PENERAPAN ALGORITMA *DECISION TREE* C4.5 DALAM MENDIAGNOSIS PENYAKIT BATU GINJAL BERDASARKAN DATA KLINIS

Tugas Akhir

Diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

pada Program Studi S1 Sains Data

Direktorat Kampus Purwokerto Universitas Telkom

21110026

Safina Octaviana Putri

PROGRAM STUDI S1 SAINS DATA

DIREKTORAT KAMPUS PURWOKERTO

UNIVERSITAS TELKOM

PURWOKERTO

2025

PENERAPAN ALGORITMA *DECISION TREE* C4.5 DALAM MENDIAGNOSIS PENYAKIT BATU GINJAL BERDASARKAN DATA KLINIS

Tugas Akhir

Diajukan untuk memenuhi salah satu syarat

memperoleh gelar sarjana

pada Program Studi S1 Sains Data

Direktorat Kampus Purwokerto Universitas Telkom

21110026

Safina Octaviana Putri

PROGRAM STUDI S1 SAINS DATA

DIREKTORAT KAMPUS PURWOKERTO

UNIVERSITAS TELKOM

PURWOKERTO

LEMBAR PENGESAHAN

PENERAPAN ALGORITMA *DECISION TREE* C4.5 DALAM MENDIAGNOSIS PENYAKIT BATU GINJAL BERDASARKAN DATA KLINIS

IMPLEMENTATION OF THE C4.5 DECISION TREE ALGORITHM IN DIAGNOSING KIDNEY STONE DISEASE BASED ON CLINICAL DATA

21110026
Safina Octaviana Putri

Tugas akhir ini telah diterima dan disahkan untuk memenuhi sebagian syarat memperoleh gelar pada Program Studi S1 Sains Data

Direktorat Kampus Purwokerto

Universitas Telkom

Purwokerto, 14 Mei 2025

Menyetujui
Pembimbing I,

Dr. Ridwan Pandiya, S.Si., M.Sc
NIP. 15820053

Penguji I,

Dr. Yogo Dwi Prasetyo, S.Si., M.Si.
NIP. 22870003

Penguji II,

Paradise, S.Kom., M.Kom
NIP. 22950016

Ketua Program Studi
Sarjana S1 Sains Data

Siti Khomsah, S.Kom., M.Cs
NIP. 23810002

Universitas Telkom

LEMBAR PERNYATAAN

Dengan ini saya, Safina Octaviana Putri, menyatakan sesungguhnya bahwa Tugas Akhir saya dengan judul Penerapan Algoritma Decision Tree C4.5 dalam Mendiagnosis Penyakit Batu Ginjal Berdasarkan Data Klinis beserta dengan seluruh isinya adalah merupakan hasil karya sendiri, dan saya tidak melakukan penjiplakan yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan, serta produk dari tugas akhir bukan merupakan produk dari *Generative AI*. Saya siap menanggung risiko/sanksi yang diberikan jika dikemudian hari ditemukan pelanggaran terhadap etika keilmuan dalam Laporan TA atau jika ada klaim dari pihak lain terhadap keaslian karya.

Purwokerto, 30 April 2025

Yang menyatakan

Safina Octaviana Putri

21110026

ABSTRAK

Penerapan Algoritma *Decision Tree* C4.5 dalam Mendiagnosis Penyakit Batu Ginjal Berdasarkan Data Klinis

Oleh

Safina Octaviana Putri

21110026

Batu ginjal merupakan penyakit yang disebabkan oleh pembentukan materi keras dalam urin akibat interaksi antara garam dan mineral. Penyakit ini dapat menyebabkan infeksi saluran kemih berulang, gangguan fungsi ginjal, hematuria (darah dalam urin), hingga berisiko menyebabkan kanker ginjal. Penyakit batu ginjal telah diderita oleh sekitar 1.499.400 orang di Indonesia sehingga diperlukan prediksi penyakit batu ginjal. Penelitian ini bertujuan untuk membangun model klasifikasi untuk memprediksi penyakit batu ginjal berdasarkan data klinis pasien dengan menggunakan algoritma *Decision Tree* C4.5 karena efisien dalam menangani data numerik dan menghasilkan model yang mudah diinterpretasikan. Data yang digunakan berasal dari RSUD Cideres Majalengka dengan 7 variabel klinis yang digunakan. Model dievaluasi menggunakan metode validasi silang (*cross-validation*) dengan 5 lipatan. Hasil penelitian menunjukkan bahwa model mampu mencapai akurasi sebesar 91,39% pada data latih dan 89,88% pada data uji. Model juga diuji pada 929 data uji dan berhasil memprediksi 835 data secara tepat, dengan akurasi keseluruhan sebesar 90%. Uji validasi silang menunjukkan akurasi rata-rata sebesar 87,44% yang menandakan performa model yang konsisten dan tidak mengalami *overfitting* maupun *underfitting*.

Kata Kunci: Batu ginjal, klasifikasi, diagnosis, deteksi dini, Decision Tree C4.

ABSTRACT

Implementation of the C4.5 Decision Tree Algorithm in Diagnosing Kidney Stone Disease Based on Clinical Data

By

Safina Octaviana Putri

21110026

Kidney stones are a disease caused by the formation of hard material in the urine due to the interaction between salts and minerals. This disease can cause recurrent urinary tract infections, impaired kidney function, hematuria (blood in the urine), to the risk of causing kidney cancer. Kidney stone disease has been suffered by around 1,499,400 people in Indonesia so that prediction of kidney stone disease is needed. This study aims to build a classification model to predict kidney stone disease based on patient clinical data using the C4.5 Decision Tree algorithm because it is efficient in handling numerical data and producing models that are easy to interpret. The data used comes from RSUD Cideres Majalengka with 7 clinical variables used. The model was evaluated using the cross-validation method with 5 folds. The results showed that the model was able to achieve an accuracy of 91.39% on training data and 89.88% on test data. The model was also tested on 929 test data and successfully predicted 835 data correctly, with an overall accuracy of 90%. The cross-validation test showed an average accuracy of 87.44%, indicating consistent model performance and no overfitting or underfitting.

Keywords: Kidney stones, classification, diagnosis, early detection, Decision Tree C4.5

KATA PENGANTAR

Puji Syukur kepada Allah SWT berkat Rahmat, Hidayah, dan Karunia-Nya kepada kita semua sehingga penulis dapat menyelesaikan tugas akhir berjudul "Penerapan Algoritma *Decision Tree* C4.5 dalam Mendiagnosa Penyakit Batu Ginjal Berdasarkan Data Klinis" yang disusun sebagai salah satu syarat memperoleh gelar sarjana pada program Strata-1 di Jurusan Sains Data, Universitas Telkom Purwokerto. Penulis berharap tugas akhir ini dapat memberikan ilmu pengetahuan khususnya dalam bidang data sains dan prediksi data klinis berbasis *machine learning*. Penulis menyadari dalam penyusunan tugas akhir ini tidak lepas dari bimbingan, bantuan dan dukungan dari banyak pihak. Oleh karena itu, penulis menyampaikan ucapan terima kasih kepada semua pihak yang telah memberikan bantuan dalam penyusunan tugas akhir ini. Penulis juga menyadari bahwa tugas akhir ini masih belum sempurna, sehingga diharapkan kritik dan saran yang membangun.

UCAPAN TERIMA KASIH

Dengan mengucap syukur ke hadirat Tuhan Yang Maha Esa, penulis menyampaikan terima kasih yang sebesar-besarnya kepada semua pihak yang telah memberikan dukungan, bantuan, serta doa selama proses penyusunan tugas akhir ini hingga selesai. Karena itu, kami ingin menyampaikan ucapan terima kasih kepada:

1) Allah SWT, atas Rahmat dan hidayah-Nya.
2) Dr. Tenia Wahyuningrum, S.Kom., M.T. selaku Direktur Universitas Telkom Purwokerto.
3) Siti Khomsah, S.Kom, M.Cs. selaku Kaprodi S1 Sains Data Universitas Telkom Purwokerto.
4) Dr. Ridwan Pandiya, S.Si., M.Sc. selaku Dosen Pembimbing atas bimbingan dan saran yang telah diberikan.
5) Kedua orang tua tercinta dan keluarga, atas doa, kasih sayang dan dukungan moral yang tak pernah henti.
6) Kepada sahabat terdekat saya Rizqi Fajar yang telah membantu saya menyelesaikan tugas akhir ini.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1. Latar Belakang

Kesehatan merupakan salah satu hal yang sangat penting dalam kehidupan manusia sehingga perlu dijaga dengan baik karena dengan tubuh yang sehat maka aktivitas dapat berjalan dengan baik [1]. Berbagai upaya yang dilakukan manusia dalam menjaga kesehatan seperti mengatur pola makan, rutin berolahraga . Namun, berbagai penyakit dapat muncul sehingga menghambat aktivitas manusia. Salah satu penyakit yang banyak diderita adalah batu ginjal yaitu sebanyak 1.499.400 manusia di Indonesia yang menderita penyakit batu ginjal [2].

Batu ginjal merupakan materi keras seperti batu yang terbentuk dari urin yang berkonsentrasi [3] dengan garam dan mineral [2]. Batu ginjal dapat menyebabkan infeksi berulang, gangguan ginjal dan juga hematuria [4] yang apabila tidak ditangani sedini mungkin dapat menyebabkan kerusakan ginjal seperti gagal ginjal bahkan dapat menyebabkan kanker ginjal [5]. Gejala yang dialami oleh penderita batu ginjal adalah rasa sangat nyeri pada bagian pinggul, muntah secara terus menerus dan berdarah saat berkemih yang disebabkan oleh batu ginjal yang bergerak di dalam saluran ureter melalui saluran urin [3]. Penyakit ini dapat diatasi dengan baik apabila dapat dideteksi sedini mungkin untuk mencegah komplikasi seperti infeksi dan hematuria. Deteksi dini dapat dilakukan dengan memanfaatkan teknik *data mining* untuk mengolah data klinis dengan menemukan pola tersembunyi dan informasi yang relevan dengan diagnosis penyakit batu ginjal [6].

Data mining atau penambangan data merupakan suatu teknik yang digunakan untuk mencari informasi dari suatu kumpulan data yang besar [6]. Data mining akan melakukan sebuah proses ekstrasi dengan memanfaatkan teknik kecerdasan buatan, statistika dan juga teknik matematika dan *machine learning*. Setelah proses ekstrasi data dilakukan, informasi akan didapatkan dari proses data mining tersebut, informasi ini akan membantu dalam prediksi dalam banyak bidang. *Data mining* memiliki fungsi sebagai *clustering, association,* dan *classification* [7].

Classification atau klasifikasi merupakan salah satu teknik dalam *data mining* yang merupakan proses analisis data input dengan model klasifikasi berdasarkan fitur dalam dataset. Klasifikasi memungkinkan memberikan label pada data baru yang belum diketahui labelnya dengan memahami karakteristik kelas [1][8]. Beberapa model klasifikasi pada *data mining* diantaranya adalah *support vector machine, decision tree, random forest, naïve bayes, neural network* [9].

Metode klasifikasi yang umum digunakan untuk klasifikasi penyakit adalah *decision tree* atau pohon keputusan yang merupakan salah satu metode untuk membuat pohon keputusan berdasarkan data pelatihan untuk pengklasifikasian dan prediksi yang mudah dipahami aturannya [1][8]. Dalam pohon keputusan, pembagian himpunan data besar akan dilakukan menjadi himpunan terkecil dalam proses klasifikasi atau prediksi [8]. Pembagian himpunan data besar didasarkan pada nilai *gain* terbesar yang dihasilkan variabel. Nilai *gain* yang terbesar akan digunakan model pohon keputusan sebagai akar pohon [7].

Salah satu algoritma dalam pohon keputusan yang sangat popular digunakan adalah algoritma C4.5 [8]. Algoritma C4.5 memiliki sampel pelatihan yang akan digunakan untuk membangun sebuah pohon dan kebenarannya telah diuji [9]. Untuk membangun sebuah algoritma C4.5 diperlukan atribut sebagai akar. Beberapa kelebihan algoritma ini adalah pohon keputusan yang mudah untuk diinterpretasikan, efisien dalam penanganan data yang bertipe numerik dan diskrit. Kekurangan algoritma pohon keputusan adalah rentan terhadap *overfitting* apabila data memiliki fitur yang terlalu banyak [10] dapat ditangani dengan melakukan pemangkasan pohon keputusan atau disebut dengan *pruning* di algoritma C4.5.

Pada proses klasifikasi, sebuah model yang dilatih menggunakan data *training* tertentu dapat menunjukkan performa yang sangat baik secara kebetulan, namuna model tersebut belum tentu mampu bekerja dengan baik pada data yang belum pernah dilihat sebelumnya. Oleh karena itu, diperlukan proses validasi model untuk memastikan bahwa performa yang dihasilkan model benar-benar menunjukkan kemampuan model secara umum. Beberapa teknik validasi model yang umum digunakan dalam pembelajaran mesin adalah *cross-validation, hold-out validation, dan bootstrapping*. Salah satu teknik yang paling populer adalah

cross-validation yang bekerja dengan membagi data ke dalam beberapa subset untuk kemudian digunakan secara bergiliran sebagai data latih dan data uji. Teknik ini membantu memastikan bahwa hasil yang diperoleh benar-benar mewakili pola dari data yang sesungguhnya, serta dapat meminimalkan risiko kesalahan interpretasi atau pengambilan kesimpulan yang tidak akurat terhadap data.

Penelitian sebelumnya [1] menggunakan algoritma *Decision Tree* dan menerapkan metode C4.5 dengan membandingkan hasil akurasi, namun belum menerapkan teknik evaluasi yang bertujuan untuk mengidentifikasi adanya *overfitting* atau *underfitting*. Teknik seperti *cross-validation* belum digunakan untuk memastikan bahwa model yang dihasilkan memiliki kemampuan generalisasi yang baik terhadap data baru. Selain itu, penelitian tersebut menggunakan dataset dari RSUD Tangerang dengan 18 variabel, yang memiliki karakteristik dan struktur data berbeda dibandingkan dengan dataset yang digunakan dalam penelitian ini. Perbedaan tersebut mencakup jumlah variabel, sumber data, serta definisi dari masing-masing atribut yang digunakan sebagai basis klasifikasi. Penelitian ini menggunakan data dari RSUD Cideres dengan 7 variabel utama.

1.2. Rumusan Masalah

Penerapan algoritma *Decision Tree* C4.5 pada klasifikasi penyakit batu ginjal memiliki akurasi yang tinggi, namun penerapan metode ini pada sumber data yang berbeda dapat menimbulkan sebuah perbedaan penanganan data karena adanya karakteristik yang berbeda antara dataset dan perbedaan distribusi variabel. Penelitian ini berupaya untuk menganalisis efektivitas metode *Decision Tree* C4.5 pada dataset yang berasal dari sumber yang berbeda.

1.3. Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Menerapkan algoritma *Decision Tree* C4.5 untuk mengklasifikasikan penyakit batu ginjal berdasarkan data klinis.
2. Mengukur akurasi algoritma *Decision tree* C4.5 dalam melakukan prediksi pada penyakit batu ginjal.
3. Menemukan variabel yang memiliki pengaruh paling signifikan.

1.4. Manfaat Penelitian

Menerapkan algoritma *Decision Tree* C4.5 dalam klasifikasi penyakit batu ginjal diharapkan mampu menghasilkan sebuah diagnosis dengan melakukan prediksi berdasarkan data prediktor yang ada dalam dataset. Hal ini diharapkan mampu mempermudah dalam mendeteksi dini penyakit batu ginjal dalam mengambil diagnosis.

BAB II
KAJIAN PUSTAKA

Pada penelitian terdahulu, algoritma *Decision Tree* telah banyak diterapkan pada berbagai masalah terutama pada kasus klasifikasi penyakit untuk mendiagnosis penyakit. Beberapa penelitian menunjukkan bahwa penggunaan algoritma *Decision Tree* terutama pada algoritma *Decision Tree* C4.5 yang menghasilkan akurasi yang lebih baik dibandingkan dengan akurasi model lainnya. Namun, penelitian mengenai penyakit batu ginjal belum banyak diterapkan pada data yang lain sehingga pada penelitian ini data yang akan digunakan adalah data penyakit batu ginjal dari RSUD Cideres Majalengka.

2.1. Kajian Penelitian

Tujuan penelitian ini adalah mengklasifikasikan penyakit batu ginjal berdasarkan dataset klinis, diawali dengan mengkaji hasil-hasil riset terkait klasifikasi penyakit batu ginjal. Berdasarkan penelitian [1] menggunakan dataset dari RSUD di Tangerang dengan 18 variabel yang digunakan. Hasil penelitian menunjukkan algoritma *Decision Tree* menghasilkan akurasi sebesar 96,7%.

Penelitian [11] pada konteks penyakit yang sama menggunakan data penyakit batu ginjal dari situs Kaggle yang terdiri dari 8 variabel diantaranya adalah ID, *Gravity, pH, Urea, Calcium* dan sebagainya. Penelitian [11] menggunakan algoritma *Decision Tree* yang menghasilkan akurasi sebesar 72%.

Penelitian [12] yang memprediksi penyakit ginjal menggunakan algoritma *Decision Tree*. Penelitian ini menggunakan data yang berasal dari *UCI Machine Learning* yang berjumlah 337 data dan penelitian ini menggunakan 5 jenis atribut. Hasil penelitian ini menjukkan bahwa algoritma *Decision Tree* mampu melakukan klasifikasi dengan baik, ditunjukkan dengan akurasi yang diperolah pada penelitian ini adalah sebesar 89,05%, *Recall* sebesar 100%, *Precision* sebesar 77,96%.

Algoritma *Decision Tree* juga digunakan pada penelitian [13] untuk memprediksi penyakit *cerebrovascular* dengan membandingkan dua algoritma *Machine Learning*, yaitu *Decision Tree* dan *Naïve Bayes*. Dataset yang digunakan dalam penelitian ini adalah dataset *Stroke Prediction Dataset* dari situs Kaggle yang berjumlah 5110 baris dan 12 kolom. Hasil dari penelitian ini menunjukkan akurasi dalam algoritma *Naïve Bayes* sebesar 91% sedangkan algoritma *Decision Tree* menunjukkan akurasi yang lebih tinggi yaitu sebesar 95,3%. Berdasarkan hasil penelitian, algoritma *Decision Tree* mampu melakukan klasifikasi dengan baik menggunakan data penyakit *cerebrovascular*.

Penelitian [14] menggunakan algoritma *Decision Tree* untuk memprediksi penyakit *stroke*. Data yang digunakan dalam penelitian ini adalah dataset yang berasal dari situs Kaggle yang berjumlah 4906 baris setelah dilakukan pemrosesan data. Algoritma *Decision Tree* menunjukkan akurasi yang tinggi yaitu sebesar 96,05%. Hasil menunjukkan bahwa algoritma *Decision Tree* mampu melakukan klasifikasi dengan baik menggunakan data penyakit *stroke*.

Penelitian ini didukung penelitian [15] yang menggunakan algoritma *Decision Tree* untuk konteks yang sama yaitu untuk klasifikasi untuk memprediksi penyakit diabetes. Penelitian ini menggunakan dataset yang berasal dari Kaggle dengan total 2000 baris yang memiliki variabel *Pregnancies, Glucose, Blood, Pressure, BMI, Age* dan *Outcome* yang semuanya bertipe data numerik. Hasil dari penelitian ini menunjukkan akurasi model *Decision Tree* sebesar 96% yang menunjukkan bahwa algoritma dapat mengklasifikasikan data dengan baik.

Penelitian [16] menggunakan perbandingan algoritma *Decision Tree* dan *Naïve Bayes* untuk memprediksi penyakit Diabetes. Dataset yang digunakan berasal dari Rumah Sakit Sylhet, Bangladesh. Hasil akurasi dalam penelitian ini menujukkan algoritma *Decision Tree* menghasilkan akurasi 96,36% yang lebih unggul dibandingkan dengan algoritma *Naïve Bayes* yaitu 90,45%. Penelitian ini menunjukkan bahwa algoritma *Decision Tree* lebih baik dalam melakukan klasifikasi penyakit.

Penelitian [17] menggunakan algoritma *Decision Tree* C4.5 untuk memprediksi penyakit Diabetes dengan menggunakan data yang berasal dari *UCI Machine Learning*. Hasil penelitian ini menunjukkan bahwa root node yang ditemukan adalah "sering buang air kecil" dan akurasi yang didapatkan adalah sebesar 95,51%. Penelitian ini menunjukkan bahwa algoritma *Decision Tree* dapat digunakan untuk memprediksi penyakit dengan baik.

Penerapan *Decision Tree* [18] juga digunakan pada diagnosis penyakit jantung dengan mengklasifikasikan data klinis penyakit jantung. Dataset yang digunakan berasal dari Kaggle yang berisi 12 atribut. Hasil akurasi yang didapatkan dalam penelitian ini adalah 80,43% dengan eror klasifikasi sebesar 19,57%, *recall true* normal sebesar 80,49, *recall true* penyakit jantung sebesar 80,395, *precision* normal 75,74 dan *precision* penyakit jantung 83,67%. Penelitian ini menunjukkan bahwa algoritma *Decision Tree* baik diterapkan pada klasifikasi penyakit jantung.

Penelitian [19] yang dilakukan dengan menggunakan data penyakit jantung menggunakan *Decision Tree* C4.5. Data yang digunakan merupakan data klinis dengan 12 atribut yaitu usia, jenis kelamin, jenis nyeri dada, tekanan darah saat istirahat, kadar kolesterol, gula darah puasa, hasil elektrokardiogram saat istirahat, denyut jantung maksimum, angina saat berolahraga, penurunan *oldpeak*, kemiringan segmen ST. Hasil akurasi yang didapatkan dalam penelitian ini adalah sebesar 84%. Dalam penelitian ini, *Decision Tree* dapat dikatakan baik dalam melakukan klasifikasi penyakit jantung.

Penelitian [20] menggunakan algoritma *Decision Tree* C4.5 untuk memprediksi penyakit diabetes. Data yang digunakan dalam penelitian ini adalah data klinis penyakit diabsetes yang berasal dari Kaggle.com. Hasil penelitian ini menunjukkan bahwa akurasi yang didapatkan sebesar 85%, *precision* sebesar 92% dan *recall* sebesar 85%. Penelitian ini menunjukkan bahwa *Decision Tree* diterapkan dengan baik untuk memprediksi penyakit diabetes.

Ragkuman kajian dari penelitian-penelitian terdahulu ini disajikan dalam bentuk Tabel 2.1.

Tabel 2.1 Ringkasan Penelitian Terdahulu

Latar Belakang Penelitian			Desain Riset dan Metodologi			
Referensi	Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan	GAP Penelitian
[1]	Penyakit batu ginjal yang dapat mengakibatkan masalah yang serius apabila tidak ditangani secara dini.	Mencari atribut yang paling signifikan dan memprediksi penyakit batu ginjal.	Algoritma *Decision Tree* C4.5	Data penyakit batu ginjal 2016-2019 RSUD di Tangerang.	Akurasi tertinggi yang didapatkan adalah 95,71%.	Perbedaan sumber dataset yang digunakan dan metode evaluasi yang digunakan.
[11]	Penyakit batu ginjal yang berbahaya apabila tidak ditangani secara dini sehingga perlu adanya pediksi penyakit batu ginjal.	Memprediksi penyakit batu ginjal untuk dapat melakukan diagnosis awal.	Algoritma *Decision Tree* C4.5	Data penyakit batu ginjal berasal dari Kaggle.	Akurasi *Decision Tree* yang didapatkan sebesar 72%.	Perbedaan sumber dataset yang digunakan dan metode evaluasi yang digunakan.
[12]	Penyakit ginjal yang sering terlambat terdeteksi menyebabkan angka kematian yang tinggi.	Mendeteksi dini dan mendiagnosis penyakit ginjal dengan hasil evaluasi algoritma yang digunakan	Algoritma *Decision Tree* C4.5	Data penyakit ginjal dari UCI Machine Learning	Hasil akurasi 89,05% dengan recall 100%, precision 77,96%.	Perbedaan sumber dataset yang digunakan.
[13]	Penyakit *cerebrovascular* perlu deteksi dini penyakit	Mendeteksi dini dan memprediksi penyakit *cerebrovascular*.	*Decision Tree* C4.5	Data penyakit *cerebrovascular*	Akurasi *Decision Tree* sebesar 95,3%, sedangkan	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.

Latar Belakang Penelitian			Desain Riset dan Metodologi			
Referensi	Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan	GAP Penelitian
	cerebrovascular untuk mengurangi angka kematian.		dan Naïve Bayes.	berasal dari Kaggle.	akurasi *Naïve Bayes* sebesar 91,3%.	
[14]	Serangan stroke termasuk 10 penyakit berbahaya di dunia dan perlu deteksi dini.	Mendeteksi dini dan memprediksi penyakit stroke.	Algoritma *Decision tree* C4.5	Data penyakit stroke dari Kaggle.	Hasil akurasi yang didapatkan sebesar 96,05%.	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.
[15]	Penyakit diabetes merupakan penyakit berbahaya ditandai sebesar 21,3 juta manusia yang terkena penyakit diabetes menurut WHO.	Mendeteksi dini dengan memprediksi penyakit diabetes mellitus.	Algoritma *Decision Tree* C4.5	Data penyakit diabetes mellitus dari Kaggle.	Hasil akurasi yang didapatkan sebesar 96% dengan *precision* sebesar 99%, *recall* 95% dan f1-*score* sebesar 97%.	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.
[16]	Penyakit diabetes merupakan penyakit berbahaya ditandai dengan adanya 463 juta manusia yang terkena penyakit diabetes.	Mendeteksi dini dengan memprediksi penyakit diabetes.	*Decision Tree* C4.5 dan *Naïve Bayes*	Data penyakit diabetes berasal dari UCI Machine Learning	Hasil akurasi *Decision Tree* C4.5 sebesar 96,36 dan hasil akurasi *Naïve Bayes* sebesar 90,45%.	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.
[17]	Diabetes merupakan penyakit paling berbahaya ke-9 di dunia menurut WHO	Mendeteksi dini penyakit diabetes dengan melakukan	Algoritma *Decision Tree* C4.5	Data penyakit diabetes berasal dari UCI *Machine Learning*	Hasil akurasi algoritma *Decision Tree* C4.5 adalah sebesar 95,51%.	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.

Latar Belakang Penelitian			Desain Riset dan Metodologi			
Referensi	Masalah Penelitian	Tujuan	Metode	Data	Kesimpulan	GAP Penelitian
	sehingga diperlukan adanya deteksi dini penyakit diabetes	prediksi penyakit diabetes.				
[18]	Tingginya angka kematian akibat penyakit jantung serta perlunya prediksi data yang akurat untuk memahami mendiagnosis penyakit jantung.	Memprediksi data penyebab penyakit jantung menggunakan algoritma *decision tree* C4.5.	Algoritma *Decision Tree* C4.5	Data penyakit jantung yang berasal dari Kaggle.	Tingkat akurasi model mencapai 80,43%, dengan *error classification* sebesar 19,57%.	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.
[19]	Angka kematian akibat penyakit jantung yang tinggi sehingga perlu deteksi dini.	Memprediksi data penyebab penyakit jantung	Algoritma *Decision Tree* C4.5	Data penyakit jantung dari situs Kaggle.	Menghasilkan akurasi sebesar 85%	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.
[20]	Tingginya angka kasus diabetes dan kematian akibat penyakit	Mendeteksi dini penyakit diabetes dengan melakukan prediksi	Algoritma *Decision Tree* C4.5	Data penyakit jantung yang berasal dari Kaggle.com	Menghasilkan akurasi 85%, *precision* 92%, dan *recall* 85%.	Penerapan *Decision Tree* C4.5 pada domain yang berbeda.

2.2. Dasar Teori

Beberapa teori yang melandasi penelitian ini sebagai berikut.

2.2.1 Penyakit Batu Ginjal

Batu ginjal merupakan materi keras seperti batu yang terbentuk dari urin yang berkonsentrasi [3] dengan garam dan mineral [2]. Penyebab penyakit ini adalah karena dehidrasi, infeksi saluran kemih, gangguan metabolik dalam tubuh dan juga konsumsi garam yang berlebih [21].

Selain disebabkan oleh gaya hidup yang tidak sehat, penyakit batu ginjal juga dapat disebabkan oleh faktor keturunan [1]. Gejala yang dialami oleh penderita batu ginjal adalah rasa sangat nyeri pada bagian pinggul, muntah secara terus menerus dan berdarah saat berkemih yang disebabkan oleh batu ginjal yang bergerak di dalam saluran ureter melalui saluran urin [3]. Batu ginjal dapat menyebabkan infeksi berulang, gangguan ginjal dan juga hematuria atau darah dalam urin [4] yang apabila tidak ditangani sedini mungkin dapat menyebabkan kerusakan ginjal seperti gagal ginjal bahkan dapat menyebabkan kanker ginjal [5].

2.2.2 *Data Mining*

Data mining atau penambangan data merupakan suatu teknik yang digunakan untuk mencari informasi dari suatu kumpulan data yang besar [6]. *Data mining* akan melakukan sebuah proses ekstrasi dengan memanfaatkan teknik kecerdasan buatan, statistika, teknik matematika dan *machine learning*. Setelah proses ekstrasi data dilakukan, informasi akan didapatkan dari proses *data mining* tersebut, informasi ini akan membantu dalam prediksi dalam banyak bidang. *Data mining* memiliki fungsi sebagai *clustering, association*, dan *classification* [7].

Pemanfaatan *data mining* memiliki banyak keuntungan yaitu dapat membantu dalam identifikasi pola dan membantu dalam pengambilan keputusan, menemukan faktor penyebab masalah yang mungkin belum dapat diidentifikasi dengan metode konvensional. Namun, *data mining* juga memiliki keterbatasan yaitu memiliki ketergantungan dalam kualitas data yang digunakan, apabila kualitas data kurang baik maka hasil yang dihasilkan dalam proses *data mining* akan tidak akurat bahkan salah sehingga data perlu penanganan seperti menghapus nilai *null*, menangani *outlier* dan menangani *imbalance class* terlebih dahulu [7].

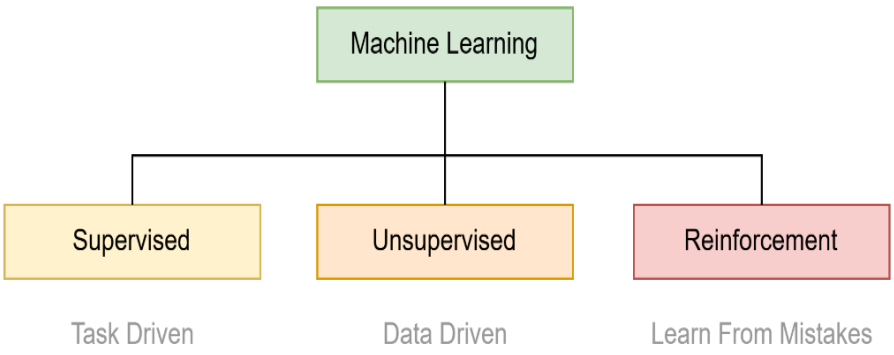
Data mining banyak diterapkan dalam berbagai bidang, seperti bidang kesehatan. *Data mining* dapat digunakan dalam menganalisis data pasien dan memprediksi risiko penyakit, menemukan pola penyakit dan memprediksi hasil penyakit [7].

Dalam proses *data mining*, diperlukan sebuah model prediktif maupun model deskriptif sebagai alat bantu *data mining*. Salah satu alat bantu yang digunakan adalah *machine learning* yang dapat belajar dari data lama[7].

2.2.3 *Machine Learning*

Machine Learning merupakan salah satu teknik dalam *data mining* yang berfokus pada pengembangan algoritma dan model yang memungkinkan komputer untuk mempelajari pola data dan dapat dilakukan pengambilan keputusan tanpa pemrograman berulang kali. *Machine Learning* membutuhkan sebuah data pelatihan dan akan diproses untuk menghasilkan sebuah *output* [22].

Fokus utama dalam pembelajaran mesin adalah mengolah data dengan berbagai tipe menjadi suatu keputusan yang bermakna dengan sedikit campur tangan manusia melalui sebuah pemrograman [23]. Sebuah program atau sebuah kasus dapat dikatakan masuk dalam kategori *Machine Learning* apabila program komputer tersebut telah mempelajari sebuah *experience E* terhadap tugas (*task T*) yang ada dan kemudian akan diukur untuk hasil kinerjanya dengan *Performance Measure P* [24]. *Machine Learning* memiliki tiga sub bagian diantaranya adalah *Supervised Learning*, *Unsupervised Learning* dan *Reinforcement* yang ditunjukkan pada Gambar 2.1 di bawah ini:



Gambar 2. 1 Tipe Pembelajaran Mesin

Unsupervised Learning adalah pembelajaran yang tidak diawasi atau pembelajaran dengan aturan asosiasi. Tujuan dari *Unsupervised Learning* adalah menggali pola atau informasi tersembunyi dari dataset berukuran besar, sehingga hubungan antar atribut dapat diidentifikasi tanpa adanya label target. Sub bagian kedua adalah *Reinforcement Learning*, yang dalam prosesnya tidak menggunakan data berlabel sebagai input, melainkan melibatkan agen yang mengamati lingkungan, memilih tindakan, dan mengeksekusi tindakan tersebut secara mandiri untuk memaksimalkan suatu nilai penghargaan (*reward*) [25].

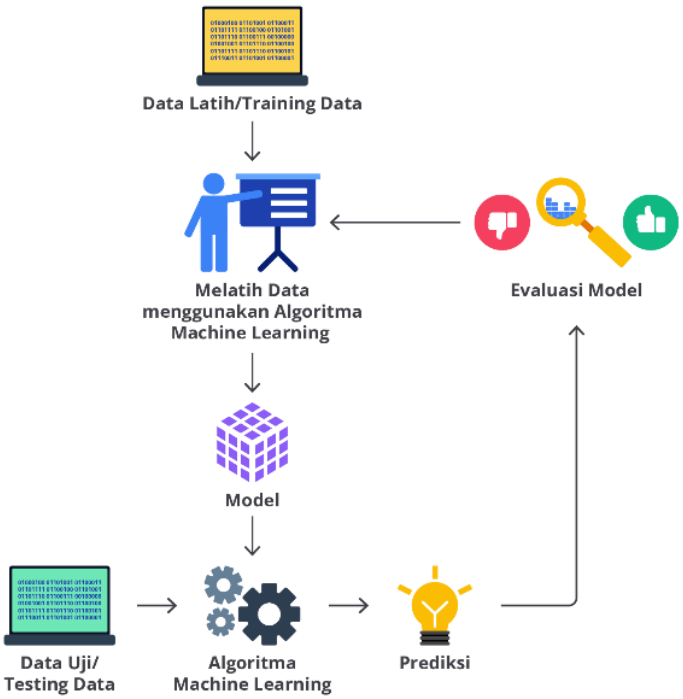
Supervised Learning adalah sub bagian terakhir yang merupakan pembelajaran yang terarah dalam pembelajaran mesin yang paling umum digunakan apabila data telah diannotasi. Tujuan dari pembelajaran terarah (*supervised learning*) adalah menemukan pola dalam data untuk diimplementasikan ke dalam model klasifikasi maupun regresi [26]. Salah satu bentuk paling umum dari *supervised learning* adalah klasifikasi, yang berperan penting dalam mengelompokkan data berdasarkan label atau kategori tertentu.

Klasifikasi merupakan proses menganalisis data masukan (*input*) menggunakan model klasifikasi berdasarkan fitur yang terdapat dalam dataset. Klasifikasi memungkinkan pemberian label pada data baru yang belum diketahui labelnya, dengan memahami karakteristik masing-masing kelas [1][8]. Selain itu, klasifikasi mampu menangani rentang data yang luas, bahkan dalam beberapa kasus lebih luas dibandingkan dengan yang dapat ditangani oleh regresi [7]

Dalam klasifikasi, pembangunan model klasifikasi berfungsi untuk menyelesaikan masalah atribut dalam data yang digunakan. Tahap pembangunan model merupakan fase pelatihan klasifikasi, dengan data latih akan dilakukan analisis. Setelah data latih dianalisis, model akan diterapkan untuk mengetahui seberapa akurat klasifikasi yang dilakukan dengan menggunakan data uji. Apabila model menghasilkan akurasi yang baik, maka aturan klasifikasi dapat diterapkan untuk prediksi data baru yang belum diketahui kelasnya [7].

Alur kerja dalam model klasifikasi seperti Gambar 2.2 yaitu dengan *input* data latih yang telah berlabel, melatih data menggunakan *machine learning*,

menentukan model dan melakukan prediksi. Setelah dilakukan prediksi, model akan dievaluasi untuk melihat keakuratan model yang diterapkan.



Gambar 2. 2 Alur Kerja Klasifikasi

(Sumber : Dicoding.com)

2.2.4 Analisis Eksplorasi Data

Eksploratory Data Analysis (EDA) adalah sebuah proses awal dalam sebuah proses *data science* untuk menemukan pola tersembunyi, mendeteksi adanya kelas tidak seimbang, mendeteksi adanya anomali, melihat *outlier*, melihat *missing value* dan melihat distribusi data dalam suatu set data. Tujuan dari adanya sebuah *Eksploratory Data Analysis* adalah untuk memahami sebuah set data yang akan digunakan [26].

Analisis eksplorasi data yang pertama dilakukan adalah melihat korelasi antar fitur yang digunakan untuk penelitian. Korelasi dengan koefisien mendekati 1 atau -1 memiliki hubungan yang kuat. Korelasi dapat dihitung menggunakan Persamaan 2.1 di bawah ini:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\{n\sum x^2 - (\sum x)^2\{n\sum y^2 - (\sum y)^2}} \quad (2.1)$$

Dimana:

n = Jumlah x dan y

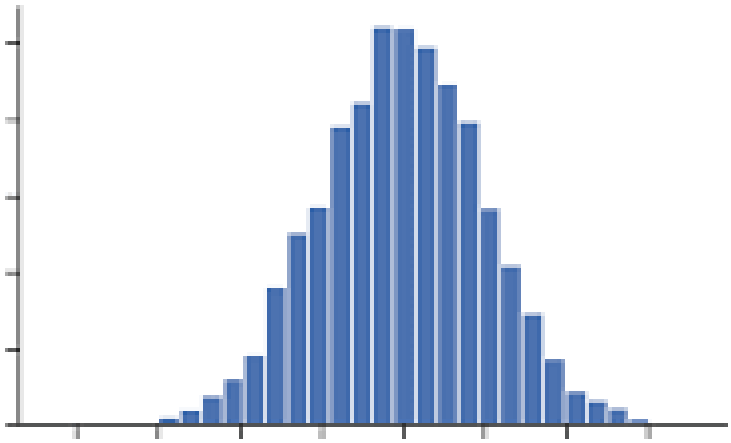
$\sum x$ = Jumlah variabel x

$\sum y$ = Jumlah variabel y

$\sum x^2$ = Jumlah variabel x yang dikuadratkan

$\sum y^2$ = Jumlah variabel y yang dikuadratkan

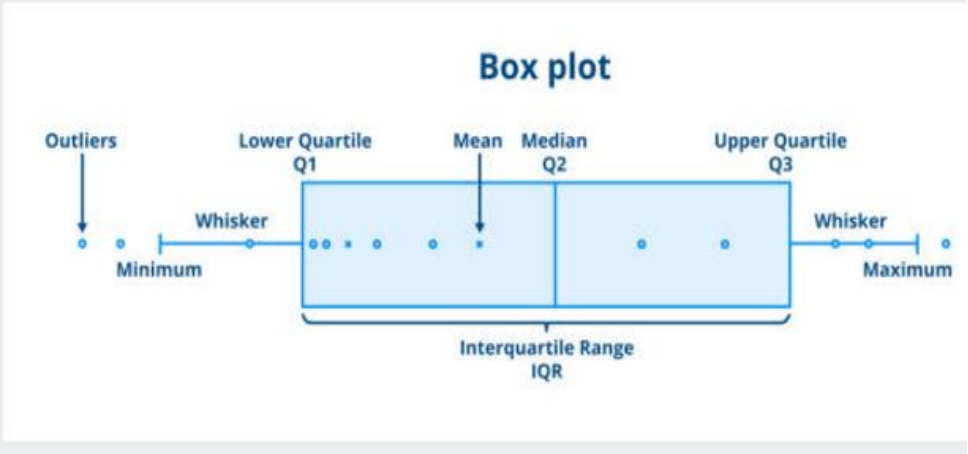
Dalam melihat distribusi data yang ada pada suatu *dataset*, terdapat dua kemungkinan distribusi dalam data tersebut. Distribusi yang pertama adalah distribusi normal yang apabila data berada simetris di sekitar nilai tengahnya dan memuncak di tengah [27]. Distribusi normal dalam dataset dapat dilihat seperti Gambar 2.3 di bawah ini:



Gambar 2. 3 Distribusi Normal

(Sumber : Gramedia.com)

Apabila asumsi distribusi normal tidak dipenuhi, maka diperlukan adanya deteksi *outlier*. Nilai yang dikatakan *outlier* adalah nilai yang lebih besar atau lebih kecil dari nilai maksimum atau minimum, sehingga dapat menyebabkan kesalahan dalam pengambilan kesimpulan [27]. Pada Gambar 2.4 di bawah ini menunjukkan data yang terdapat *outlier*.



Gambar 2. 4 Deteksi *Outlier Box Plot*

(Sumber : Parameterd.com)

Setelah karakteristik data telah diketahui, maka dalam proses *Machine Learning*, data akan diimplementasikan algoritma yang sesuai dengan karakteristiknya untuk meminimalkan potensi adanya kesalahan [26].

2.2.5 *Decision Tree* (Pohon Keputusan)

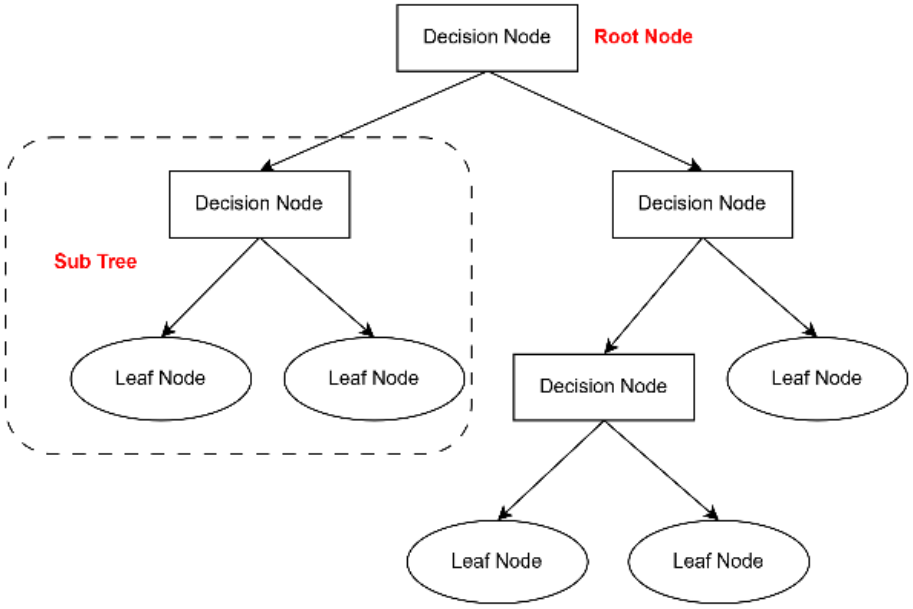
Algoritma *Decision Tree* adalah algoritma *Machine Learning* yang menggunakan struktur pohon dengan menggunakan *root node. Root node* digunakan untuk melihat atribut yang memiliki nilai paling penting dengan menghitung *gain* dan *entropy* [13]. Beberapa metode dalam *decision tree* adalah C4.5, CART, ID3.

Decision Tree adalah C4.5 yang merupakan salah satu algoritma yang bisa digunakan untuk klasifikasi data, baik data berbentuk kategorik maupun data numerik. Algoritma ini merupakan pengambangan dari algoritma CART. Algoritma C4.5 membangun pohon keputusan berdasarkan data pelatihan dengan

efektivitas dan keakuratannya telah dibuktikan melalui berbagai penelitian sebelumnya [9].

Untuk membangun sebuah algoritma C4.5 diperlukan atribut sebagai akar. Beberapa kelebihan algoritma ini adalah pohon keputusan yang mudah untuk diinterpretasikan, efisien dalam penanganan data yang bertipe numerik dan diskrit [10], dapat menangani nilai data yang hilang [13].

Kekurangan algoritma pohon keputusan adalah rentan terhadap *overfitting* apabila data memiliki fitur yang terlalu banyak [10] dapat ditangani dengan melakukan *pruning* atau pemangkasan di algoritma C4.5. Gambar 2.5 di bawah ini adalah struktur dari algoritma *Decision Tree*.

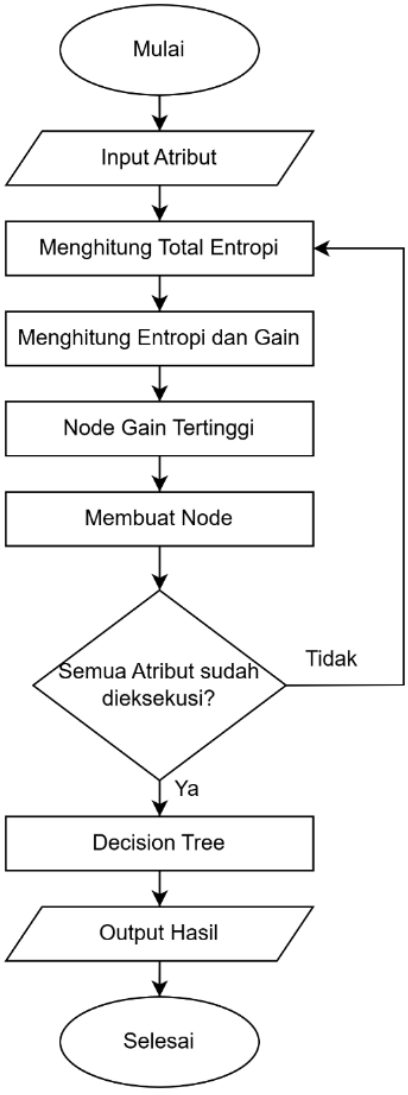


Gambar 2. 5 Struktur *Decision Tree* C4.5

Gambar 2.5 menunjukan bahwa terdapat *root node* yang dihitung dengan mencari nilai *gain* dan *entropy* yang kemudian akan diseleksi. Nilai *gain* yang paling tinggi akan digunakan sebagai akar dalam pohon yang berperan penting dalam proses klasifikasi. Setelah atribut yang menjadi akar telah diketahui, kemudian buat cabang untuk setiap nilai dan ulangi proses menghitung *entropy* dan

gain sampai semua kasus pada cabang mencapai batas kedalaman pohon yang telah ditentukan. Cabang dalam algoritma pohon keputusan disebut dengan *sub-tree*. Dalam setiap node pohon keputusan akan ditampilkan nilai dari ambang batas, nilai *entropy*, jumlah *sample* dan jumlah data terbagi dalam kelas 0 dan dalam kelas 1.

Proses dalam algoritma *Decision tree* C4.5 ditunjukkan dengan diagram alir proses kerja *Decision Tree* seperti Gambar 2.6 di bawah ini:



Gambar 2. 6 Proses Kerja *Decision tree* C4.5

1. Menentukan atribut yang akan digunakan untuk penelitian, dalam penelitian ini atribut yang digunakan adalah asupan cairan, asupan sodium, asupan protein, kadar asam urat urin, pH urin, dan fungsi ginjal yang direkomendasikan oleh tim dokter RSUD Cideres Majalengka.
2. Langkah kedua adalah menghitung total entropi dan untuk menghitung nilai *gain*, digunakan Persamaan 2.2 sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2(pi) \quad (2.2)$$

Keterangan rumus :

S : Himpunan kasus yang ada.

n : Jumlah kelas dalam dataset.

pi : Perbandingan jumlah data pada kelas ke_i terhadap total data (S)

3. Menghitung nilai *entropy* dan *gain* dari masing-masing atribut yang digunakan dalam penelitian. Semua atribut akan dihitung nilai entropi dan *gain* untuk melihat nilai masing-masing atribut dan dilakukan pencarian nilai terbesar *gain*.

   Nilai proporsi data (*pi*) terhadap seluruh dataet *S* dapat dihitung dengan Persamaan 2.3 di bawah ini:

$$p_i = \frac{\text{Jumlah data kelas } ke_i}{\text{Jumlah data dalam } S} \quad (2.3)$$

Setelah nilai *entropy* diketahui, selanjutnya dihitung nilai *gain* dengan menggunakan Persamaan 2.4 di bawah ini:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad (2.4)$$

Keterangan rumus:

S : Himpunan kasus yang ada.

A : Atribut A.

n : Jumlah kelas dalam dataset

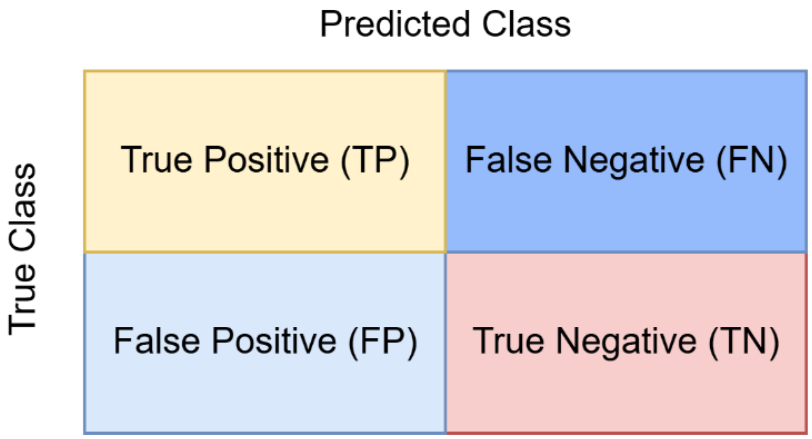
S_i : Jumlah kasus ke_i pada partisi

4. Nilai terbesar *gain* ditentukan dan dijadikan sebagai akar dari algoritma *Decision Tree C4.5*.
5. Setelah *root node* diketahui, dilanjutkan dengan menghitung nilai *entropy* dan *gain* dari atribut selanjutnya dan tentukan nilai *gain* tertinggi untuk membentuk *sub tree.*
6. Setelah semua atribut selesai dihitung nilai *entropy* dan *gain*, tahap selanjutnya adalah membangun model *Decision Tree C4.5* untuk memprediksi penyakit batu ginjal. Setelah implementasi, data baru tanpa label yang dimasukkan dalam model akan secara langsung diprediksi apakah termasuk dalam kelas "Batu Ginjal" atau "Tidak Batu Ginjal".

2.2.6 *Confusion Matrix*

Dalam *Machine Learning*, terdapat beberapa metrik yang digunakan untuk mengukur keakuratan model yang digunakan. Beberapa metrik pengukuran tersebut adalah Akurasi (*Accuracy)*, Matriks Konfusi (*Confusion Matrix)* dan *Classification Report.*

Confusion Matrix adalah cara untuk mengukur seberapa akurat model *Machine Learning* yang digunakan. *Confusion Matrix* digunakan untuk memecahkan permasalahan dalam klasifikasi biner maupun klasifikasi multikelas [28]. Dalam klasifikasi biner, matriks konfusi memiliki empat komponen perhitungan seperti Gambar 2.7 di bawah ini:



Gambar 2. 7 Komponen Matriks Konfusi

Dimana:

a. *True Positive* (*TP*) merupakan nilai dalam kelas positif yang diklasifikasikan dan diprediksi dengan benar oleh model.
b. *True Negative* (*TN*) merupakan nilai dalam kelas negatif yang diklasifikasikan dan diprediksi dengan benar oleh model.
c. *False Positive* (*FP*) merupakan nilai yang diprediksi dalam kelas positif, namun seharusnya bagian dari kelas negative.
d. *False Negative* (*FN*) merupakan nilai yang dipresiksi dalam kelas negative, namun seharusnya bagian dari kelas positif.

Model klasifikasi dapat dikatakan baik apabila nilai dari TN dan TP lebih besar dari FN dan FP. Dengan menggunakan Matriks Konfusi maka akan terlihat jumlah nilai mana yang diprediksi benar dan salah.

Dalam klasifikasi, metrik evaluasi *Accuracy* juga digunakan untuk mengukur keakuratan model yang digunakan. Untuk menghitung nilai akurasi digunakan Persamaan 2.5 seperti di bawah ini:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.5)$$

Metrik *Classification Report* juga digunakan dalam mengukur seberapa baik model yang digunakan. Dalam metrik ini terdapat empat komponen diantaranya adalah *Precision, Recall, F1-Score* dan *Support* [17].

a. *Precision* menunjukkan proporsi jumlah benar yang dinyatakan positif dalam model. *Precision* kelas positif dihitung dengan menggunakan Persamaan 2.6 dan *precision* kelas negatif dihitung dengan menggunakan Persamaan 2.7 di bawah ini:

$$Precision_1 = \frac{TP}{TP+FP} \quad (2.6)$$

$$Precision_0 = \frac{TN}{TN+FN} \quad (2.7)$$

b. *Recall* menunjukkan seberapa baik model mampu mengklasifikasikan data positif dari data yang sebenarnya positif. *Recall* dihitung dengan Persamaan 2.8 di bawah ini:

$$Recall_1 = \frac{TP}{TP+FN} \quad (2.8)$$

$$Recall_0 = \frac{TN}{TN+FP} \quad (2.9)$$

c. *F1-Score* akan menunjukkan sebuah pengukutan antara *recall* dan *precision* yang seimbang dan berguna pada saat keduanya perlu diperhatikan secara bersamaan untuk mencari *mean* yang optimal dari kedua *recall* dan *precision*. *F1-Score* dapat dihitung dengan Persamaan 2.8 di bawah ini:

$$F1\text{-}Score = \frac{Precision \times Recall}{Precision + Recall} \quad (2.10)$$

d. *Support* merupakan sebuah informasi mengenai banyak data pada tiap kelas data yang telah dilakukan evaluasi.

2.2.7 *Cross Validation*

Teknik *cross validation* merupakan sebuah proses validasi model yang berguna untuk memastikan bahwa hasil yang diperoleh model menunjukan data asli, karena model dapat secara kebetulan mempelajari data *training* yang terlalu

bagus sehingga tidak dapat mengenali data yang sebelumnya belum pernah diliat. Teknik ini dapat mencegah sebuah kesalahan dalam interpretasi atau kesimpulan dari data. *Cross validation* akan membagi data menjadi k-*fold* atau k-lipatan dan semua lipatan akan secara bergantian digunakan sebagai data *training* dan data *testing* sehingga meminimalkan risiko kesalahan interpretasi model.

BAB III
METODOLOGI PENELITIAN

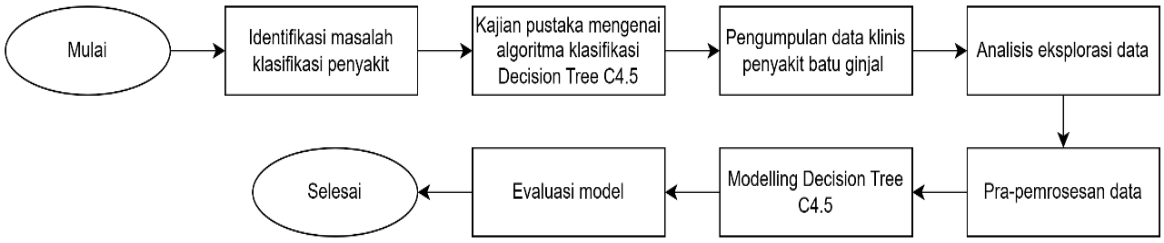
3.1 Subyek dan Obyek Penelitian

Subjek dalam penelitian ini adalah data penyakit batu ginjal tahun 2021-2024 yang diperoleh dari Rumah Sakit Umum Daerah Cideres Majalengka. Data ini akan diklasifikasikan berdasarkan 11 variabel prediktor untuk memprediksi penyakit batu ginjal untuk dapat diketahui prediksi penyakit batu ginjal. Objek penelitian ini adalah proses klasifikasi untuk memprediksi data penyakit ginjal menggunakan algoritma *Decision Tree* C4.5.

3.2 Bahan Penelitian

Data yang akan digunakan dalam penelitian ini adalah data klinis penyakit ginjal tahun 2021-2024 yang diperoleh dari Rumah Sakit Umum Daerah Cideres Majalengka.

3.3 Alur Penelitian



Gambar 3. 1 Diagram Alir Penelitian

a. Identifikasi masalah dan klasifikasi penyakit

Penerapan algoritma *Decision Tree* C4.5 pada klasifikasi penyakit batu ginjal memiliki akurasi yang tinggi [1], namun penerapan metode ini pada sumber data yang berbeda dapat menimbulkan sebuah perbedaan penanganan data karena adanya karakteristik yang berbeda antara dataset dan perbedaan distribusi variabel. Penelitian ini berupaya untuk menganalisis efektivitas metode *Decision Tree* C4.5 pada dataset yang berasal dari sumber yang

berbeda, penelitian [1] menggunakan data bersumber dari RSUD Tangerang dan penelitian ini menggunakan data yang bersumber dari RSUD Cideres Majalengka.

b. Kajian Pustaka

Pada alur penelitian bagian kajian pustaka, dilakukan eksplorasi mengenai kajian pustaka mengenai topik algoritma *Decision Tree* C4.5, penyakit batu ginjal dan teknik evaluasi pada *Machine Learning*.

c. Pengumpulan data

Pengumpulan data primer dilakukan pada Rumah Sakit Umum Daerah Cideres Majalengka sebagai sumber utama data penyakit batu ginjal. Data yang diperoleh terdiri dari 2322. Data yang diperoleh memiliki fitur sebagai berikut:

Tabel 3.1 Tipe Data

Deskripsi	Tipe Data
No	Numerik
Medical No	Numerik
Patient Name	String
Sex	Kategori
Usia	Numerik
Asupan Cairan	Numerik
Asupan Sodium	Numerik
Asupan Protein	Numerik
Kadar Asam Urat Urin	Numerik
pH Urin	Numerik
Fungsi Ginjal	Numerik
Status Batu Ginjal	Kategori

d. Analisis Eksplorasi Data

Setelah data diperoleh, dilakukan Analisis Eksplorasi Data untuk mengetahui karakteristik dalam data. Analisis yang dilakukan adalah untuk melihat apakah data *imbalance, outlier, missing value* dan melihat tipe dari masing-masing variabel untuk dilakukan pelabelan bila perlu.

Eksplorasi data yang pertama dilakukan dalam penelitian ini adalah melihat tipe data, jumlah data dan variabel yang digunakan. Eksplorasi data ini ditunjukkan pada Tabel 3.2 di bawah ini:

Tabel 3. 2 Data Klinis

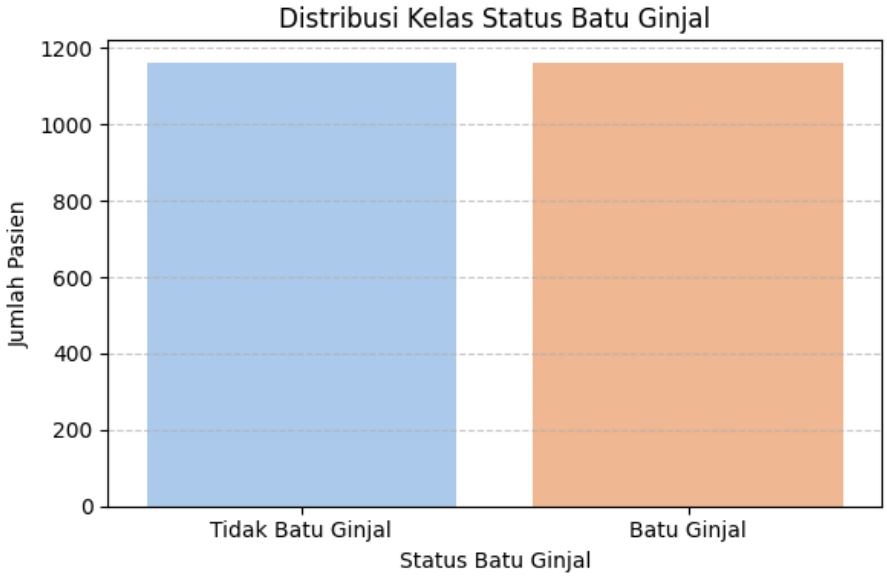
No	Medical No	Patient Name	Sex	Usia	Asupan Cairan	Asupan Sodium	…	Status Batu Ginjal
1	38****	S***** W***	L	42	2.57495	3393.75	…	Batu Ginjal
2	38****	A** R*****	L	33	1.17382	3389.09	…	Batu Ginjal
3	11****	H***** A******	P	43	2.01482	3976.75	…	Batu Ginjal
4	38****	Y**** B**** M*****	L	48	1.42394	1700.01	…	Tidak Batu Ginjal
…	…	…	…	…	…	…	…	…
2322	42****	W***** *****	P	29	2.51518	1736.30	…	Tidak Batu Ginjal

Eksplorasi data yang kedua dilakukan adalah statistik deskriptif untuk mengatahui pola dan karakteristik dari dataset. Pada statistik deskriptif memuat beberapa parameter seperti *count, mean, standar deviasi,* nilai terkecil dalam data, kuartil 1, kuartil 2 (*Median*), kuartil 3 dan nilai maksimal dalam data. Berikut adalah beberapa statistik deskriptif yang dilakukan :

Tabel 3. 3 Statistik Deskriptif

	Asupan Cairan	Asupan Sodium
Count	2232	2322
Mean	2.201105	2595.460507
Std	0.718283	836.502789
Min	1.200790	1501.156460
25%	1.632595	1794.303600
50%	2.106105	2468.793770
75%	2.701798	3399.302180
Max	1.000	3999.353580

Analisis yang selanjutnya dilakukan adalah melihat jumlah *record* masing-masing kelas dengan menggunakan grafik histogram untuk melihat apakah data memiliki kelas yang seimbang. Apabila terlihat kelas tidak seimbang maka perlu dilakukan teknik penanganan *imbalance.* Grafik histogram ditunjukkan dalam Gambar 3.2 berikut:



Gambar 3. 2 Distribusi Kelas Target

e. *Preprocessing*

Preprocessing dalam data yang karakteristiknya sudah diketahui untuk memastikan data yang digunakan sudah bersih dan siap diolah. Tahap *preprocessing* ini memuat beberapa proses seperti menghapus nilai atau kolom yang hilang, melakukan pelabelan pada variabel kategorik (jenis kelamin dan status batu ginjal).

Tahap *preprocessing* yang pertama dilakukan adalah menghapus kolom null dan kolom yang tidak digunakan dalam penelitian. Terdapat 3 kolom null dengan nama kolom "*Unamed*15", "*Unamed*16", "*Unamed*17", "*Unamed*18". Kolom yang tidak digunakan sebagai prediktor atau target adalah kolom "No", "*Patient Name*", "*Medical No".*

Tahap kedua *preprocessing* adalah memberikan label pada variabel yang memiliki tipe data kategorik. Dalam data yang digunakan, terdapat dua variabel dengan tipe data kategorik diantaranya adalah variabel *Sex* dan Status

Batu Ginjal. Terdapat dua cara untuk melakukan pelabelan yaitu *Label Encoder* dan *One Hot Encoder*. Pada penelitian ini akan menggunakan *Label Encoder* untuk menghindari pertambahan dimensi data.

Tabel 3. 4 Proses Encode Data

Variabel	Sebelum Encode Data	Sesudah Encode Data
Sex	L	0
	P	1
Status Batu Ginjal	Tidak Batu Ginjal	0
	Batu Ginjal	1

f. *Modelling Decision Tree*

Penelitian ini akan menerapkan algoritma *Decision Tree* untuk memprediksi penyakit batu ginjal. Jenis algoritma yang digunakan adalah *Decision tree* C4.5. Langkah awal dalam algoritma ini adalah menghitung proporsi data ke-i menggunakan Persamaan 2.3 di bawah ini:

$$p_i = \frac{\text{Jumlah data kelas } ke_i}{\text{Jumlah data dalam } S} \quad (2.3)$$

$$p_0 = \frac{1162}{2322} = 0.50088$$

$$p_1 = \frac{1160}{2322} = 0.49956$$

Setelah nilai Pi diketahui, dilanjutkan dengan menghitung nilai *entropy* dari dataset menggunakan Persamaan 2.2 di bawah ini:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2(pi) \quad (2.2)$$

$$Entropy(S) = [(-0.50088) \times log_2(0.50088)] + [(-0.49956) \times log_2(0.4966)]$$

$$Entropy(S) = 0.9999$$

Selanjutnya, nilai *entropy sample* digunakan untuk menghitung nilai *entropy* dari semua variabel yang digunakan. Setelah nilai *entropy* dari semua variabel diperoleh, dilakukan perhitungan nilai *gain* dengan menggunakan Persamaan 2.4. Setelah dilakukan perhitungan *entropy* diperoleh nilai dari variabel asupan sodium adalah 0.236 dan 0.6721 dan dilakukan perhitungan *gain* menggunakan Persamaan 2.4 di bawah ini :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad (2.4)$$

$$Gain(S, A) = 0.9999 - \left(\frac{|956|}{|2322|} \times 0.236\right) + \left(\frac{|1366|}{|2322|} \times 0.6721\right)$$

$$Gain(S, A) = 0.50726$$

Hasil perhitungan menunjukkan nilai *gain* dari variabel asupan sodium yang merupakan akar atau *root node* dalam penelitian ini.

g. Evaluasi

Tahap evaluasi dalam penelitian ini akan mengevaluasi kinerja algoritma *Decision Tree* C4.5 dengan menggunakan *Confusion Matrix* yang terdiri dari TP, TN, FP dan nilai FN. Nilai tersebut kemudian digunakan untuk menghitung *accuracy*, *precision*, *recall*, *f1-score* dan *support*.

h. Analisis Hasil dan Kesimpulan

Hasil klasifikasi dengan menggunakan algoritma *Decision Tree* digunakan untuk memprediksi penyakit batu ginjal. Dengan nilai *Accuracy*, *Confusion Matrix*, *Classification Report* yang didapatkan diharapkan mampu diketahui seberapa baik model dapat melakukan prediksi pada data baru yang belum diketahui kelasnya.

BAB IV
HASIL DAN PEMBAHASAN

4.1 *Exploratory Data Analysis*

Tahap *Exploratory Data Analysis* terdiri dari beberapa bagian, yaitu data *understanding*, statistik deskriptif, dan visualisasi data. *Data understanding* dilakukan dengan tujuan untuk menggali informasi secara mendalam, mengungkap pola, menemukan kejanggalan, serta mendapatkan gambaran awal dari data. *Data understanding* yang dilakukan adalah melihat dimensi data dan tipe data dari semua variabel yang digunakan.

Analisis eksplorasi data yang pertama dilakukan adalah melihat dimensi dari data yang digunakan, dengan perintah df.*shape* yang menghasilkan jumlah baris dan jumlah kolom. Dataset terdiri dari 2322 baris dan 19 kolom. Adapun penjelasan tipe data dan penjelasan satuan variabel penelitian disajikan pada Tabel 4.1.

Tabel 4. 1 Tipe dan Satuan Data

Deskripsi	Satuan	Tipe Data
No	-	Numerik
Medical No	-	Numerik
Patient Name	-	String
Sex	-	Kategori
Usia	-	Numerik
Asupan Cairan	liter/hari	Numerik
Asupan Sodium	mg/hari	Numerik
Asupan Protein	g/hari	Numerik
Kadar Asam Urat Urin	mg/dL	Numerik
pH Urin	-	Numerik
Fungsi Ginjal	mL/min	Numerik
Status Batu Ginjal	-	Kategori

Variabel *Sex* dan Status Batu Ginjal memiliki tipe data ketegori, variabel nama pasien memiliki tipe data string dan untuk variabel lainnya memiliki tipe data numerik. Tampilan 4 baris pertama dari dataset dijelaskan dalam Tabel 4.2 di bawah ini :

Tabel 4. 2 Tampilan Data

No	Medical No	Patient Name	Sex	Usia	Asupan Cairan	Asupan Sodium	…	Status Batu Ginjal
1	38****	S***** W***	L	42	2.57495	3393.75	…	Batu Ginjal
2	38****	A** R*****	L	33	1.17382	3389.09	…	Batu Ginjal
3	11****	H***** A******	P	43	2.01482	3976.75	…	Batu Ginjal
4	38****	Y**** B**** M*****	L	48	1.42394	1700.01	…	Tidak Batu Ginjal
…	…	…	…	…	…	…	…	…
2322	42****	W***** *****	P	29	2.51518	1736.30	…	Tidak Batu Ginjal

Analisis eksplorasi yang selanjutnya dilakukan adalah melihat jumlah kolom yang memiliki nilai *null*. Data yang digunakan memiliki kolom *Unnamed*: 15, *Unnamed*: 16, *Unnamed*: 17 dan *Unnamed*: 18 yang *null* dan dianggap tidak digunakan sehingga perlu dihapus.

4.2 Data *Preprocessing*

Tahapan pra-pemrosesan data dilakukan untuk mempersiapkan data mentah yang didapatkan sehingga dapat digunakan untuk proses pemodelan. Pada tahap ini memuat penghapusan kolom *null* yang dianggap tidak digunakan dalam proses analisis dan dilakukan *encoding* data kategorikal.

Pada dataset, terdapat sebanyak 4 kolom *null* yaitu *Unnamed*:15, *Unnamed*:16, *Unnamed*:17 dan *Unnamed*:18 yang dihapus. Kemudian dilakukan *label encode* variabel Y dari yang awalnya memiliki 2 kelas yaitu "Batu Ginjal" dan "Tidak Batu Ginjal" menjadi 1 dan 0, hasil encode disajikan dalam tabel 4.3 di bawah ini :

Tabel 4. 3 Hasil *Labelling*

Variabel	Sebelum Encode Data	Sesudah Encode Data
Sex	L	0
	P	1
Status Batu Ginjal	Tidak Batu Ginjal	0
	Batu Ginjal	1

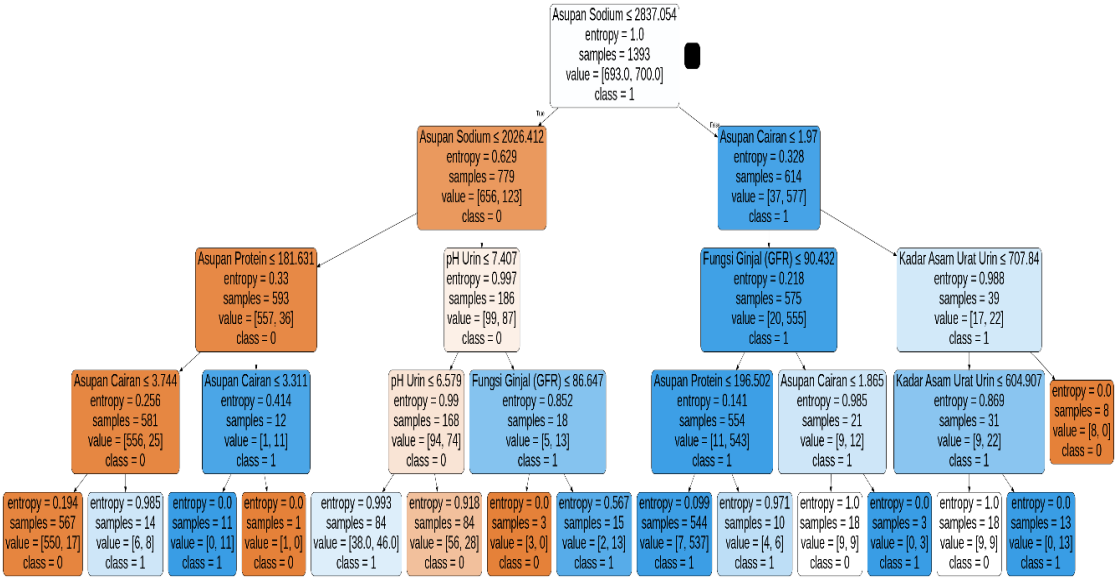
Pra pemrosesan selanjutnya yang dilakukan adalah membagi dataset menjadi X dan Y. Dalam tahap ini, variabel X yang digunakan adalah Usia, Asupan Cairan, Asupan Sodium, Asupan Protein, Kadar Kalsium Urin, Kadar Asam Urat Urin, pH Urin, Kadar Oksalat dalam Urin, dan Fungsi Ginjal.

Setelah data sudah dibagi menjadi X dan Y, variabel tersebut akan dibagi menjadi data pelatihan dan data pengujian. Data pelatihan yang digunakan sebesar 60% dan data pengujian sebesar 40%. Jumlah data yang digunakan dalam pengujian adalah sebesar 929 baris dan 10 kolom, sedangkan jumlah data pelatihan adalah sebesar 1393 baris dan 10 kolom.

4.3 *Modelling*

Tahap *modelling* dilakukan dengan menerapkan algoritma *Decision Tree* C4.5 menggunakan modul *DecisionTreeClassifier* dari *scikit-learn* untuk melatih model. Model *Decision Tree* menghasilkan sebuah pohon keputusan yang berisi *root node*, *sub tree* dan *leaf node* yang masing-masing akan berisi nilai *entropy*, jumlah *sample*, *value* dan hasil klasifikasi kelas.

Berikut adalah hasil pohon keputusan pada penelitian ini:



Gambar 4. 1 Hasil *Decision Tree*

Pohon keputusan di atas menghasilkan aturan atau rule dalam penelitian ini, yaitu:

1. *Root Node*

   Node akar (*root node*) yang dihasilkan adalah Asupan Sodium, dengan nilai *entropy* sebesar 1. Nilai *entropy* ini menunjukkan ketidakmurnian bahwa data pada *node* tersebut masih bercampur atau belum terpisah dengan jelas antara kategori batu ginjal dan tidak batu ginjal. Ketidakmurnian *node* akar dibuktikan dengan adanya data yang terbagi hampir rata di kedua kelas. Pada *node* ini terdapat *sample* sebanyak 1393 yang artinya terdapat 1393 data yang diproses oleh model untuk selanjutnya dibagi menjadi kelas target. Sebanyak 1393 sample terbagi menjadi 2 *value* yaitu 700 dikategorikan sebagai kelas batu ginjal dan sebanyak 693 dikategorikan sebagai kelas tidak batu ginjal. Pada *node* ini, prediksi mayoritas adalah kelas 1 atau kelas batu ginjal karena data terbagi lebih banyak pada kelas 1 (sebanyak 700 data).

   Setelah *root node* diketahui, pohon keputusan akan melanjutkan proses untuk mencari nilai *gain* tertinggi untuk selanjutnya digunakan sebagai *node* setiap cabang. *Node* akan terbagi menjadi 2 cabang yaitu cabang *true* (kiri) dan cabang *false* (kanan). Kedua cabang tersebut merupakan *sub-tree* dari pohon keputusan.

2. Cabang *True* (Kiri)

   Cabang kiri atau *sub-tree true* menghasilkan aturan sebagai berikut :

   Aturan dijelaskan dalam Gambar 4.2 di bawah ini:

```
|--- Asupan Sodium <= 2837.05
|   |--- Asupan Sodium <= 2026.41
|   |   |--- Asupan Protein <= 181.63
|   |   |   |--- Asupan Cairan <= 3.74
|   |   |   |   |--- class: 0
|   |   |   |--- Asupan Cairan >  3.74
|   |   |   |   |--- class: 1
|   |   |--- Asupan Protein >  181.63
|   |   |   |--- Asupan Cairan <= 3.31
|   |   |   |   |--- class: 1
|   |   |   |--- Asupan Cairan >  3.31
|   |   |   |   |--- class: 0
|   |--- Asupan Sodium >  2026.41
|   |   |--- pH Urin <= 7.41
|   |   |   |--- pH Urin <= 6.58
|   |   |   |   |--- class: 1
|   |   |   |--- pH Urin >  6.58
|   |   |   |   |--- class: 0
|   |   |--- pH Urin >  7.41
|   |   |   |--- Fungsi Ginjal (GFR) <= 86.65
|   |   |   |   |--- class: 0
|   |   |   |--- Fungsi Ginjal (GFR) >  86.65
|   |   |   |   |--- class: 1
```

Gambar 4. 2 *Sub-Tree True*

Cabang *true* memiliki aturan sebagai berikut:

1) **Aturan 1**

   Jika :

   - Asupan Sodium ≤ 2837.05
   - Asupan Sodium ≤ 2026.41
   - Asupan Protein ≤ 181.63
   - Asupan Cairan ≤ 3.74

   Maka: **class: 0** (tidak batu ginjal)

   Aturan 1 menghasilkan kelas tidak batu ginjal apabila fitur memiliki nilai asupan sodium yang berkisar antara ≤ 2026.41 sampai ≤ 2837.05, asupan protein ≤ 181.63 dan asupan cairan ≤ 3.74. Dalam aturan ini, terapat 550 kasus yang terklasifikasi menjadi kelas 0.

2) **Aturan 2**

   Jika :

   - Asupan Sodium ≤ 2837.05

- Asupan Sodium ≤ 2026.41
- Asupan Protein ≤ 181.63
- Asupan Cairan > 3.74

Maka: class: 1 (batu ginjal)

Aturan 2 menghasilkan kelas batu ginjal apabila fitur memiliki nilai asupan sodium yang berkisar antara ≤ 2026.41 sampai ≤ 2837.05, asupan protein ≤ 181.63 dan asupan cairan > 3.74. Dalam aturan ini terdapat 8 kasus yang terklasifikasi sebagai kelas 1.

3) **Aturan 3**

Jika :

- Asupan Sodium ≤ 2837.05
- Asupan Sodium ≤ 2026.41
- Asupan Protein > 181.63
- Asupan Cairan ≤ 3.31

Maka: class: 1 (batu ginjal)

Aturan 3 menghasilkan kelas batu ginjal apabila asupan sodium ≤ 2837.05, asupan sodium ≤ 2026.41, asupan protein > 181.63, asupan cairan ≤ 3.31. Dalam aturan 3 terdapat 11 kasus.

4) **Aturan 4**

Jika:

- Asupan Sodium ≤ 2837.05
- Asupan Sodium ≤ 2026.41
- Asupan Protein > 181.63
- Asupan Cairan > 3.31

Maka: class: 0 (tidak batu ginjal)

Aturan 4 menghasilkan klasifikasi kelas 0 apabila asupan sodium ≤ 2026.41 sampai ≤ 2837.05, asupan protein > 181.63, asupan cairan > 3.31. Dalam aturan 4 hanya terdapat 1 kasus.

5) **Aturan 5**

Jika:

- Asupan Sodium ≤ 2837.05
- Asupan Sodium > 2026.41
- pH Urin ≤ 7.41
- pH Urin ≤ 6.58

Maka: class: 1 (batu ginjal)

Aturan 5 menghasilkan klasifikasi pada kelas 1 apabila Asupan Sodium ≤ 2837.05, Asupan Sodium > 2026.41, pH Urin pH Urin ≤ 6.58 sampai 7.41. Aturan 5 terdapat sebanyak 46 kasus kelas 1.

6) **Aturan 6**

Jika:

- Asupan Sodium ≤ 2837.05
- Asupan Sodium > 2026.41
- pH Urin ≤ 7.41
- pH Urin > 6.58

Maka: class: 0 (tidak batu ginjal)

Aturan 6 menghasilkan klasifikasi kelas 0 apabila asupan sodium ≤ 2837.05, asupan sodium > 2026.41 dan pH urin antara > 6.58 sampai 7.41. Aturan 6 terdapat sebanyak 56 kasus.

7) **Aturan 7**

Jika:

- Asupan Sodium ≤ 2837.05
- Asupan Sodium > 2026.41
- pH Urin > 7.41
- Fungsi Ginjal (GFR) ≤ 86.65

Maka: class: 0 (tidak batu ginjal)

Aturan 7 menghasilkan klasifikasi pada kelas 0 apabila asupan sodium ≤ 2837.05, asupan sodium > 2026.41, pH urin >7.41 dan fungsi ginjal ≤ 86.65. Aturan 7 terdapat sebanyak 3 kasus.

8) **Aturan 8**

Jika:

- Asupan Sodium ≤ 2837.05
- Asupan Sodium > 2026.41
- pH Urin > 7.41
- Fungsi Ginjal (GFR) > 86.65

Maka: class: 1 (batu ginjal)

Aturan terakhir pada cabang true menghasilkan klasifikasi kelas 1 dengan aturan apabila asupan sodium berada antara > 2026.41 sampai 2837.05, pH urin > 7.41 dan fungsi ginjal >86.65. Aturan 8 terdapat sebanyak 13 kasus.

3. Cabang *False* (Kanan)

Cabang kanan atau *sub-tree false* menghasilkan aturan yang dijelaskan dalam Gambar 4.3 di bawah ini:

```
|--- Asupan Sodium >  2837.05
|   |--- Asupan Cairan <= 1.97
|   |   |--- Fungsi Ginjal (GFR) <= 90.43
|   |   |   |--- Asupan Protein <= 196.50
|   |   |   |   |--- class: 1
|   |   |   |--- Asupan Protein >  196.50
|   |   |   |   |--- class: 1
|   |   |--- Fungsi Ginjal (GFR) >  90.43
|   |   |   |--- Asupan Cairan <= 1.87
|   |   |   |   |--- class: 0
|   |   |   |--- Asupan Cairan >  1.87
|   |   |   |   |--- class: 1
|   |--- Asupan Cairan >  1.97
|   |   |--- Kadar Asam Urat Urin <= 707.84
|   |   |   |--- Kadar Asam Urat Urin <= 604.91
|   |   |   |   |--- class: 0
|   |   |   |--- Kadar Asam Urat Urin >  604.91
|   |   |   |   |--- class: 1
|   |   |--- Kadar Asam Urat Urin >  707.84
|   |   |   |--- class: 0
```

Gambar 4. 3 *Sub-Tree False*

Cabang *false* memiliki aturan sebagai berikut:

1) **Aturan 9**

   Jika:

   - Asupan Sodium > 2837.05
   - Asupan Cairan ≤ 1.97
   - Fungsi Ginjal (GFR) ≤ 90.43
   - Asupan Protein ≤ 196.50

   Maka: class: 1 (batu ginjal)

   Aturan 9 menghasilkan klasifikasi pada kelas 1 apabila asupan sodium > 2837.05, asupan cairan ≤ 1.97, fungsi ginjal (GFR) ≤ 90.43 dan asupan protein ≤ 196.50. Pada aturan 9 menghasilkan sebanyak 537 kasus terklasifikasi sebagai kelas batu ginjal.

2) **Aturan 10**

   Jika:

   - Asupan Sodium > 2837.05
   - Asupan Cairan ≤ 1.97
   - Fungsi Ginjal (GFR) ≤ 90.43
   - Asupan Protein > 196.50

   Maka: class: 1 (batu ginjal)

   Aturan 10 menghasilkan kelas batu ginjal apabila asupan sodium > 2837.05, asupan cairan ≤ 1.97, fungsi Ginjal (GFR) ≤ 90.43, asupan protein > 196.50. Aturan 10 menghasilkan klasifikasi kelas 1 sebanyak 6 kasus.

3) **Aturan 11**

   Jika:

   - Asupan Sodium > 2837.05

- Asupan Cairan ≤ 1.97
- Fungsi Ginjal (GFR) > 90.43
- Asupan Cairan ≤ 1.87

Maka: class: 0 (tidak batu ginjal)

Aturan 11 menghasilkan klasifikasi kelas tidak batu ginjal asupan sodium > 2837.05, asupan cairan ≤ 1.97, fungsi ginjal (GFR) > 90.43 dan asupan cairan ≤ 1.87. Aturan 11 menghasilkan sebanyak 9 kasus terklasifikasi sebagai kelas 0.

4) **Aturan 12**

Jika:

- Asupan Sodium > 2837.05
- Asupan Cairan ≤ 1.97
- Fungsi Ginjal (GFR) > 90.43
- Asupan Cairan > 1.87

Maka: class: 1 (batu ginjal)

Aturan 12 menghasilkan klsifikasi kelas batu ginjal apabila asupan sodium > 2837.05, asupan cairan ≤ 1.97, , fungsi ginjal (GFR) > 90.43 dan asupan cairan > 1.87. Aturan 12 menghasilkan sebanyak 3 kasus terklasifikasi sebagai kelas 1.

5) **Aturan 13**

Jika:

- Asupan Sodium > 2837.05
- Asupan Cairan > 1.97
- Kadar Asam Urat Urin ≤ 707.84
- Kadar Asam Urat Urin ≤ 604.91

Maka: class: 0 (tidak batu ginjal)

Aturan 13 menghasilkan klasifikasi kelas 0 apabila asupan sodium > 2837.05, asupan cairan > 1.97, kadar asam urat urin ≤ 707.84 dan kadar asam urat urin ≤ 604.91. Aturan 13 manghasilkan sebanyak 9 kasus yang terklasifikasi sebagai kelas 0.

6) **Aturan 14**

Jika:

- Asupan Sodium > 2837.05
- Asupan Cairan > 1.97
- Kadar Asam Urat Urin ≤ 707.84
- Kadar Asam Urat Urin > 604.91

Maka: class: 1 (batu ginjal)

Aturan 13 menghasilkan klasifikasi kelas 1 apabila asupan sodium > 2837.05, asupan cairan > 1.97, kadar asam urat urin ≤ 707.84 dan kadar asam urat urin > 604.91. Aturan 14 manghasilkan sebanyak 913 kasus yang terklasifikasi sebagai kelas 1.

7) **Aturan 15**

Jika:

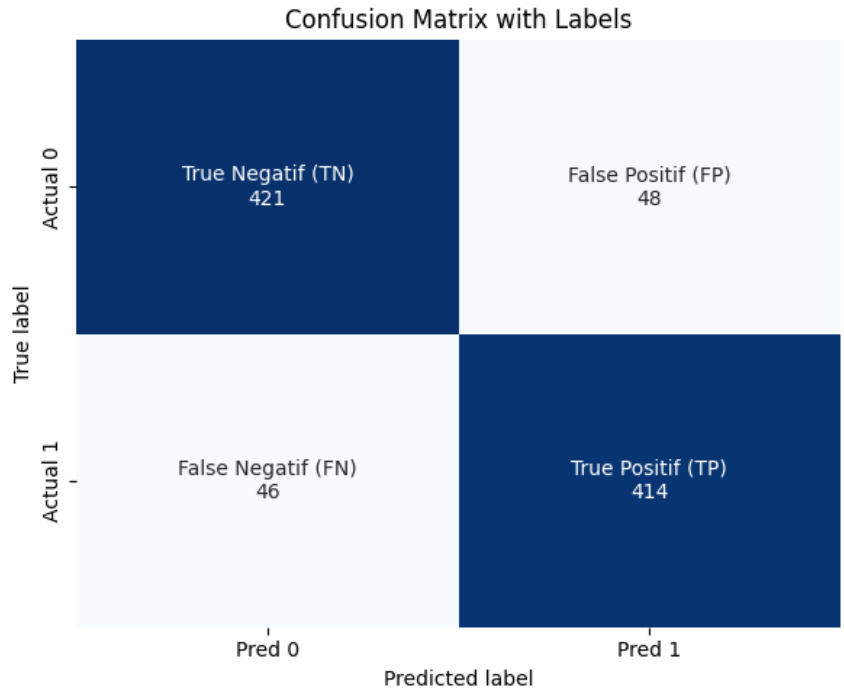
- Asupan Sodium > 2837.05
- Asupan Cairan > 1.97
- Kadar Asam Urat Urin > 707.84

Maka: class: 0 (tidak batu ginjal)

Aturan 15 menghasilkan klasifikasi kelas tidak batu ginjal apabila asupan sodium > 2837.05, asupan cairan > 1.97, kadar asam urat urin > 707.84. Aturan 15 menghasilkan sebanayak 8 kasus yang terklasifikasi menjadi kelas tidak batu ginjal.

4.4 Evaluasi

Model *Decision Tree* C4.5 yang dikembangkan menggunakan kriteria pemisahan *entropy*, dan menghasilkan hasil matriks konfusi yang digunakan untuk menghitung akurasi model. Hasil matriks konfusi pada penelitian ini terdapat pada Gambar 4.4 di bawah :



Gambar 4. 4 Matriks Konfusi

Berdasarkan hasil matriks konfusi, akurasi model dan *classification report* dapat dihitung dengan menggunakan nilai TN, FN, FP dan nilai TP. Berikut adalah perhitungan *classification report* dan akurasi model:

1. *Precision*

   Hasil *precision* digunakan untuk mengukur seberapa tepat model dalam mengklasifikasikan data positif (batu ginjal). Berdasarkan matriks konfusi didapatkan nilai TP sebesar 350 dan FP sebesar 7 sehingga dapat dilakukan perhitungan *precision.* Perhitungan *precision* kelas positif atau kelas 1 menggunakan Persamaan 2.6 di bawah ini:

$$Precision_1 = \frac{TP}{TP+FP} \quad (2.6)$$

$$Precision_1 = \frac{414}{414+48}$$

$$Precision_1 = 0.8961 \approx 0.90$$

Nilai *precision* kelas positif sebesar **0.90** menunjukkan bahwa 90% dari prediksi positif yang dibuat oleh model adalah benar. *Precision* kelas *negative* dapat dihitung dengan menggunakan Persamaan 2.7 di bawah ini:

$$Precision_0 = \frac{TN}{TN+FN} \quad (2.7)$$

$$Precision_0 = \frac{421}{421+46}$$

$$Precision_0 = 0.9015 \approx 0.90$$

Nilai *precision* kelas negatif sebesar **0.90** menunjukkan bahwa 90% dari prediksi negatif yang dibuat oleh model adalah benar.

2. *Recall*

   Recall atau sensitivitas mengukur kemampuan model dalam mendeteksi data positif secara benar dari keseluruhan data positif yang sebenarnya. Semakin tinggi nilai *recall*, semakin baik model dalam menangkap semua kasus positif. Nilai *recall* kelas positif dapat dihitung dengan Persamaan 2.8 di bawah ini:

$$Recall_1 = \frac{TP}{TP+FN} \quad (2.8)$$

$$Recall_1 = \frac{414}{414+46}$$

$$Recall_1 = 0.90$$

Hasil menujukkan *recall* sebesar 90% yang artinya model mampu mendeteksi 90% dari semua kasus positif yang sebenarnya. *Recall* pada kelas negatif dapat dihitung dengan Persamaan 2.9 di bawah ini:

$$Recall_0 = \frac{TN}{TN+FP} \quad (2.9)$$

$$Recall_0 = \frac{421}{421 + 48}$$

$$Recall_0 = 0.8977 \approx 0.90$$

Hasil menujukkan *recall* sebesar 90% yang artinya model mampu mendeteksi 90% dari semua kasus negatif yang sebenarnya.

3. F1-*Score*

F1-*score* merupakan harmonic mean dari *precision* dan *recall*. Metrik ini digunakan ketika dibutuhkan keseimbangan antara *precision* dan *recall*, khususnya ketika distribusi kelas tidak seimbang.

$$F1\text{-}Score_1 = 2 \times \frac{Precision_1 \times Recall_1}{Precision_1 + Recall_1} \quad (2.10)$$

$$F1\text{-}Score_1 = 2 \times \frac{0.9015 \times 0.8977}{0.9015 + 0.8977}$$

$$F1\text{-}Score_1 = 0.8996 \approx 0.90$$

Nilai F1-score sebesar 0.90 menunjukkan bahwa model memiliki keseimbangan yang sangat baik antara *precision* dan *recall* pada kelas positif.

$$F1\text{-}Score_0 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.10)$$

$$F1\text{-}Score_0 = 2 \times \frac{0.8961 \times 0.90}{0.8961 + 0.90}$$

$$F1\text{-}Score_0 = 0.8980 \approx 0.90$$

Nilai F1-score sebesar 0.90 menunjukkan bahwa model memiliki keseimbangan yang sangat baik antara *precision* dan *recall* pada kelas negatif. Hasil dari *classification report* disajikan dalam Tabel 4.4 di bawah ini:

Tabel 4. 4 *Classification Report*

Kelas	*Precision*	*Recall*	F1-*Score*
0	0.90	0.90	0.90
1	0.90	0.90	0.90

4. *Accuracy*

   Accuracy model dihitung dengan menggunakan nilai-nilai yang ada pada matriks konfusi. *Accuracy* dihitung dengan menggunakan Persamaan 2.5 di bawah ini :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.5)$$

$$Accuracy = \frac{414 + 421}{414 + 421 + 48 + 46}$$

$$Accuracy = 0.8999 \approx 0.90$$

Berdasarkan hasil evaluasi menggunakan *confusion matrix*, hasil akurasi dari model klasifikasi pohon keputusan sebesar 90%. Artinya, dari sebanyak 929 data uji, sebanyak 91.29% yang dilakukan model sesuai dengan label aktual. Akurasi 89.88% juga menunjukkan bahwa model mengklasifikasinya sebanyak 835 dari total data uji.

Tabel 4. 5 Hasil Akurasi

Jenis Pengujian	Hasil Akurasi
Training-set	91.39%
Test-set	89.88%
Model Accuracy	90%

Untuk menguji kemampuan generalisasi model, dilakukan prediksi menggunakan data baru sebanyak 10 baris yang sebelumnya tidak dimasukkan ke dalam proses pelatihan model. Data baru ini digunakan untuk mengevaluasi performa model terhadap data yang belum pernah dilihat sebelumnya.

Asupan Cairan (liter/hari)	Asupan Sodium (mg/hari)	Asupan Protein (g/hari)	Kadar Asam Urat Urin (mg/dL)	pH Urin	Fungsi Ginjal (GFR) mL/min	Status Batu Ginjal
2.57	1605.91	63.01	515.34	7.88	118.82	Tidak Batu Ginjal
1.17	3621.76	84.34	626.29	5.06	72.49	Batu Ginjal
2.01	1734.93	62.00	377.03	6.04	118.82	Tidak Batu Ginjal
1.42	3347.70	87.64	723.07	4.66	80.01	Batu Ginjal
1.92	3241.51	97.44	650.38	5.05	77.75	Batu Ginjal
2.72	1104.76	56.01	338.95	7.37	114.40	Tidak Batu Ginjal
1.06	3253.68	95.94	691.97	5.29	77.98	Batu Ginjal
2.81	1495.98	64.25	346.37	6.08	91.22	Tidak Batu Ginjal
2.67	1006.66	68.14	319.40	6.08	115.15	Tidak Batu Ginjal
1.03	3098.31	101.13	693.57	5.60	66.02	Batu Ginjal

Gambar 4. 5 Data Prediksi

Hasil prediksi dari model menunjukkan bahwa kelas target yang dihasilkan sesuai dengan data yang digunakan untuk prediksi yang ditunjukkan pada Tabel 4.6 berikut:

Tabel 4. 6 Hasil Prediksi Model

Status Batu Ginjal (Data Aktual)	Prediksi Model
Tidak Batu Ginjal	Tidak Batu Ginjal
Batu Ginjal	Batu Ginjal
Tidak Batu Ginjal	Tidak Batu Ginjal
Batu Ginjal	Batu Ginjal
Batu Ginjal	Batu Ginjal
Tidak Batu Ginjal	Tidak Batu Ginjal
Batu Ginjal	Batu Ginjal
Tidak Batu Ginjal	Tidak Batu Ginjal
Tidak Batu Ginjal	Tidak Batu Ginjal
Batu Ginjal	Batu Ginjal

Hasil prediksi model terhadap 10 data baru menunjukkan bahwa seluruh prediksi sesuai dengan label aktual pasien.

4.5 Analisis Hasil dan Kesimpulan

Model Decision Tree dibangun menggunakan kriteria pemisahan berbasis entropy, dan mampu memberikan tingkat akurasi yang tinggi pada data *training* maupun data *testing*. Model menghasilkan akurasi keseluruhan sebesar 90% yang artinya model menunjukkan performa yang konsisten dalam mengenali kedua kelas, baik kelas 0 maupun kelas 1. Sebanyak 929 data uji yang digunakan, model mampu memprediksi dengan benar sebanyak 835 data. Dalam data latih, akurasi yang diperoleh sebesar 91.39% menunjukkan bahwa model mampu mempelajari pola data dengan baik pada saat pelatihan. Sedangkan pada data uji, akurasi yang dihasilkan sebesar 89.88% yang artinya model memiliki kemampuan untuk generalisasi yang baik pada data baru yang belum pernah dipelajari oleh model. Perbandingan akurasi pada data pelatihan dengan akurasi pada data pengujian menunjukkan bahwa pada model tersebut tidak menunjukkan adanya *overfitting* dan *underfitting*.

Model juga diuji dengan validasi silang menggunakan *k-fold cross-validation* sebanyak 5 lipatan. Hasil nilai akurasi pada masing-masing lipatan adalah *Fold* 1: 88,89%, *Fold* 2: 88,89%, *Fold* 3: 84,23%, *Fold* 4: 87,42%, *Fold* 5: 87,8% yang menunjukkan bahwa model mampu melakukan generalisasi dengan baik. Model rata-rata akurasi validasi silang sebesar 87,44% yang menunjukkan konsisten pada data pelatihan, pengujian, dan validasi silang dan menunjukkan bahwa model tidak mengalami *underfitting* dan tidak mengalami *overfitting* karena perbedaan akurasi antara pelatihan dan pengujian tidak jauh.

Hasil visualisasi pohon keputusan menunjukkan bahwa variabel asupan sodium menjadi node akar karena memiliki nilai *gain* yang tinggi dan node tersebut tidak murni seutuhnya milik satu kelas atau karena nilainya adalah 1. Node akar ini membagi data menjadi 2 kelas yaitu *sub tree true* dan *false* yang masing-masing memiliki aturan yang berbeda. Dari hasil pohon keputusan yang diperoleh, dapat dikatakan bahwa pada *sub tree* false, asupan cairan dan kadar asam urat urin memiliki pengaruh yang penting dalam proses klasifikasi. Sedangkan dalam *sub tree true*, asupan cairan dan pH urin memiliki pengaruh paling penting dalam proses klasifikasi di cabang ini.

Hasil prediksi pada 10 data pasien terbaru dengan 5 pasien memiliki label aktual batu ginjal dan sebanyak 5 pasien memiliki label aktual tidak batu ginjal. Model menunjukkan hasil yang sesuai dengan label aktual,sehingga dapat dikatakan bahwa model menunjukkan bahwa algoritma *Decision Tree* C4.5 mampu melakukan generalisasi dengan baik pada data yang benar-benar baru dan belum pernah dikenali oleh model.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil evaluasi pada model *Decision Tree* C4.5 untuk mengklasifikasikan penyakit batu ginjal yang menggunakan kriteria *entropy*, hasil yang diperoleh menunjukkan hasil yang baik.

1. Akurasi pada data pelatihan mencapai 91.39%, dan pada data pengujian sebesar 89.88% yang menunjukkan bahwa pada model tidak menunjukkan adanya *overfitting* dan *underfitting*. Nilai *precision* dan *recall* sebesar 90%, yang mengindikasikan bahwa model memiliki ketepatan dan sensitivitas yang sangat baik dalam mengenali masing-masing kelas. Nilai *F1-score* sebesar 90% menunjukkan bahwa model memiliki keseimbangan yang baik antara *precision* dan *recall*.
2. Evaluasi menggunakan teknik *cross-validation* menghasilkan rata-rata akurasi sebesar 87.44%, yang artinya model ini cukup stabil dan tidak terlalu bergantung pada data tertentu.

Dengan demikian, dapat disimpulkan bahwa metode *Decision Tree* C4.5 berhasil membentuk model klasifikasi yang akurat dan andal untuk mendeteksi status penyakit batu ginjal berdasarkan atribut-atribut yang tersedia.

5.2 Saran

Untuk penelitian selanjutnya, penulis menyarankan untuk menggunakan data yang lebih bervariasi dengan lebih banyak atribut yang relevan seperti riwayat keluarga atau jenis batu ginjal agar model klasifikasi dapat memberikan hasil yang lebih komprehensif. Selain itu, meskipun algoritma *Decision Tree* C4.5 telah menunjukkan performa yang sangat baik, akan lebih baik jika dilakukan perbandingan dengan algoritma klasifikasi lainnya seperti *Random Forest* karena algoritma ini efisien dalam menangani data dengan banyak atribut. Penggunaan penanganan *imbalance class* seperti SMOTE juga dapat dipertimbangkan apabila ditemukan ketidakseimbangan jumlah data antar kelas.

DAFTAR PUSTAKA

[1] Y. Widiastiwi and I. Ernawati, "Klasifikasi Penyakit Batu Ginjal Menggunakan Algoritma Decision Tree C4.5 Dengan Membandingkan Hasil Uji Akurasi," *J. IKRA-ITH Inform.*, vol. 5, no. 2, p. 128, 2021.

[2] E. Hadibrata and Suharmanto, "Faktor-Faktor Yang Berhubungan Dengan Terjadinya Batu," *Jurnal Penelitian Perawat Profesional.*, vol. 4, no. 3, pp. 1041–1046, 2022.

[3] I. Russari, "Sistem Pakar Diagnosa Penyakit Batu Ginjal Menggunakan Teorema Bayes," *Jurnal Riset Komputer*, vol. 3, no. 1, pp. 18–22, 2016, doi: 10.30865/jurikom.v3i1.44.

[4] A. Lidesna Shinta Amat and H. Pieter Louis Wungouw, "Deteksi Dini Batu Ginjal (Early Detection of Kidney Stones)," *Jurnal Pengabdian Masyarakat.*, vol. 2, no. 1, pp. 24–27, 2022.

[5] G. A. Ali Muhammad Kavosh, Ali Reza Aminsharifi, Vahid Keshtkar, Abdosaleh Jafari, "Is screening of Staghorn Stones cost-effective?," *Urology Journal.*, vol. 16, no 4. July, pp. 337–342, 2019.

[6] P. Rahayu et al., Manfaat Data Mining, *Buku Ajar Data Mining*, Edisi Pertama, Jambi : PT. Sonpedia Publishing Indonesia, 2024.

[7] M. M. Muttaqin, Wahyu Wijaya Widiyanto, A. W. Green Ferry Mandias, Stenly Richard Pungus, S. A. H. Wiranti Kusuma Hapsari, E. F. B. Aslam Fatkhudin, Pasnur, and N. S. Mochammad Anshori, Suryani. 2023. Pengenalan Data Mining, Sumatera Utara : Yayasan Kita Menulis.

[8] N. N. Habibah, A. Nazir, I. Iskandar, F. Syafria, L. Oktavia, and I. Syurfi, "Pemodelan Klasifikasi Untuk Menentukan Penyakit Diabetes dengan Faktor Penyebab Menggunakan Decision Tree C4.5 Pada Wanita," *Jurnal Sistem Komputer dan Informasi.*, vol. 4, no. 4, p. 654, 2023.

[9] B. A. Candra Permana and I. K. Dewi Patwari, "Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi

Penyakit Diabetes," *Jurnal Informasi dan Teknologi.*, vol. 4, no. 1, pp. 63–69, 2021.

[10] M. A. U. Yosef Mulyanto Dawa, Abdul Aziz, "Optimasi Algoritma C4.5 Berbasis Particle Swarn Optimization (PSO) untuk Menentukan Wholesales Penjualan," *Jurnal Riset Mahasiswa Bidang Teknologi Informasi.*, vol. 6, pp. 21–26, 2023.

[11] J. Barale, "Comparative Analysis of Algorithms to Predict the kidney Stone," December, 2023.

[12] K. Imelda, "Penerapan Metode Klasifikasi Dengan Algoritma Decision Tree C4.5 Untuk Mendiagnosa Awal Penyakit Ginjal Kronis," *Jurnal Pelita Teknologi Bangsa,* vol. 15, no. 1, pp. 7-14, 2024.

[13] K. L. Kohsasih and Z. Situmorang, "Analisis Perbandingan Algoritma C4.5 Dan Naïve Bayes Dalam Memprediksi Penyakit Cerebrovascular," *Jurnal. Informatika.*, vol. 9, no. 1, pp. 13–17, 2022.

[14] R. Estian Pambudi, "Klasifikasi Penyakit Stroke Menggunakan Algoritma Decision Tree C4.5," *Jurnal Teknika.*, vol. 16 , No 02, pp. 221-226 , 2022.

[15] A. Afifuddin and L. Hakim, "Deteksi Penyakit Diabetes Mellitus Menggunakan Algoritma Decision Tree Model Arsitektur C4.5," *Jurnal Krisnadana,* vol. 3, no. 1, pp. 25–33, 2023.

[16] M. M. S. Jogo, M. K. Biddinika, and A. Fadlil, "Klasifikasi Penyakit Diabetes dengan Algoritma Decision Tree dan Naïve Bayes," *Jurnal Elektronika Kendali Telekomunikasi Tenaga Listrik Komputer Vol.*, vol. 6, no. 2, pp. 113–118, 2023.

[17] N. Sunanto and G. Falah, "Penerapan Algoritma C4.5 Untuk Membuat Model Prediksi Pasien Yang Mengidap Penyakit Diabetes," *Rabit Jurnal Teknologi dan Sistem Informasi Univrab,* vol. 7, no. 2, pp. 208–216, 2022.

[18] I. M. Agus Oka Gunawan, I. D. A. Indah Saraswati, I. D. G. Riswana Agung, and I. P. Eka Putra, "Klasifikasi Penyakit Jantung Menggunakan Algoritma Decision Tree Series C4.5 Dengan Rapidminer," *Jurnal Teknologi dan Sistem Informasi Bisnis,* vol. 5, no. 2, pp. 73–83, 2023.

[19] A. Rahmat, M. Syafiih, and M. Faid, “Implementasi Klasifikasi Potensi Penyakit Jantung Dengan Menggunakan Metode C4.5 Berbasis Website (Studi Kasus Kaggle.Com),” INFOTECH J., vol. 9, no. 2, pp. 393–400, 2023.

[20] B. A. R. P. Wahyu, A. F. Farozi, C. P. Mahendra, and R. K. Hapsari, “Klasifikasi Penderita Penyakit Diabetes Berdasarkan Decision Tree,” INTEGER J. Inf. Technol., vol. 8, no. 1, pp. 80–89, 2023.

[21] A. Ardita, D. Permatasari, and R. M. Sholihin, “Diagnostik Urolithiasis” *Jurnal Farmasi dan Kesehatan.*, vol. 10, no. 1, pp. 35-46, 2021.

[22] C. Chazar and B. Erawan, “Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine,” *Inf. (Jurnal Informatika dan Sistem Informasi)*, vol. 12, no. 1, pp. 67–80, 2020.

[23] D. Kurniawan, Belajar Dari Data, *Pengenalan Machine Learning dengan Python*, Edisi Digital. Jakarta: PT Elex Media Komputindo, 2020.

[24] T. M. Mitchell, Well-Posed Learning Problems, *Machine Learning*. New York: McGraw-Hill Science/Engineering/Math, 1997.

[25] A. Géron, The Machine Learning Landscape, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*, Second Edition. Sebastopol : *O'Reilly Media*, Inc., 2019.

[26] J. C. Mestika, M. O. Selan, and M. I. Qadafi, “Menjelajahi Teknik-Teknik Supervised Learning untuk Pemodelan Prediktif Menggunakan Python,” *Buletin Ilmiah Ilmu Komputer dan Multimedia.*, vol. 99, no. 99, pp. 216–219, 2022.

[27] B. Pasaribu, A. Ahman, H. F. Muhtadi, S. F. Diba, N. Anggara, and W. Kanti, “Kesalahan Umum dalam Analisis Data: Data Normal dan Tidak Normal,” *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, vol. 7, no. 3, pp. 2413–2418, 2024.

[28] A. Purwanto, “Analisa Perbandingan Kinerja Algoritma C4.5 dan Algoritma K-Nearest Neighbors untuk Klasifikasi Penerima Beasiswa," *Jurnal Teknoinfo.*, vol. 17, no. 1, pp. 236–243, 2023.

LAMPIRAN

Lampiran 1

PEMERINTAH KABUPATEN MAJALENGKA
RUMAH SAKIT UMUM DAERAH CIDERES
Jalan Raya Cideres-Kadipaten No.180 Kecamatan Dawuan Kabupaten Majalengka 45453
Telp (0233) 661003-662082 Faks (0233) 662082 Pos-el rsudcideres@majalengkakab.go.id

Majalengka, 18 Maret 2025

Nomor : 000.9/13/DIKLATPEP--P4SDM/2025
Sifat : Biasa
Lampiran : -
Hal : **Jawaban Permohonan Penelitian Mahasiswa**

Yth.
Kepala Bagian Pelayanan Akademik
Pusat Telkom University Purwokerto

di
Tempat

Dipermaklumkan dengan hormat, menindaklanjuti surat pengantar dari Kepala Bagian Pelayanan Akademik Pusat Telkom University Purwokerto Tanggal 22 Januari 2025 Nomor : 284/AKD09/KP-WDI/2025 perihal Permohonan Izin Penelitian dengan nama :

Nama : Safina Octaviana Putri
NIM : 21110026

Maka dengan ini kami memberikan izin untuk melakukan penelitian yang berjudul "Penerapan algoritma decision tree dalam mendiagnosa penyakit batu ginjal berdasarkan data klinis" di RSUD Cideres Kabupaten Majalengka dengan ketentuan sebagai berikut :

1. Pelaksanaan Penelitian tidak mengganggu pelayanan di Unit Kerja tempat penelitian di lakukan;
2. Setelah melaksanakan kegiatan penelitian agar melaporkan hasil penelitian kepada Direktur RSUD Cideres Cq. Kepala Bagian Diklat Litbang dan PEP berbentuk *soft copy*.
3. Tidak mempublikasikan hasil penelitian tanpa seizin direktur RSUD Cideres;

Demikian surat ini kami sampaikan, atas kerjasamanya kami ucapkan terima kasih.

A.n Direktur RSUD Cideres
Wakil Direktur Administrasi Umum
dan Keuangan RSUD Cideres
Kabupaten Majalengka,

Isye Hendrayanti