prediksi model kami yang benar.

Secara formal, akurasi memiliki definisi sebagai berikut :

$$\text{Akurasi} = \frac{\text{Number of Correect Prediction}}{\text{Total Number of Prediction}} \quad (2)$$

Untuk klasifikasi biner, akurasi juga dapat dihitung dalam hal positif dan negatif sebagai berikut :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Dimana

TP =

True

Positif TN = True Negatif

FP = False Positif

FN = False Negatif