PENERAPAN DATA MINING DENGAN MENGGUNAKAN ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT DIABETES

PROPOSAL SKRIPSI

Disusun Oleh :

Muhammad Hanifudin

8020190047

Untuk Persyaratan Penelitian Dan Penulisan Tugas Akhir

Sebagai Akhir Proses Studi Strata 1

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS DINAMIKA BANGSA

2022

IDENTITAS PROPOSAL PENELITIAN

Judul Proposal : Penerapan Data Mining Dengan Menggunakan Metode C4.5 Untuk Prediksi Penyakit Diabetes

Program Studi : Teknik Informatika

Jenjang Pendidikan : Strata 1 (S1)

Peneliti :

a. Nama Lengkap : Muhammad Hanifudin
b. NIM : 8020190047
c. Jenis Kelamin : Laki-Laki
d. Tempat/Tgl Lahir : Sungai Pinang/03 Juni 2001
e. Alamat : Jl. Lingkar Selatan, RT.33, Kec. Palmerah, Jambi
f. No. Telepon : 082183857981
g. Email : hanif14azaz@gmail.com

ABSTRAK

Penyakit diabetes merupakan salah satu masalah Kesehatan terbesar di seluruh dunia dengan perkiraan sedikitnya 463 juta penderita. Angka tersebut akan semakin meningkat. Jumlah ini bisa bertambah apabila masyarakat tidak mengetahui faktor-faktor yang dapat menyebabkan adanya diabetes. Penelitian ini bertujuan untuk membuat model prediksi menggunakan metode C4. 5 yang menghasilkan sebuah pohon keputusan dengan pengujian nya menggunakan WEKA dan Rapidminer agar pencegahan terhadap penyakit diabetes dapat diklasifikasikan sedini mungkin. Hasil penelitian ini akan dijadikan referensi untuk melihat apakah seorang beresiko terkena diabetes yang berdasarkan pada atribut yang telah ditetapkan pada dataset.

Kata Kunci : Klasifikasi, *Algoritma C4.5*, Penyakit Diabetes

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa yang telah memberikan berkat dan rahmat-Nya sehingga penulis dapat menyelesaikan proposal skripsi yang berjudul “Penerapan Data Mining Dengan Menggunakan Metode C4.5 Untuk Prediksi Penyakit Diabetes” dengan sebaik-baiknya.

Penulis menyadari bahwa dalam penulisan proposal skripsi ini banyak mengalami kendala dan kekurangan, namun berkat bantuan, bimbingan, dan petunjuk dari banyak pihak sehingga kendala-kendala yang dihadapi tersebut dapat diatasi dan penulis dapat menyelesaikan proposal skripsi ini dengan baik.

Penulis berharap agar proposal skripsi ini dapat bermanfaat bagi kita semua. Akhir kata penulis mengucapkan terima kasih.

Jambi, September 2022

Penulis

DAFTAR ISI

BAB I
PENDAHULUAN

1.1. LATAR BELAKANG

Menurut WHO diabetes adalah penyakit yang paling mematikan ke-9 yang ada di dunia. Salah satu yang menyebabkan meningkatnya jumlah penderita diabetes adalah karena keterlambatan diagnosis penyakit diabetes itu sendiri.

Diabetes adalah penyakit tanpa gejala subjektif yang jelas, dan orang tidak menyadari bahwa mereka menderita diabetes. Menurut [1] diabetes adalah penyakit yang mempengaruhi pankreas dengan ketidakmampuan untuk memproduksi insulin dengan benar. Insulin merupakan salah satu hormon yang di hasilkan dari pankreas, yang bertugas sebagai pintu untuk menyalurkan glukosa dari makan yang diserap untuk di alirkan ke dalam sel-sel darah agar tubuh dapat menghasilkan energy untuk di gunakan [2]. Diabetes menjadi berbahaya apabila kadar gula dalam darah sudah melebihi batas normal. Jika diabetes tidak diobati dengan benar maka dapat menyebabkan berbagai macam masalah yang akan menjadi kompilkasi berbagai penyakit yang akan mengancam nyawa pasien.

Data mining digunakan untuk melakukan proses ekstraksi informasi yang tersembunyi dari dataset yang banyak dan terdapat beberapa teknik dalam data mining seperti klasifikasi ,clustering, regresi dan asosiasi yang akan di gunakan dalam data pada bidang medis [3]. Algoritma C4.5 merupakan salah satu Algoritma yang dapat digunakan untuk melakukan mengklasifikasikan atau mengelompokkan kumpulan data. Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (Decision Tree). Pohon keputusan dapat membantu memecahkan masalah dalam pengambilan keputusan mengenai alternatif-alternatif yang ada pada masalah tersebut.

Oleh karena itu di butuhkan prediksi lebih awal untuk penyakit diabetes berdasarkan dari atribut yang terdapat pada dataset menurut [4]

Salah satu cara untuk klasifikasi penyakit diabetes dengan menggunakan data mining. Didalam melakukan prediksi data dalam penelitian ini, metode yang di gunakan adalah Decision Tree C4.5. Metode Decision Tree dengan menggunakan algoritma C4.5 bisa melakukan prediksi dari berbagi informasi berdasarkan data yang digunakan untuk menghitung kemungkinan terjadinya penyakit berdasarkan atributnya dapat digunakan dan juga bisa dilihat seberapa efektifnya algoritma Decision Tree untuk deteksi penyakit diabetes**.**

1.2. RUMUSAN MASALAH

Berdasarkan uraian latar belakang, maka dapat dirumuskan masalah yang akan dibahas dalam penelitian ini adalah bagaimana mendeteksi penyakit diabetes menggunakan metode C4.5.

1.3. BATASAN MASALAH

Adapun batasan masalah yang terdapat dalam penelitian ini antara lain :

1. Data yang digunakan bersumber dari *dataset* penyakit diabetes.
2. Variabel yang digunakan dalam penelitian ini adalah pregnencies, glucose, blood preasure, skin thickness, insulin, bmi, diabetes pedigree function, age, outcome.
3. Metode yang digunakan yaitu *algoritma C4.5*

1.4. TUJUAN DAN MANFAAT PENELITIAN

1.4.1. Tujuan Penelitian

Adapun tujuan dari penelitian yang akan dilakukan oleh penulis, yaitu :

1. Menerapkan metode Algoritma C4.5 untuk klasifikasi penyakit diabetes
2. Mendapatkan akurasi yang tepat untuk melakukan klasifikasi penyakit diabetes dengan menggunakan *Algoritma C4.5*.
3. Mengetahui tingkat akurasi *Algoritma C4.5* dalam klasifikasi penyakit diabetes.

1.4.2. Manfaat Penelitian

Adapun beberapa manfaat yang akan didapat dalam melakukan penelitian ini, yaitu :

1. Dapat memberikan kemudahan dalam klasifikasi penyakit diabetes dalam dunia medis.
2. Penelitian ini dapat dijadikan sebagai referensi bagi peneliti berikutnya yang berkaitan dengan menggunakan metode *Algoritma C4.5* di dunia pendidikan.

BAB II
LANDASAN TEORI

2.1. DEFINISI PENYAKIT

Penyakit adalah kondisi buruk pada organ tertentu yang mempengaruhi struktur atau fungsi sebagian atau seluruh tubuh makhluk hidup dan bukan merupakan akibat langsung dari cedera eksternal.

2.2. DATA MINING

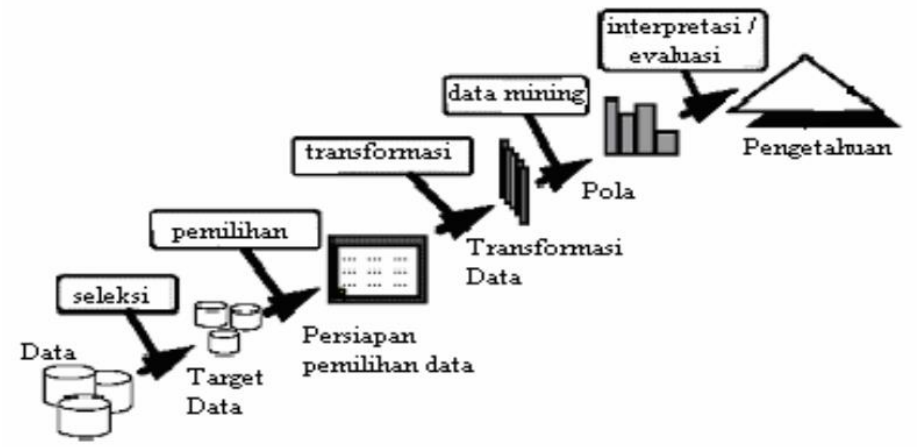
Data mining sering juga disebut *Knowledge Discovery in Database,* adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. *Data Mining* juga dapat diartikan sebagai menambang data atau upaya untuk menggali informasi yang berharga dan berguna pada database yang sangat besar [5].

Hal penting yang terkait dengan data mining adalah :

1. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan di proses berupa data yang sangat besar.
3. Tujuan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

2.2.1. Tahapan Proses dalam *Data Mining*

Ada beberapa tahapan dalam proses *data mining*. Diagram dibawah ini menggambarkan beberapa tahap/proses yang berlangsung dalam *data mining*.



Gambar 2.1. Fase – Fase

Tahapan dalam proses *data mining* dapat dijelaskan sebagai berikut :

1. **Seleksi data**

   Pemilihan (seleksi) data baru sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *data mining* dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. ***Pre-processing/Cleaning*** **(pemilihan data)**

   Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

3. **Transformasi**

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*.

4. ***Data Mining***

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode dan algoritma sesuai dengan kebutuhan dan tujuan.

5. **Interpretasi/Evaluasi**

   Pola informasi yang dihasilkan dari informasi *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini mencakup pemeriksaan apakah pola atau

informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

2.2.2. Pengelompokkan *Data Mining*

Pengelompokkan *data mining* dapat diklasifikasikan berdasarkan fungsi yang dilakukan atau berdasarkan jenis aplikasi yang menggunakannya [6] :

1. Klasifikasi

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

Contoh lain klasifikasi dalam bisnis dan penelitian adalah :

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
c. Mendiagnosa penyakit seorang pasien untuk mendapatkan termasuk kategori apa.

2. Pengklusteran (*Clustering*)

Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan suatu dengan yang lainnya dan memiliki ketidakmiripan dengan *record* dalam kluster lain.Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.

Contoh pengklusteran dalam bisnis dan penelitian adalah :

a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.
b. Untuk tujuan audit akutansi, yaitu melakukan pemisahan terhadap prilaku finansial dalam baik dan mencurigakan.

3. Asosiasi

Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

Contoh asosiasi dalam bisnis dan penelitian adalah :

a. Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respon positif terhadap penawaran *upgrade* layanan yang diberikan.
b. Menemukan barang dalam supermarket yang dibeli secara bersamaan dan barang yang tidak pernah dibeli bersamaan.

4. Deskripsi

Terkadang analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

5. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan dara sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel

prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

6. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa mendatang.

2.3. KLASIFIKASI

Klasifikasi adalah tugas pembelajaran sebuah fungsi target f yang memetakan setiap himpunan atribut x ke salah satu label *class* y yang telah didefinisikan sebelumnya [7].

Klasifikasi dapat juga diartikan suatu proses untuk menemukan suatu model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan dapat menggunakan model untuk memprediksi kelas objek yang label *class*-nya tidak diketahui.

2.4. *DECISION TREE & ALGORITMA C4.5*

Pohon keputusan (*Decision Tree)* merupakan salah satu metode data mining yang banyak diterapkan sebagai solusi untuk mengklasifikasikan masalah. *Decision tree* adalah suatu struktur yang berbentuk seperti pohon dimana setiap simpul (*node*) pohon mewakili atribut yang telah diuji. Setiap cabang mewakili hasil yang telah diuji dan simpul daun (*leaf node*) mewakili kelas atau distribusi kelas [8]. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk table dengan atribut dan record. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon [9]. *Decision tree* banyak digunakan karena dapat menggabungkan sebuah pola/informasi kedalam sebuah bentuk pohon keputusan.

Terdapat tiga jenis *node* pada *decision tree*:

1. *Root node*, merupakan simpul yang tidak memiliki input tetapi memiliki output yang lebih dari satu.
2. *Internal node*, memiliki satu input dan memiliki output lebih dari dua
3. *Leaf* atau *terminal node*, mempunyai satu input dan tidak mempunyai output.Pada setiap *decision tree*, setiap leaf memiliki sebuah nama kelas.

Algoritma C4.5 merupakan salah satu metode yang sering digunakan untuk membuat Decision Tree berdasarkan training data yang telah disediakan. Algoritma C4.5 memetakan atribut menjadi kelas yang dapat diterapkan untuk klasifikasi baru [10].

2.5. PENELITIAN SEJENIS

Tabel 2.1. Penelitian Sejenis

No.	Judul, Penulis dan Tahun	Metode	Atribut	Akurasi
1	"Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus". Achmad Ridwan. 2020. [2].	*Naïve Bayes*	Age, Sex, Polyuria, Polydipsia, Sudden weight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis , Muscle stiffness, Alopecia, Obesity, Class	Hasil evaluasi klasifikasi ini menghasilkan akurasi untuk mengklasifisikan Dataset Diabetes yaitu 90,20%.
2	"Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5". Sanni Ucha Putri, Eka	*Algoritma C4.5*	Usia(Tahun), Tekanan Darah(NmHg), Denyut Nadi(s/menit), Berat Badan(Kg), Kadar gula darah(mg/dl), Variabel	Dengan menerapkan Algoritma C4.5 pada software Rapidminner menghasilkan hasil yang sama pada perhitungan manual.

	Irawan, Fitri Rizky. 2021. [4].			Berdasarkan pengolahan data menggunakan software RapidMiner didapat nilai akurasi sistem sebesar 90,00%, artinya bahwa rule yang dihasilkan tingkat kebenaran mendekati 100%.
3	"Penerapan Algoritma C4.5 Untuk Membuat Model Prediksi Pasien Yang Mengidap Penyakit Diabetes". Sunanto, Ghazi Falah. 2022. [1]	*Algoritma C4.5*	Usia, Jenis Kelamin, Sering buang air kecil, Sering haus, Penurunan berat badan secara tiba-tiba, Lemas ,Banyak makan, Genital thrush, Pandangan kabur, Gatal-gatal, Mudah marah, Luka susah sembuh, Partial paresis, Muscle stiffness, Kebotakan, Kegemukan, dan Class.	Prediksi penyakit menggunakan algoritma Decision TreeC4.5 memiliki hasil yang bagus dapat dilihat dari hasil perhitungan confusion matrix yang mendapatkan hasil accuracy sebesar 95,51% dan juga hanya mendapatkan error classification sebesar 4,49%..
4	"Penerapan Algoritma C4.5 Untuk Klasifikasi Data Rekam Medis berdasarkan International Classification Diseases (ICD-	*Algoritma C4.5*	Jenis kelamin, Umur pasien, Bulan masuk pasien ke rumah sakit, Group ICD	Hasil analisa menunjukan bahwa algoritma C4.5 berhasil mengelompokan penyakit ke dalam 13 jenis kategori dari 21 jenis kategori

	10)". Yudha Aditya Fiandra, Sarjon Defit, Yuhandri. 2017. [11].			yang menjadi label tujuan berdasarkan ICD (International Code Diseases) atau kode penyakit internasional, sehingga dapat dikatakan bahwa Algoritma C4.5 berhasil mendefinisikan 61,9% dari kategori label tujuan yang ada.
5	"Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Mellitus di Rumah Sakit UNHAS Menggunakan Algoritma C4.5". Dewi Rahma Ente, Sri Astuti Thamrin, Hedi Kuswanto, Samsul Arifin, Andreza. 2020. [12].	*Algoritma C4.5*	Jenis Kelamin, Usia, Berat badan, Tinggi badan, GDP, HDL, LDL, Kolestrol total, Trigliserida	Pengukuran akurasi data latih dan data uji dari algoritma C4.5 dengan validasi silang lipat 10 setelah proses seleksi atribut dan nilai akurasi nya memiliki rentang antara 50% sampai dengan 100% dengan tingkat akurasi rata-rata prediksi yaitu 98,5%.
6	"Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes". Andika Prasatya, Riki Ruli.A. Siregar,	*K-Means* dan *Algoritma C4.5*	*Time in Hospital, Numbers of Lab Procedures, Numbers of Procedures, Number of Diagnoses,*	Hasil prediksi yang didapatkan akan dilakukan proses validasi untuk mengetahui tingkat keakurasian dengan

	Rakhmat Arianto. 2020.[13]		*Gender, Age, Admission Type, Discharge Disposition, Admission Source, Diagnosis 1, Diagnosis 2, Diagnosis 3, A1c test result, Change of Medications, Diabetes Medications, Readmitted*	menggunakan K-Fold Cross Validation. Nilai akurasi yang didapatkan sebesar 72%.
7	"Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5". Fida Maisa Hana. 2020.[14]	*Algoritma C4.5*	Umur, Jenis Kelamin, *Polyuria, Polydipsia, Suddenweight loss, Weakness, Polyphagia, Genital thrush, Visual blurring, Itching, Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesitas,* Kelas	Dari hasil Pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%

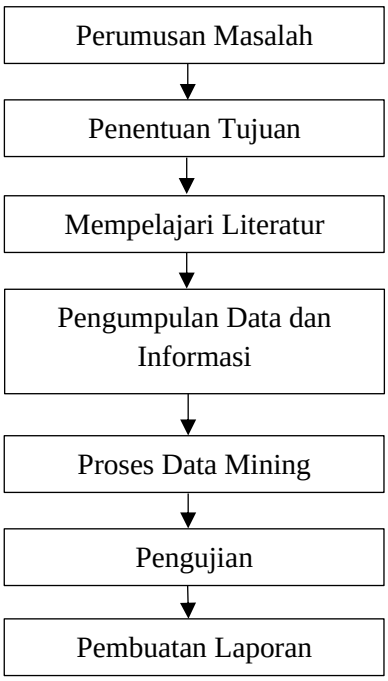
8	"Application Of Data Mining To Identify Diabetes Mellitus Using The Support Vector Machine (SVM) Algorithm And kNN". Windania Purba, Yessy, Riski Nofarianus Gulo. 2022.[15]	*Support Vector Machine (SVM)* dan *K-Nearest Neighbor* (*KNN)*	Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes pedigree Function, Age.	From the results of this study, it can be concluded that the Support Vector Machine (SVM) algorithm produces an accuracy value of 76% while the accuracy value of the K-Nearest Neighbor (KNN) algorithm is 75%.
9	"Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree". Shuja Mirza, Dr. Sonu Mittal, Dr. Majid Zaman. 2018.[16]	*SMOTE* dan *Decision tree*	*Age, Fasting, Post-Prandial glucose, Waist, BMI, Systolic, Diastolic, Hba1c, Gender, History, Class*	
10	"Klasifikasi Resiko Diabetes Tahap Awal Menggunakan Metode rat dan Algoritma C4.5". Febby Putri Marshanda, Mishabatuz Zolam, Bijanto. 2022.[17]	*Naïve Bayes* dan *Algoritma C4.5*	*Age, Sex, Polyuria, Polydipsia, Sudden Weight Loss, Weakness, Polyphagia, Genital Thrush, Visuall Blurring, Itching, Irritability, Delayed Healing, Partial Paresis, Muscle Stiness,*	Hasil akurasi pada algoritma C4.5 untuk menentukan klasifikasi resiko diabetes tahap awal dengan hasil nilai akurasi terbaik sebesar 95.96% presisi sebesar 93,24% dan recall sebesar

			Alopecia, Obesity, Class	96,60% menggunakan data dari UCI dataset.

BAB III
METODOLOGI PENELITIAN

3.1. KERANGKA KERJA PENELITIAN

Untuk membantu dalam penyusunan penelitian ini, maka perlu adanya kerangka kerja (*framework*) yang jelas tahapan-tahapannya. Kerangka kerja ini merupakan langkah- langkah yang akan dilakukan dalam penyelesaian masalah yang akan dibahas. Adapun kerangka kerja penelitian yang akan digunakan adalah sebagai berikut :



Gambar 3.1. Tahapan Penelitian

Berdasarkan kerangka kerja penelitian yang telah di gambar di atas, maka dapat di uraikan pembahasan masing masing tahapan dalam penelitian adalah sebagai berikut :

1. **Perumusan Masalah**

   Masalah yang dirumuskan dalam penelitian ini adalah bagaimana penerapan metode *Decision Tree* dengan *Algoritma C4.5* dalam klasifikasi penyakit diabetes.

2. **Penentuan Tujuan**

   Tujuan yang akan dicapai dalam penelitian ini adalah ingin mengetahui berapa besar tingkat akurasi *Algoritma C4.5* dalam klasifikasi penyakit diabetes.

3. **Mempelajari Literatur**

   Mempelajari literatur-literatur yang dapat mencapai tujuan penelitian, literatur-literatur bersumber dari buku-buku perpustakaan Universitas Dinamika Bangsa Jambi dan jaringan internet. Literatur-literatur yang digunakan nanti dilampirkan dalam daftar pustaka.

4. **Pengumpulan Data dan Informasi**

   Dalam pengumpulan data, penulis mendapatkan dataset online yang terdapat pada sebuah website di internet.

5. **Proses Data Mining**

   Data Mining adalah proses pengektrasian *knowledge* yang tersimpan dalam dataset bervolume besar. Untuk mendapatkan *knowledge* dalam dataset digunakanlah *Algoritma C4.5*.

6. **Pengujian**

   Pada tahap ini dilakukan pengujian dari hasil yang didapat dari tahap sebelumnya sebagai pedoman untuk mendapatkan hasil klasifikasi penyakit diabetes.

7. **Pembuatan Laporan**

   Pada tahap ini dilakukan pembuatan laporan, membuat hasil akhir dari suatu kegiatan penelitian berdasarkan data dan fakta yang telah diamati pada saat meneliti.

3.2. ANALISA TEKNIK *ALGORITMA C4.5*

Pada bagian ini data dan informasi yang diperoleh dan diproses dengan menggunakan Metode *Decision Tree* dengan *Algoritma C4.5* untuk mendapatkan hasil yang sesuai.

3.3. METODE PENELITIAN

Metode pengumpulan data yang digunakan dalam penelitian adalah mencari dan mengambil *dataset* online dari sebuah website yang bernama "*www.kaggle.com*" untuk dijadikan pengujian dalam penelitian ini. Pengumpulan data dilakukan untuk memperoleh hasil dan informasi yang dibutuhkan dalam hal mencapai tujuan penelitian.

3.4. ALAT PENELITIAN

Penulis menggunakan beberapa alat/piranti yang digunakan untuk melakukan pengolahan data/bahan penelitian, yaitu:

1. Hardware, dengan spesifikasi sebagai berikut:
    a. Laptop, dengan processor Ryzen 3200U (2.40 GHz)
    b. RAM : 8.00 GB
    c. SSD : 500 GB
2. Software, dengan keterangan sebagai berikut:
    a. OS Windows 11 (64 bit)
    b. Microsoft Excel & Word 2019
    c. WEKA
    d. RapidMiner Studio Professional 7.1.001.

3.5. JADWAL PENELITIAN

No	Jenis Kegiatan	Bulan																							
		Sept				Okt				Nov				Des				Jan				Feb			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Perumusan Masalah																								
2	Penentuan Tujuan																								
3	Mempelajari Literatur																								
4	Pengumpulan Data & Informasi																								
5	Proses Data Mining																								
6	Pengujian																								
7	Pembuatan Laporan																								

DAFTAR PUSTAKA

[1] N. Sunanto and G. Falah, "Penerapan Algoritma C4.5 Untuk Membuat Model Prediksi Pasien Yang Mengidap Penyakit Diabetes," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 7, no. 2, pp. 208–216, 2022, doi: 10.36341/rabit.v7i2.2435.

[2] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 15–21, 2020, doi: 10.47970/siskom-kb.v4i1.169.

[3] C. Harvinder and C. Anu, "Implementation of decision tree C4. 5 algorithm," *Int. J. Sci. Res. Publ.*, vol. 3, no. 10, 2013.

[4] S. Ucha Putri, E. Irawan, F. Rizky, S. Tunas Bangsa, P. A. -Indonesia Jln Sudirman Blok No, and S. Utara, "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5," *Januari*, vol. 2, no. 1, pp. 39–46, 2021.

[5] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and a I. Verkamo, "Fast discovery of association rules," *Advances in knowledge discovery and data mining*, vol. 12. pp. 307–328, 1996.

[6] E. T. L. Kusrini and E. Taufiq, "Algoritma data mining," *Yogyakarta Andi Offset*, 2009.

[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques, 2nd editionMorgan Kaufmann Publishers," *San Fr. CA, USA*, 2006.

[8] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," *J. Pilar Nusa Mandiri*, vol. 14, no. 2, p. 217, 2018, doi: 10.33480/pilar.v14i2.926.

[9] Y. Mardi, "Data Mining: Classification Using the C4. 5 Algorithm," *J. Informatics Educ.*, vol. 2, no. 2, pp. 213–219, 2017.

[10] A. Novandya, "Penerapan Algoritma Klasifikasi Data Mining C4.5 pada Dataset Cuaca Wilayah Bekasi," *KNiST*, pp. 368–372, 2017.

[11] Y. A. Fiandra, S. Defit, and Y. Yuhandri, “Penerapan Algoritma C4.5 untuk Klasifikasi Data Rekam Medis berdasarkan International Classification Diseases (ICD-10),” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 2, pp. 82–89, 2017, doi: 10.29207/resti.v1i2.48.

[12] D. R. Ente, S. A. Thamrin, S. Arifin, H. Kuswanto, and A. Andreza, “Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5,” *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 80–88, 2020, doi: 10.29244/ijsa.v4i1.330.

[13] A. Prasatya, R. R. A. Siregar, and R. Arianto, “Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes,” *Petir*, vol. 13, no. 1, pp. 86–100, 2020, doi: 10.33322/petir.v13i1.925.

[14] F. M. Hana, “Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5,” *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.

[15] W. Purba and R. N. Gulo, “APPLICATION OF DATA MINING TO IDENTIFY DIABETES MELLITUS USING THE SUPPORT VECTOR MACHINE (SVM) ALGORITHM AND KNN,” vol. 10, no. 2, pp. 994–1000, 2022.

[16] S. Mirza, S. Mittal, and M. Zaman, “Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree,” *Int. J. Appl. Eng. Res.*, vol. 13, no. 11, pp. 9277–9282, 2018, [Online]. Available: http://www.ripublication.com.

[17] F. P. Marshanda and M. Zolam, “Klasifikasi Resiko Diabetes Tahap Awal Menggunakan Metode Naive Bayes dan Algoritma C4 . 5,” *AMRI (Analisa, Meetode, Rekayasa, Iformatika)*, vol. 1, no. 1, pp. 9–15, 2022, doi: 10.12487/AMRI.v1i1.xxxxx.

LAMPIRAN DATASET

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	0	0	39.4	0.257	43	1
1	97	66	15	140	23.2	0.487	22	0
13	145	82	19	110	22.2	0.245	57	0
5	117	92	0	0	34.1	0.337	38	0
5	109	75	26	0	36	0.546	60	0
3	158	76	36	245	31.6	0.851	28	1
3	88	58	11	54	24.8	0.267	22	0
6	92	92	0	0	19.9	0.188	28	0
10	122	78	31	0	27.6	0.512	45	0
4	103	60	33	192	24	0.966	33	0
11	138	76	0	0	33.2	0.42	35	0
9	102	76	37	0	32.9	0.665	46	1
2	90	68	42	0	38.2	0.503	27	1
4	111	72	47	207	37.1	1.39	56	1
3	180	64	25	70	34	0.271	26	0
7	133	84	0	0	40.2	0.696	37	0
7	106	92	18	0	22.7	0.235	48	0
9	171	110	24	240	45.4	0.721	54	1
7	159	64	0	0	27.4	0.294	40	0
0	180	66	39	0	42	1.893	25	1
1	146	56	0	0	29.7	0.564	29	0
2	71	70	27	0	28	0.586	22	0
....	....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....	....
1	103	80	11	82	19.4	0.491	22	0
1	101	50	15	36	24.2	0.526	26	0
5	88	66	21	23	24.4	0.342	30	0