BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terkait Terdahulu

Tabel 2.1 berikut ini merupakan penelitian terkait dengan penelitian yang sedang dilakukan saat ini :

Tabel 2.1 Penelitian Terkait

No	Nama Penulis	Judul/Tahun Terbit	Uraian
1	Widhi Ramdhani , David Bona , Rafi Bagus Musyaffa , Chaerur Rozikin	Klasifikasi Penyakit Kangker Payudara Menggunakan Algoritma K-Nearest Neighbor Jurnal Ilmiah Wahana Pendidikan, Agustus 2022	Nilai persentasi pengklasifikasian penyakit kangker payudara dengan persentase 62,7% dalam kategori kangker jinak dan 37,3% dalam kategori kangker ganas. Selain itu, dilakukan evaluasi atau pengujian nilai performa dari pemodelan algoritma KNN yang digunakan dengan melakukan beberapa percobaan nilai K [6]. Pengujian dengan nilai akurasi tertinggi diperoleh dari nilai k=21 dan k=11 dengan nilai akurasi mencapai 98%.
2	.Nurul Khairina, Theofil Tri Saputra Sibarani, Rizki Muliono, Zulfikar Sembiring, Muhathir	Identifikasi Pneumonia Menggunakan Metode K-Nearest Neighbors Menggunakan Ekstraksi Fitur Hog Jite, 5 (2) January 2022	Identifikasi lebih detail tentang penyakit Pneunomia dapat bermanfaat pada perkembangan ilmu pengetahuan kususnya dunia kedokteran. Gejala penyakit Pneumonia yang diawali dari batuk biasa, sering sekali dianggap hal yang ringan dan tidak membutuhkan penanganan yang serius, padahal penyakit Pneumonia

			dapat menyebabkan kematian [7] .Penelitan ini mengidentifikasi gejala penyakit Pneumonia dengan metode K-Nearest Neighbor dengan tiga klasifikasi yaitu Fine KNN, Cosine KNN dan Cubic KNN serta dikombinasikan dengan fitur Histogram of Oriented Gradient . Hasil penelitian menunjukkan tingkat akurasi klasifikasi set gambar X-ray Pneumonia menggunakan metode KNN dan HOG memiliki hasil yang berbedabeda. Fine KNN mencapai akurasi sebesar 80,67, Cosine KNN mencapai akurasi sebesar 84.93333, dan Cubic KNN mencapai akurasi sebesar 83.13333. Dari tiga klasifikasi KNN, dapat dilihat bahwa klasifikasi Cosine KNN memiliki nilai yang paling akurat yaitu sebesar 84.93333.
3	Nisa Hanum Harani, Cahyo Prianto	Penerapan algoritma Adaboost guna menentukan pola masuknya calon mahasiswa TRANSFORMTIKA, Vol.18, No.1, July 2020	Penerapan Metode klasifikasi pohon keputusan (decission tree) dan adaboost dapat meningkatkan akurasi hingga 91,35% [8]. Model kombinasi ini dianggap paling akurat jika dibandingkan dengan metode klasifikasi yang hanya menggunakan algoritma pohon keputusan saja (61,4%) . Hasil akurasi menunjukan bahwa model yang dihasilkan dapat melakukan prediksi dengan tepat dalam menentukan pola mahasiswa yang akan benar – benar masuk

			Perguruan Tinggi (PT)
4	Rizki Tri Prasetio , Sari Susanti	Prediksi Harapan Hidup Pasien Kanker Paru Pasca Operasi Bedah Toraks Menggunakan Boosted K-Nearest Neighbor JURNAL RESPONSIF, Vol.1 No.1 Agustus 2019	Prediksi ini dilakukan dengan menganalisa kondisi pasien sebelum dan sesudah operasi z. Data yang digunakan pada penelitian ini merupakan data sekunder yang berisi 470 data dengan sebaran 400 data pasien yang hidup (survival) dan 70 data pasien yang meninggal (die). Adaptive Boost digunakan sebagai optimasi level algoritma pada algoritma k-nearest neighbor. Hasil penelitian menunjukan bahwa metode yang diusulkan menghasilkan akurasi prediksi harapan hidup sebesar 85.11% menggunakan validasi 10 fold cross validation dengan parameter k pada algoritma k-nearest neighbor bernilai 5.
5	Sherly, Delima Sitanggang	Model Prediksi Obesitas Dengan Menggunakan Support Vector Machine Jusikom Prima (Jurnal Sistem Informasi Dan Ilmu Komputer Prima) Vol. 5 No. 2, Februari 2022	manfaat dari penelitian ini adalah untuk memperoleh model prediksi data yang dapat membantu memprediksi nilai persentase lemak pada badan sehingga dapat digunakan untuk kelengkapan data serta penyajian informasi tanpa perlu memperhatikan faktor bentuk badan yang beragam [9]. proses implementasi algoritma svr dapat dengan baik melakukan regresi dengan tingkat akurasi akhir 71.80% dan mse 17.76. sistem prediksi yang dihasilkan dengan algoritma mampu

			membantu dalam penentuan otomatis persentase lemak pada badan tanpa perlu pengukuran densitas badan yang memerlukan pengukuran dalam air dikarenakan volume tubuh manusia yang beragam dan bervariasi.persentase lemak pada badan merupakan informasi yang penting baik untuk keperluan diagnosa maupun sebagai informasi peringatan yang dikarenakan apabila persentase berlebih dapat menyebabkan penyakit beresiko tinggi seperti type-2 diabetes dan penyakit jantung lainnya.

Berdasarkan hasil review beberapa jurnal, dapat disimpulkan bahwa dalam beberapa kasus, tingkat akurasi cenderung lebih tinggi ketika menggunakan algoritma K-Nearest Neighbor (K-NN). K-NN telah terbukti menghasilkan tingkat akurasi yang tinggi dalam berbagai penelitian. Dalam konteks ini, Adaboost dapat digunakan sebagai metode untuk meningkatkan kinerja algoritma K-NN. Adaboost merupakan metode ensemble learning yang digunakan untuk meningkatkan performa model dengan menggabungkan beberapa model lemah menjadi satu model yang kuat. Dengan menerapkan Adaboost pada algoritma K-NN, dapat dilakukan optimasi pada level algoritma tersebut sehingga tingkat akurasi yang didapatkan dapat lebih tinggi lagi. Dengan kombinasi antara K-NN dan Adaboost, kelemahan atau keterbatasan dari K-NN dalam beberapa kasus dapat diatasi, sehingga dapat meningkatkan performa keseluruhan dari model prediktif. Hal ini menunjukkan bahwa penggunaan Adaboost sebagai teknik ensemble dapat memberikan kontribusi yang signifikan dalam meningkatkan akurasi prediksi, terutama ketika digunakan bersama dengan algoritma yang telah terbukti efektif seperti K-NN.

2.2 Obesitas

Obesitas adalah sebuah kondisi kronis yang diakibatkan karena konsumsi kalori berlebihan, obesitas dapat ditandai dengan adanya penumpukan lemak dalam tubuh yang sangat tinggi. Terjadinya obesitas dipengaruhi oleh asupan makanan yang melebihi kebutuhan tubuh, kurangnya aktivitas fisik, dan faktor genetic . Ketidakseimbangan antara asupan kalori dan penggunaan energi tubuh adalah kondisi di mana tubuh menerima lebih banyak kalori dari makanan daripada yang dibutuhkan untuk aktivitas sehari-hari. Hal ini menyebabkan penimbunan lemak yang berlebihan dalam tubuh. Kondisi ini sangat berpotensi meningkatkan risiko terkena berbagai penyakit serius, termasuk diabetes tipe 2, penyakit jantung, dan tekanan darah tinggi. Obesitas, sebagai hasil dari ketidakseimbangan energi ini, dianggap sebagai faktor risiko utama untuk pengembangan berbagai penyakit kronis. Ketika tubuh terlalu banyak menimbun lemak, sel-sel lemak yang berlebihan bisa merusak fungsi organ-organ penting, seperti pankreas dan jantung. Akibatnya, obesitas dapat meningkatkan kemungkinan terkena penyakit jantung koroner, serangan jantung, stroke, serta masalah kesehatan lainnya.

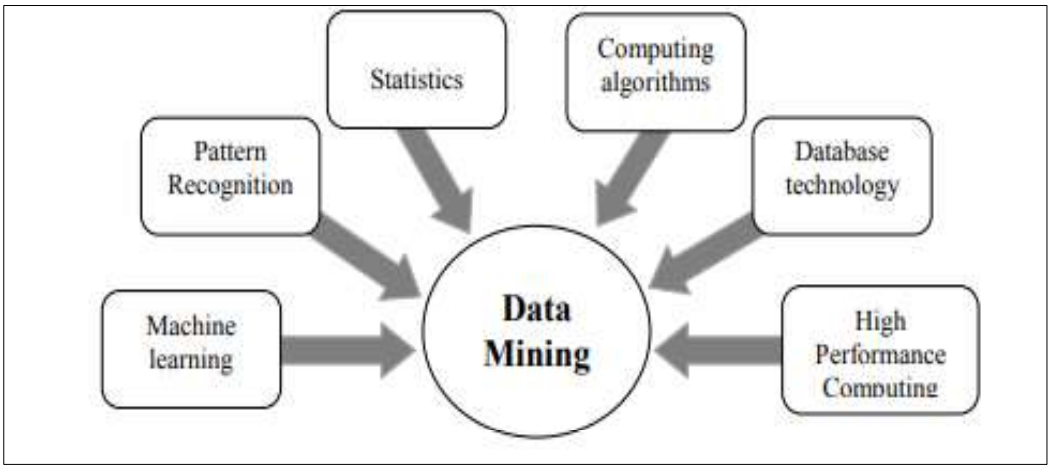
Untuk mengatasi masalah ini, upaya pencegahan obesitas sangat ditekankan. Pencegahan obesitas melibatkan perubahan gaya hidup yang sehat, dimana fokus utamanya adalah mencapai keseimbangan energi yang optimal. Langkah-langkah pencegahan yang direkomendasikan meliputi:

- Mengadopsi pola makan sehat dengan mengonsumsi makanan bergizi dan seimbang.
- Mengurangi konsumsi makanan tinggi lemak, gula, dan garam.
- Rutin berolahraga dan meningkatkan aktivitas fisik sehari-hari.
- Memperhatikan porsi makan dan menghindari makan berlebihan.
- Memiliki waktu istirahat yang cukup dan mengelola stres dengan baik.
- Konsultasi dengan profesional kesehatan untuk mendapatkan saran dan dukungan yang sesuai.

2.3 *Data Mining*

Data mining dikenal sejak tahun 1990-an, ketika adanya suatu pekrjaan yang memanfaatkan data menjadi suatu hal yang lebih penting dalam berbagai bidang, seperti marketing dan bisnis, sains, dan teknik, serta seni dan hiburan. Sebagian ahli menyatakan bahwa *data mining* merupakan suatu lankah untuk menganalisis pengetahuan dalam basis data atau biasa disebut *Knowledge Discovery in Database* (KDD) . *Data mining* merupakan proses untuk menemukan pola data dan pengetahuan yang menarik dari kumpulan data yang sangat besar [10].

Data mining, secara sederhana merupakan suatu langkah ekstraksi untuk mendapatkan informasi penting yang sifatnyaimplisit dan belum diketahui. Data miningmempunyai hubungan dengan berbagai bidang seperti statistic, machine learning, *computing algorithms, database technology.* Gambar 2.1 merupakan diagram hubungan *data mining* :



Gambar 2.1. Diagram Hubungan *Data Mining*

Secara sistematis, langkah utama untuk melakukan *data mining* terdiri dari tahap, yaitu sebagai berikut :

1) Ekspolasi Atau Pemrosesan Awal Data

   Eksplorasi atau pemrosesan awal data terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan missing value, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

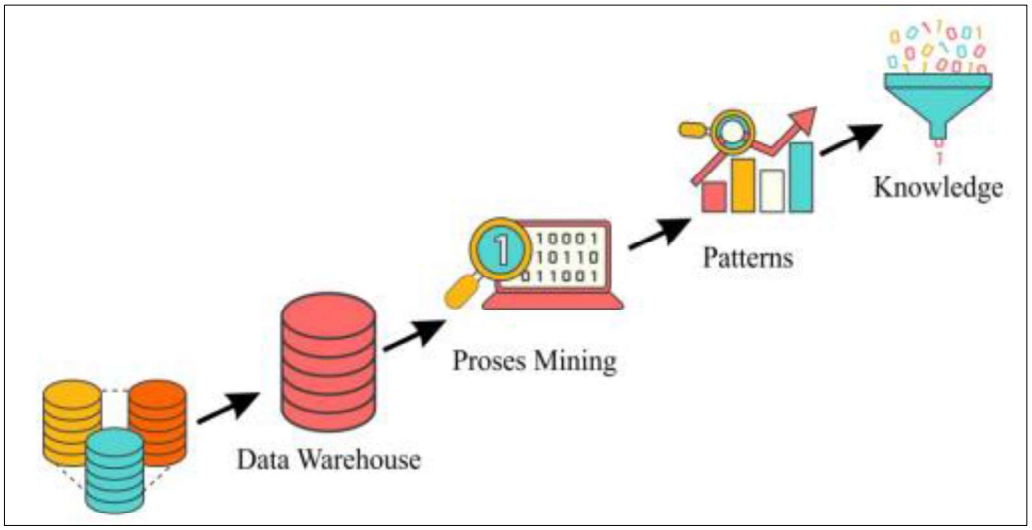
2) Membangun Model Dan Validasi

Membangun model dan validasi, merupakan melakukan analisis dari berbagai model dan memlih model sehingga menghasilkan kinerja yang terbaik. Pembangunan model dilakukan menggunakan metode-metode seperti klasifikasi, regresi,analisis cluster, dan asosiasi.

3) Penerapan

Penerapan dilakukan dengan menerapkan model yang dipilih pada data baru untuk menghasilkan kinerja yang baik pada msalah yang diinvestigasi.

Tahapan proses data mining ada beberapa yang sesuai dengan proses KDD (*Knowledge Discovery in Database*) . Gambar 2.2 merupakan proses KDD (*Knowledge Discovery in Database*) :



Gambar 2.2. Proses KDD (*Knowledge Discovery in Database*)

1. *Cleaning And Integration.*
   a. *Data Cleaning* (Pembersih data)

      Data cleaning (Pembersihan data) adalah proses yang dilakukan untuk menghilangkan noise pada data yang tidak konsisten atau bisa disebut tidak relevan. Data yang diperoleh dari database suatu perusahaan maupun hasil eksperimen yang sudah ada, tidak semuanya memiliki isian yang sempurna

misalnya data yang hilang, data yang tidak valid, atau bisa juga hanya sekedar salah ketik. Data yang tidak relevan itu dapat ditangani dengan cara dibuang atau sering disebut dengan proses cleaning. Proses cleaning dapat berpengaruh terhadap performa dari teknik *data mining*.

b. *Data Integration* (Integrasi Data)

Integrasi data merupakan proses penggabungan data dari berbagai database sehingga menjadi satu database baru. Data yang perlukan pada proses *data mining* tidak hanya berasal dari beberapa database.

2. *Selection and Transformation*
   a. *Data Selection* (Seleksi Data)

Tidak semua data yang ada didatabase akan dipakai, karena hanya data yang sesuai saja yang akan dianalisis dan diambil dari database.Misalnya pada sebuah kasus market basket analysis yang akan meneliti faktor kecenderungan pelanggan, maka tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan.

b. *Data Trnasformation* (Transformasi Data)

Transformasi data merupakan proses pengubahan data dan penggabungan data ke dalam format tertentu, *data mining* membutuhkan format data khusus sebelum diaplikasikan. Misalnya metode standar seperti analysis asosiasi dan clustering haya bias menerima inputan data yang bersifat katagorial. Karenanya data yang berupa angka numeric apabila mempunyai sifat kontinyu perlu dibagi menjdi beberapa interval. Proses ini sering disebut dengan transformasi data.

3. *Poses Mining*

Proses mining dapat disebut juga sebagai proses penambangan data. Proses mining merupakan proses utama yang menggunakan metode untuk menemukan pengetahuaan beharga yang tersembunyi dari data.

4. *Evaluation and Precentation*
   a. Evaluasi Pola *(Pattren Evaluation)*

Evaluasi pola bertugas untuk mengidentifikasi pola-pola yang menarik ke dalam knowledge based yang ditemukan. Pada tahap ini dihasilkan polapola yang khas dari model klasifikasi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai dengan hipotesa, terdapat beberapa alternative yang bias diambil seperti menjadikanya umpan baik untuk memprbaiki proses *data mining*, atau mencoba metode *data mining* lain yang lebih sesuai.

b. Presentasi Pengetahuan *(Knowledge Presentation)*

Knowledge presentation merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan atau informasi yang telah digali oleh pengguna. Tahap terakhir dari proses data mining adalah memformulasikan keputusan dari hasil analisis yang didapat.

2.4 *Particle* ADABOOST (Adaptive Boosting)

Adaboost merupakan akronim dari Adaptive Boosting termasuk kedalam Ensemble Methods /Boosting Methods yang sering dipakai. Boosting diperkenalkan oleh Freund dan Schapire tahun 1995 melalui algoritme AdaBoost dengan konsep dasar peningkatan bobot pengamatan yang salah klasifikasi [18]. Algoritme AdaBoost membangun model pohon gabungan secara sekuensial, yaitu pada setiap iterasi, bobot data dimodifikasi dengan tujuan mengoreksi data yang salah klasifikasi pada iterasi sebelumnya [11]. Secara garis besar proses yang dilakukan dalam Adaboost ialah membangun sejumlah weak learners yang tidak memiliki korelasi satu sama lain, lalu kemudian menggabungkan prediksinya. Dalam penerapannya Adaboost dikombinasikan dengan algoritma lain dengan tujuan untuk mengoptimalisasi performa yang dihasilkan. Adaboost $Hk\ (x)$ [12] didefinisikan sebagai:

$$H_k(x) = \sum_{t=1}^{T} \left(\frac{log\ 1}{\beta_t}\right) h_t^k(x) \qquad \text{.....(1)}$$

Dimana $ht\ k\ (x)$ merupakan weak learners yang memiliki nilai error terendah , sedangkan βt merupakan bobot dari weak learners tersebut. Premis akhir dalam Adaboost dihasilkan dari kombinasi weak learners yang memiliki nilai suara tertinggi

2.5 Algoritma K-NN

Algoritma KNN merupakan algoritma yang banayak digunakan dalam melakukan klasifikasi. KNN merupakan algoritma yang sederhana untuk diimplementasikan tetapi menghasilkan akurasi yang baik .Salah satu kelemahan dari algoritma ini adalah dalam penentuan nilai k, jika nikai k terlalu besar maka akan membuat hasil klasifikasi menjadi tidak jelas atau kabur, sedangkan jika nilai k terlalu kecil atau di misalkan k=1 maka akan menyebabkan hasil klasifikasi terasa kaku karena tidak ada pilihan .Maka dari itu diperlukan penelitian penentuan nilai k yang baik [6]. Algortima K-Nearest Neighbor (KNN) adalah merupakan sebuah metode untuk melakukan klasifikasi terhadap obyek baru berdasarkan (K) tetangga terdekatnya. KNN termasuk algoritma supervised learning, dimana hasil dari query instance yang baru, diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Kelas yang paling banyak muncul yang akan menjadi kelas hasil klasifikasi[13]. Jarak terdekat antara data training akan diukur dengan euclidean distance dan manhattan distance. Menurut penelitian terdahulu, euclidean distance mempunyai hasil lebih akurat dari pada manhattan distance [7].

2.6 Confusion Matrix

Matriks konfigurasi adalah tabel yang terdiri dari jumlah baris data uji yang diprediksi benar dan salah dengan model klasifikasi yang digunakan. Tabel Confusion Matrix diperlukan untuk memilih kinerja terbaik dari sebuah model klasifikasi [14]. Confusion matrix adalah matrix 2x2 yang merepresentasikan hasil klasifikasi biner pada suatu dataset. Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. Hasil dari nilai accuracy, precision dan recall bisa ditampilkan dalam persentase [15].

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

$$\text{Akurasi} = \frac{\text{Number of Correect Prediction}}{\text{Total Number of Prediction}} \quad \ldots\ldots\ldots\ldots\ldots (12)$$

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

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}$$

Dimana TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

2.6.2 Precision

Precision dalam Confusion Matrix didefinisikan sebagai rasio item terkait yang dipilih dengan semua item yang dipilih. Akurasi adalah kemungkinan bahwa item yang dipilih terkait. Dapat diartikan sebagai kecocokan antara permintaan informasi dan respons terhadap permintaan itu [16].

2.6.3 Recal

Recall adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks relevan yang ada pada koleksi [15].Recall merupakan probabilitas bahwa suatu item yang relevan akan dipilih. Recall dapat dihitung dengan jumlah rekomendasi yang relevan yang dipilih oleh user dibagi dengan jumlah semua rekomendasi yang relevan baik dipilih maupun rekomendasi yang tidak terpilih [16]

BAB III
METODOLOGI PENELITIAN

3.1 Metode Penelitian

Metodologi merupakan suatu proses yang dilakukan untuk memecahkan suatu permasalahan yang diangkat dalam suatu penelitian sehingga dapat ditemukan hasil yang akurat dan dapat diambil kesimpulan [9]. Dataset yang digunakan menggunakan dataset dari kaggle https://www.kaggle.com/code/cahyaalkahfi/klasifikasi-obesitas-dengan-keras-r/input dengan jumlah sebanyak 2111 data yang menggunkanan tools rapid miner. Atribut yang digunakan sebanyak 16 atribut yaitu Gender, Age, Height, Weight, family_history_with_overweight, FAVC (Frequence of consumption of high caloric food /Frekuensi konsumsi makanan berkalori tinggi), FCVC (Frequence of consumption of vegetables/ Frekuensi konsumsi sayuran), NCP (Number of main meals/ Jumlah makanan utama), CAEC (Consumption of food between meals/ Konsumsi makanan di antara waktu makan), SMOKE, CH2O(Consumption of water daily/ Konsumsi air setiap hari), SCC (Calories consumption monitoring/ Pemantauan konsumsi kalori), FAF (Physical activity frequency/ Frekuensi aktivitas fisik), TUE (Time using technology devices/ Waktu menggunakan perangkat teknologi), CALC (Calories consumption monitoring/ Pemantauan konsumsi kalori), MTRANS (Transportation used/ Transportasi yang digunakan). Pada Bab ini akan membahas langkah-langkah dalam proses penelitian yang akan dilakukan. Tujuan utamanya adalah untuk menganalisis dan mencari pola data yang akan digunakan sebagai dataset. Hal ini akan membantu mempermudah penelitian dan menjalankannya secara sistematis sesuai dengan tujuan yang diinginkan. Untuk mencapai hal tersebut, maka dirancang langkah-langkah berikut sebagai alur dalam tahapan penelitian yang akan dilakukan: