Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

Perbandingan Kinerja Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Darah Tinggi

Ita Arfyanti[1,*], Tommy Bustomi[2], Ivan Haristyawan[1]

[1]Prodi Sistem Informasi, STMIK Widya Cipta Dharma, Samarinda, Indonesia
[2]Prodi Teknik Informatika Multimedia, Politeknik Negeri Samarinda, Samarinda, Indonesia
Email: ita@wicida.ac.id, tbustomi@gmail.com, ivan_haristyawan@yahoo.com
Email Penulis Korespondensi: ita@wicida.ac.id
Submitted: **19/12/2024**; Accepted: **29/12/2024**; Published**: 30/12/2024**

Abstrak−Penyakit darah tinggi atau hipertensi merupakan salah satu masalah kesehatan utama di dunia. Meskipun penyakit ini dapat diobati, banyak individu yang tidak menyadari bahwa mereka mengidap hipertensi, karena gejalanya seringkali tidak terlihat atau terasa. Oleh karena itu, deteksi dini penyakit darah tinggi sangat penting untuk pencegahan komplikasi serius yang dapat membahayakan kesehatan. Dalam era digital dan kemajuan teknologi informasi, banyak data kesehatan yang dapat digunakan untuk analisis. Salah satu pendekatan yang berkembang pesat untuk membantu diagnosis penyakit adalah dengan memanfaatkan data mining. Data mining adalah proses eksplorasi dan analisis data besar (big data) untuk menemukan pola, informasi, dan pengetahuan yang tersembunyi yang dapat digunakan untuk mendukung pengambilan keputusan dan prediksi. Salah satu teknik dalam data mining yang sering digunakan untuk memprediksi kondisi atau penyakit adalah algoritma klasifikasi. Namun, perbandingan kinerja antara algoritma-algoritma klasifikasi ini dalam konteks prediksi hipertensi masih terbatas. Penelitian ini bertujuan untuk mengeksplorasi dan membandingkan kinerja algoritma klasifikasi dalam memprediksi hipertensi, dengan menggunakan dataset yang berisi informasi medis tentang faktor-faktor yang mempengaruhi tekanan darah seseorang. Algoritma Naive Bayes adalah metode klasifikasi yang berbasis pada teorema Bayes dan asumsi independensi antar fitur. Algoritma C4.5 adalah algoritma pembelajaran mesin untuk membangun pohon keputusan yang digunakan dalam klasifikasi data. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan berbasis data mining yang dapat digunakan untuk mendeteksi dan memprediksi risiko hipertensi. nilai akurasi dari algoritma Naive Bayes sebesar 87,01% dan nilai akurasi dari algoritma C4.5 sebesar 94,72%. Dari proses yang telah dilakukan dapat dikatakan bahwasannya algoritma C4.5 merupakan algoritma dengan kinerja lebih baik dibandingkan algoritma Naive Bayes. Dengan demikian model yang dipergunakan dalam proses diagnosa penyakit hipertensi yaitu model dari algoritma C4.5

Kata Kunci: Penyakit Hipertensi; Data Mining; Kinerja; Algoritma Naive Bayes; Algoritma C4.5

Abstract−High blood pressure or hypertension is one of the major health problems in the world. Although this disease can be treated, many individuals are unaware that they have hypertension, because the symptoms are often not visible or felt. Therefore, early detection of high blood pressure is very important to prevent serious complications that can endanger health. In the digital era and advances in information technology, a lot of health data can be used for analysis. One of the rapidly developing approaches to help diagnose disease is by utilizing data mining. Data mining is the process of exploring and analyzing big data to find hidden patterns, information, and knowledge that can be used to support decision making and predictions. One technique in data mining that is often used to predict conditions or diseases is the classification algorithm. However, the comparison of performance between these classification algorithms in the context of hypertension prediction is still limited. This study aims to explore and compare the performance of classification algorithms in predicting hypertension, using a dataset containing medical information about factors that affect a person's blood pressure. The Naive Bayes algorithm is a classification method based on Bayes' theorem and the assumption of independence between features. The C4.5 algorithm is a machine learning algorithm for building decision trees used in data classification. The results of this study are expected to contribute to the development of a data mining-based decision support system that can be used to detect and predict the risk of hypertension. the accuracy value of the Naive Bayes algorithm is 87.01% and the accuracy value of the C4.5 algorithm is 94.72%. From the process that has been carried out, it can be said that the C4.5 algorithm is an algorithm with better performance than the Naive Bayes algorithm. Thus, the model used in the process of diagnosing hypertension is the model of the C4.5 algorithm

Keywords: Hypertension; Data Mining; Performance; Naive Bayes Algorithm; C4.5 Algorithm

1. PENDAHULUAN

Penyakit darah tinggi atau hipertensi merupakan salah satu masalah kesehatan utama di dunia. Menurut data Organisasi Kesehatan Dunia (WHO), hipertensi menjadi faktor risiko utama untuk terjadinya penyakit jantung, stroke, dan gagal ginjal, yang menyebabkan kematian lebih dari 9 juta orang setiap tahunnya. Meskipun penyakit ini dapat diobati, banyak individu yang tidak menyadari bahwa mereka mengidap hipertensi, karena gejalanya seringkali tidak terlihat atau terasa. Oleh karena itu, deteksi dini penyakit darah tinggi sangat penting untuk pencegahan komplikasi serius yang dapat membahayakan kesehatan [1]–[3].

Dalam era digital dan kemajuan teknologi informasi, banyak data kesehatan yang dapat digunakan untuk analisis. Salah satu pendekatan yang berkembang pesat untuk membantu diagnosis penyakit adalah dengan memanfaatkan *data mining*. *Data mining* adalah teknik yang digunakan untuk mengeksplorasi dan mengekstrak pola atau informasi yang berguna dari dataset besar. Dalam konteks ini, *data mining* dapat diterapkan untuk menganalisis faktor-faktor yang mempengaruhi terjadinya hipertensi dan untuk membuat prediksi yang akurat tentang risiko seseorang mengidap hipertensi [4]–[6].

Data mining adalah proses eksplorasi dan analisis data besar *(big data)* untuk menemukan pola, informasi, dan pengetahuan yang tersembunyi yang dapat digunakan untuk mendukung pengambilan keputusan dan prediksi. Proses

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

ini melibatkan penggunaan teknik statistik, algoritma, dan machine learning untuk mengekstrak informasi yang berguna dari dataset yang besar dan kompleks. Dalam penerapannya, data mining sangat berguna di berbagai bidang, seperti kesehatan, pemasaran, keuangan, dan manufaktur, untuk mengoptimalkan proses bisnis, memprediksi tren, atau meningkatkan keputusan strategis [7]–[9].

Salah satu teknik dalam *data mining* yang sering digunakan untuk memprediksi kondisi atau penyakit adalah algoritma klasifikasi. Algoritma klasifikasi memungkinkan sistem untuk mengklasifikasikan individu ke dalam kategori tertentu, seperti "berisiko hipertensi" dan "tidak berisiko hipertensi", berdasarkan data yang dimiliki. Berbagai algoritma klasifikasi dapat digunakan dalam hal ini, seperti *Decision Tree* (DT), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN), *Random Forest* (RF), dan *Naive Bayes* (NB). Setiap algoritma memiliki kelebihan dan kekurangannya sendiri, tergantung pada karakteristik data yang digunakan, seperti ukuran dataset, jenis fitur, dan kompleksitas pola yang ada[10]–[14].

Namun, perbandingan kinerja antara algoritma-algoritma klasifikasi ini dalam konteks prediksi hipertensi masih terbatas. Beberapa algoritma mungkin lebih efektif dalam menangani data yang tidak seimbang atau memiliki data yang banyak variabelnya, sementara yang lain lebih unggul dalam hal interpretabilitas atau waktu komputasi. Oleh karena itu, penting untuk melakukan penelitian yang membandingkan berbagai algoritma klasifikasi guna mengetahui mana yang memiliki performa terbaik dalam memprediksi penyakit darah tinggi, dengan mempertimbangkan metrik-metrik evaluasi seperti akurasi, *precision*, *recall*, dan *F1-score* [15]–[17].

Penelitian ini bertujuan untuk mengeksplorasi dan membandingkan kinerja algoritma klasifikasi dalam memprediksi hipertensi, dengan menggunakan dataset yang berisi informasi medis tentang faktor-faktor yang mempengaruhi tekanan darah seseorang. Melalui perbandingan ini, diharapkan dapat ditemukan algoritma yang paling efektif dan efisien dalam memprediksi penyakit darah tinggi, yang pada akhirnya dapat mendukung upaya deteksi dini dan pencegahan hipertensi secara lebih luas. Pada penelitian ini algoritma yang digunakan untuk proses penyelesaian perbandingan adalah algoritma *Naive Bayes* dan algoritma C4.5.

Algoritma *Naive Bayes* adalah metode klasifikasi yang berbasis pada teorema Bayes dan asumsi independensi antar fitur. Algoritma ini digunakan untuk memprediksi kategori atau kelas dari data berdasarkan probabilitas, dengan menghitung kemungkinan suatu kelas berdasarkan fitur-fitur yang ada. *"Naive"* merujuk pada anggapan bahwa semua fitur saling independen, meskipun dalam kenyataannya fitur-fitur tersebut mungkin saling bergantung. Sebagai dasar pertimbangan maka diperlukan beberapa referensi terdahulu berkaitan dengan algoritma *Naive Bayes* [18][19].

Penelitian yang dilakukan oleh Maulana Fansyuri dan Devi Yunita pada tahun 2024 dengan judul penelitian Analisa Citra Wajah Untuk Identifikasi Klasifikasi Jenis Kelamin Menggunakan Algoritma Naive Bayes didapatkan hasil penelitian bahwa bahwa metode ini termasuk dalam algoritma yang sangat baik untuk diaplikasikan pada klasifikasi citra wajah berdasarkan warna dan bentuk dengan nilai akurasi sebesar 80%, sehingga penentuan jenis kelamin berdasarkan objek wajah menggunakan data hasil ekstraksi warna dan bentuk serta menggunakan metode klasifikasi Naïve Bayes sesuai dengan data citra sebenarnya[20].

Penelitian lainnya yang dilakukan oleh Nur Fidiyanto dan Afifah Nurul Izzati pada tahun 2024 dengan judul penelitian Penerapan Data Mining Klasifikasi Lahan Tanam Buah Alpukat dengan Algoritma Naïve Bayes mendapatkan hasil dari penelitian Pada penelitian ini diperoleh hasil pada lahan KTH Pedunung Lestari Sejahtera dengan ketinggian lahan 250Mdpl, suhu 18°C, curah hujan 25mm/hari dan jenis tanah humus, lebih cocok ditanami tanaman alpukat jenis miki dibanding jenis shepard. Berdasarkan hasil perhitungan pada alpukat miki diperoleh nilai “Ya” sebesar 0,75 dan “Tidak” sebesar 0,25 sedangkan jenis shepard nilai “Ya” sebesar 0 dan “Tidak” sebesar 1. Adapun nilai accuracy 50%, Precision 43% dan Recall 100%[21].

Selain itu, juga dilakukan penelitian oleh M. Julkarnain dan Mar’i Yustiardin pada tahun 2024 dengan judul penelitian Penerapan Algoritma Naive Bayes Dalam Memprediksi Lulus Tepat Waktu Mahasiswa serta hasil dari penelitian Hasil penelitian menunjukkan bahwa algoritma ini mampu memberikan prediksi yang signifikan dalam menentukan mahasiswa yang berpotensi lulus tepat waktu yang memiliki akurasi 94,31%, precision 91%, recall 95%, dan F-1 Score 93%, sehingga menghasilkan model yang dapat memprediksi kelulusan tepat waktu mahasiswa dengan mutakhir[22].

Penelitian terakhir digunakan sebagai referensi penelitian yang dilakukan oleh Fajar Ramadhan dan Henny Dwi Bhakti pada tahun 2024 dengan judul penelitian Klasifikasi Penilaian Kinerja Karyawan Menggunakan Algoritma Naïve Bayes (Studi Kasus PT. As Sabar Sukses Berkah) dimana hasil dari penelitian Hasil penelitian menunjukkan tingkat akurasi klasifikasi yang didapat dari Algoritme Naïve Bayes cukup tinggi sebesar 90%[23].

Algoritma C4.5 adalah algoritma pembelajaran mesin untuk membangun pohon keputusan yang digunakan dalam klasifikasi data. C4.5 memilih atribut terbaik untuk membagi data berdasarkan Information Gain yang diukur dengan Gain Ratio, untuk memaksimalkan pemisahan kelas yang berbeda dalam data. Beberapa penelitian yang telah dilakukan berkaitan dengan algoritma C4.5 sebagai dasar terhadap pelaksanaan penelitian[24], [25].

Penelitian pertama yang dilakukan oleh Wendy Asswan Cahyadi, dkk pada tahun 2024 dengan judul dari penelitian Penerapan Data Mining Dalam Penentuan Jurusan Siswa Dengan Metode Klasifikasi Algoritma C4 . 5 Studi Kasus SMAN 1 Leuwisadeng dimana dari penelitian didapatkan hasil berupa Hasil dari akurasi yang didapat pada program ini adalah sebesar 92,31%. Dengan demikian dapat disimpulkan bahwa data mining dengan metode klasifikasi Algoritma C4.5 dapat mempercepat proses penjurusan siswa[26].

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

Penelitian kedua yang dilakukan pada penelitian oleh Nenitrolina Ndruru dan Anita Sindar pada tahun 2024 dengan judul penelitian Penerapan Data Mining Klasifikasi Kepuasan Pelanggan Transportasi Online Menggunakan Algoritma C4.5 dimana dari penelitian yang dilakukan didapatkan hasil Aplikasi data mining untuk mengklasifikasikan kepuasan pelanggan pengguna transportasi online telah berhasil diimplementasikan pada sistem dengan menerapkan Algoritma C4.5 hasil klasifikasi kepuasan pelanggan pengguna transportasi online, jumlah pelanggan pengguna transportasi online dengan klasifikasi "Tidak Puas" sebanyak 91 orang pelanggan, sedangkan yang "Tidak Puas" sebanyak 9 orang pelanggan[27].

Penelitian selanjutnya yang dilakukan oleh Revaldo Xsanal Hakim, dkk pada tahun 2024 dengan judul penelitian Penerapan Algoritma C4.5 Untuk Prediksi Anak Stunting Di Kota Pagar Alam dimana hasil dari penelitian bahwa dapat disimpulkan bahwa akurasi Algoritma C.4.5 untuk memPrediksi anak Stunting yaitu 88,20% tergolong baik[28].

Penelitian terakhir yang dilakukan pada penelitian oleh Dandi Muhamad Musa, dkk pada tahun 2024 dengan judul penelitian Penerapan Data Mining Untuk Klasifikasi Data Penjualan Pakan Ternak Terlaris Dengan Algoritma C4.5 dimana pada penelitian didapatkan hasil Hasil penelitian menunjukkan gain tertinggi terdapat pada kategori pakan dengan nilai 0.306739968 dan entropy pakan ayam pedaging dengan nilai 0.99107606. Hal ini menunjukkan bahwa pakan ayam pedaging merupakan produk paling laris berdasarkan hasil pengolahan data[29].

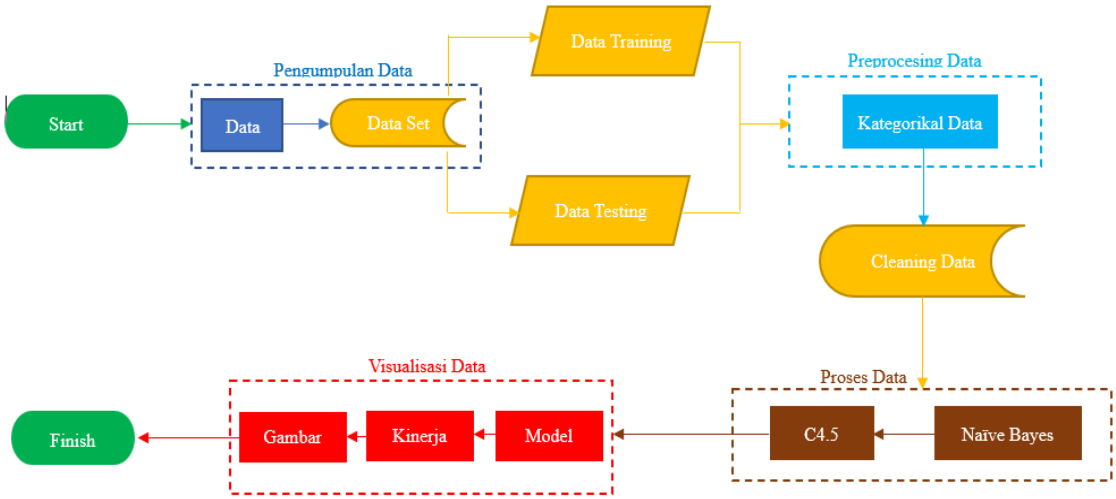
Membandingkan kinerja berbagai algoritma klasifikasi dalam memprediksi hipertensi. Menentukan algoritma klasifikasi yang paling efektif untuk digunakan dalam sistem prediksi penyakit darah tinggi. Menyediakan wawasan mengenai faktor-faktor yang mempengaruhi kinerja algoritma klasifikasi dalam prediksi penyakit darah tinggi.

Hasil dari penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan berbasis data mining yang dapat digunakan untuk mendeteksi dan memprediksi risiko hipertensi. Sistem seperti ini dapat digunakan oleh tenaga medis atau bahkan masyarakat untuk melakukan deteksi dini dan pencegahan hipertensi, sehingga dapat mengurangi angka morbiditas dan mortalitas yang disebabkan oleh hipertensi dan komplikasinya.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Pada tahapan penelitian di bawah ini merupakan alur dari tahapan penelitian serta proses pengumpulan data yang dilakukan, adalah sebagai berikut:



Gambar 1. Tahapan Penelitian

2.2 *Data Mining*

Data mining adalah proses eksplorasi dan analisis data besar untuk menemukan pola, hubungan, atau informasi yang berguna yang dapat digunakan untuk pengambilan keputusan atau prediksi. Teknik ini menggunakan algoritma dan metode statistik untuk menggali data dan menghasilkan pengetahuan yang tersembunyi di dalamnya. *Data mining* sering diterapkan dalam berbagai bidang seperti bisnis, kesehatan, dan keuangan untuk memprediksi tren, mengelompokkan data, atau mendeteksi anomali[13], [15], [24].

Data mining merupakan Proses mengungkap pola yang sebelumnya tidak diketahui dalam data besar, dengan menggunakan teknik statistik, algoritma, dan metode pembelajaran mesin *(machine learning)*, untuk mengidentifikasi informasi yang berguna dan dapat diterapkan dalam pengambilan keputusan. Dengan kata lain, data mining adalah suatu pendekatan untuk menemukan informasi yang tersembunyi dan berguna dalam dataset besar melalui analisis

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

yang mendalam. Proses ini melibatkan langkah-langkah seperti pengumpulan data, pemrosesan, dan ekstraksi pola yang relevan[20], [23], [27].

2.3 Algoritma *Naïve Bayes*

Algoritma *Naive Bayes* adalah salah satu metode klasifikasi dalam *data mining* yang berdasarkan pada prinsip probabilitas *Bayes* dan asumsi independensi antar fitur. Algoritma ini digunakan untuk memprediksi kategori atau kelas dari data berdasarkan atribut atau fitur yang dimilikinya[20], [21].

Secara umum, algoritma *Naive Bayes* bekerja dengan menghitung probabilitas posterior dari setiap kelas, dan memilih kelas yang memiliki probabilitas terbesar. Algoritma ini disebut *"naive"* (naif) karena mengasumsikan bahwa semua fitur yang digunakan untuk klasifikasi saling independen atau tidak saling bergantung satu sama lain, yang seringkali tidak realistis dalam kehidupan nyata. Meskipun demikian, meskipun asumsi ini mungkin tidak selalu berlaku, *Naive Bayes* sering kali memberikan hasil yang baik, terutama ketika data memiliki banyak fitur. Adapun langkah kerja dari algoritma *Naive Bayes* adalah[22], [23]:

a. Menghitung Probabilitas Prior (P(class)): Probabilitas awal atau distribusi kelas sebelum mempertimbangkan fitur. Ini dihitung berdasarkan jumlah contoh dari setiap kelas dalam dataset pelatihan.
b. Menghitung Probabilitas Likelihood (P(feature|class)): Probabilitas dari setiap fitur yang muncul dalam kelas tertentu. Probabilitas ini dihitung berdasarkan frekuensi fitur dalam data yang ada.
c. Menghitung Probabilitas Posterior (P(class|features)): Dengan menggunakan teorema Bayes, kita menghitung probabilitas kelas untuk suatu data berdasarkan fitur yang dimilikinya. Formula teorema Bayes adalah:

$$P(class|features) = \frac{P(feature|class) * P(class)}{P(features)} \quad (1)$$

Di sini, P(class|features)P(class|features)P(class|features) adalah probabilitas posterior (probabilitas kelas berdasarkan fitur), P(features|class)P(features|class)P(features|class) adalah probabilitas likelihood (kemungkinan fitur terjadi dalam kelas), dan P(class)P(class)P(class) adalah probabilitas prior (probabilitas kelas dalam dataset).
d. Klasifikasi: Setelah menghitung probabilitas posterior untuk setiap kelas, Naive Bayes memilih kelas dengan probabilitas tertinggi sebagai hasil klasifikasi.

2.4 Algoritma C4.5

Algoritma C4.5 adalah algoritma klasifikasi yang digunakan dalam data mining untuk menghasilkan pohon keputusan *(decision tree)*. Pohon keputusan ini digunakan untuk memetakan input (fitur) ke dalam kelas atau kategori yang sesuai. C4.5 dikembangkan oleh Ross Quinlan pada tahun 1993 sebagai pengembangan dari algoritma sebelumnya yang disebut ID3 (Iterative Dichotomiser 3). C4.5 sering digunakan karena kemampuannya menghasilkan model yang mudah dipahami, serta memiliki beberapa keunggulan dalam hal pengolahan data yang lebih baik dan pengurangan overfitting. C4.5 membangun pohon keputusan berdasarkan pembagian data yang optimal. Algoritma ini bekerja dengan langkah-langkah berikut[26], [28]:

a. Pemilihan Atribut Terbaik: Algoritma C4.5 memilih atribut yang paling relevan untuk membagi data pada setiap langkah pembentukan pohon keputusan. Pemilihan atribut ini dilakukan berdasarkan rasio informasi *(Information Gain Ratio)*, yang merupakan ukuran untuk menilai kualitas pembagian data. Dalam C4.5, rasio informasi digunakan untuk memilih atribut yang paling informatif untuk memisahkan data menjadi subset yang lebih homogen.
b. Membangun Cabang Pohon: Setelah atribut terbaik dipilih, data dibagi berdasarkan nilai atribut tersebut, dan cabang pohon dibuat untuk setiap nilai yang mungkin dari atribut tersebut. Proses ini berulang pada setiap cabang hingga seluruh data terklasifikasi.
c. Penyusunan Daun *(Leaf Node)*: Setiap daun pohon keputusan mewakili kelas atau kategori akhir yang diprediksi. Daun ini berisi hasil klasifikasi berdasarkan distribusi data yang ada pada cabang tersebut.
d. Pemangkasan *(Pruning)*: Salah satu fitur penting dari C4.5 adalah pemangkasan pohon *(pruning)*. Pemangkasan dilakukan untuk mengurangi kompleksitas model dan menghindari *overfitting*, yaitu ketika model terlalu cocok dengan data pelatihan dan tidak dapat menggeneralisasi dengan baik pada data baru. C4.5 melakukan pemangkasan dengan cara menghapus cabang yang tidak memberikan kontribusi signifikan terhadap akurasi prediksi.

Algoritma C4.5 adalah algoritma untuk membangun pohon keputusan dengan menggunakan Informasi Gain Ratio sebagai ukuran untuk memilih atribut terbaik dalam mempartisi data. Rumus utama yang digunakan dalam C4.5 untuk memilih atribut terbaik terdiri dari dua konsep utama: Information Gain dan Split Information. Berikut adalah penjelasan dan rumus-rumus terkait[27], [29]:

a. Informasi Gain digunakan untuk mengukur pengurangan ketidakpastian (entropy) setelah membagi data berdasarkan atribut tertentu. Entropy mengukur ketidakpastian atau ketidakteraturan dalam dataset. Rumus Entropy untuk suatu set data S dengan kelas C adalah:

$$Entropy(S) = -\sum_{i=1}^{k} p_i log_2(p_i) \quad (2)$$

pi adalah probabilitas relatif dari kelas iii dalam set data S dan k adalah jumlah kelas dalam dataset.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

Information Gain untuk atribut A adalah pengurangan Entropy setelah membagi dataset berdasarkan nilai-nilai atribut A. Rumusnya adalah:

$$IG(S,A) = Entropy(S) - \sum \frac{|S_v|}{|S|} * Entropy(S_v) \quad (3)$$

Sv adalah subset data yang memiliki nilai v pada atribut A, |Sv| adalah jumlah elemen dalam subset Sv dan |S'| adalah jumlah total elemen dalam dataset S serta Values(A) adalah himpunan nilai yang mungkin untuk atribut A.

b. Split Information mengukur informasi yang diperoleh dari membagi dataset berdasarkan atribut. Ini penting untuk mencegah atribut yang memiliki banyak nilai yang tidak relevan mendominasi pemilihan. Rumus Split Information untuk atribut A adalah:

$$SI(S,A) = \sum \frac{|S_v|}{|S|} \, log_2 \, (\frac{|S_v|}{|S|}) \quad (4)$$

Sv adalah subset data yang memiliki nilai v pada atribut A, |Sv| adalah jumlah elemen dalam subset Sv dan |S'| adalah jumlah total elemen dalam dataset S serta Values(A) adalah himpunan nilai yang mungkin untuk atribut A.

c. Untuk memilih atribut terbaik, C4.5 menggunakan Gain Ratio yang mengkombinasikan Information Gain dan Split Information. Gain Ratio digunakan untuk mengatasi masalah jika atribut dengan banyak nilai mendominasi pembagian data, meskipun mungkin tidak memberikan informasi yang relevan. Rumus Gain Ratio untuk atribut A adalah:

$$Gain\ Ratio\ (S,A) = \frac{IG\ (S,A)}{SI\ (S,A)} \quad (5)$$

3. HASIL DAN PEMBAHASAN

3.1 Analisa Masalah

Penyakit darah tinggi atau biasa disebut hipertensi merupakan salah satu penyakit utama di dunia serta khususnya di Indonesia. Penyakit darah tinggi tersebut merupakan salah satu penyakit yang memicu terhadap komplikasi pada penyakit lainnya. Penyakit tersebut dapat ditangani dengan baik, namun sering kali masyarakat lalai dan tidak mengetahui jika terkena penyakit tersebut. Oleh sebab itu perlu kiranya diketahui sejak dini terhadap penyakit darah tinggi tersebut. Penyakit tersebut dapat diketahui dengan melakukan prediksi. Dimasa sekarang ini penyelesaian terhadap perasalahan kesehatan sudah dapat diselesaikan dengan memanfaatkan teknologi informasi. Dimana salah satu cara yang dapat dilakukan adalah dengan menggunakan data mining. Data mining sendiri merupakan salah satu cara yang digunakan untuk menyelesaikan proses berdasarkan dengan data-data dimasa lampau. Dimana proses penyelesaian pada penelitian dapat dilakukan dengan memperhatikan data - data penyakit darah tinggi sebelumnya untuk kemudian dilakukan proses kembali untuk melakukan pengambilan keputusan. Pada data mining proses penyelesaian dapat dilakukan dengan menggunakan algoritma Naive Bayes dan algoritma C4.5 dimana kedua algoritma tersebut termasuk dalam teknik klasifikasi data minig. Perbandingan kinerja klasifikasi perlu dilakukan untuk mengetahui terhadap kinerja algortima mana yang lebih baik yang dimana nantinya hasil yang didapatkan dipergunakan dalam proses pengambilan keputusan.

Pada peneltian ini, data yang digunakan merupakan dataset penyakit hipertensi yang dimana dataset didapatkan dari UCI Repository *Machine Learning*. Pada dataset tersebut terdapat 9 atribut yaitu Usia, Jenis Kelamin, Merokok, Bekerja, Berumah Tangga, Aktivitas Bergadang, Aktivitas Berolahraga, Asuransi serta Penyakit Bawaan. Pada dataset juga terdapat 2 nilai kelas yaitu Ya dan Tidak. Selain atribut pada dataset juga terdapat jumlah data atau record, adapun jumlah data pada dataset sebanyak 10000 data. Adapun sampel pada dataset penyakit darah tinggi dapat dilihat berikut:

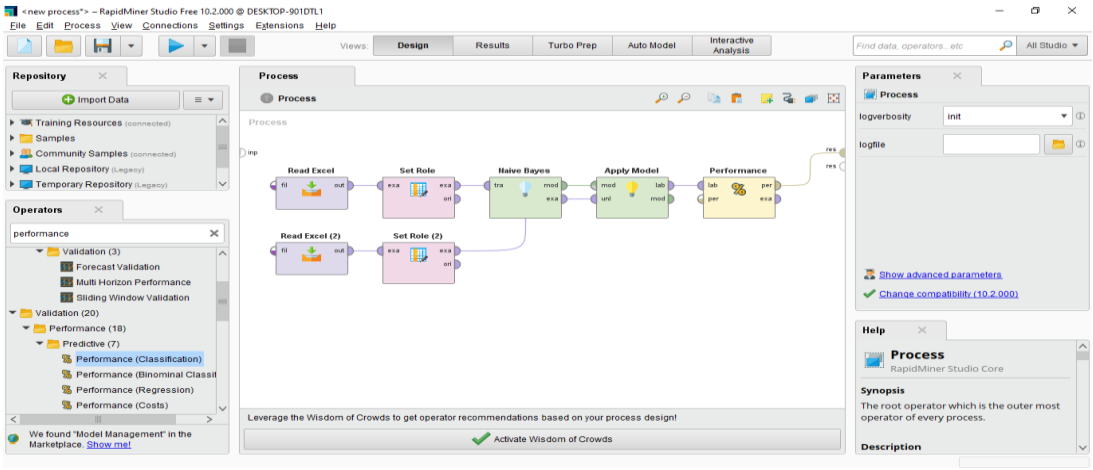
Tabel 1. Sampel Dataset Darah Tinggi

No	Usia	Jenis_Kelamin	Merokok	Bekerja	Rumah_Tangga	Aktivitas_Begadang	Aktivitas_Olahraga	Asuransi	Penyakit_Bawaan	Hasil
1	Tua	Pria	Pasif	Tidak	Ya	Ya	Sering	Ada	Tidak	Ya
2	Tua	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Ada	Tidak
3	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
4	Tua	Pria	Aktif	Ya	Tidak	Tidak	Jarang	Ada	Ada	Tidak
5	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Ya
6	Muda	Wanita	Pasif	Ya	Tidak	Tidak	Sering	Tidak	Ada	Tidak
7	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak	Ya
8	Muda	Pria	Aktif	Tidak	Ya	Ya	Sering	Tidak	Tidak	Tidak
9	Tua	Wanita	Aktif	Ya	Ya	Ya	Jarang	Ada	Ada	Ya
10	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada	Ya
11	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada	Ya
12	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak	Tidak
13	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
14	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada	Tidak
15	Muda	Wanita	Pasif	Ya	Tidak	Ya	Sering	Tidak	Ada	Ya
16	Muda	Wanita	Pasif	Ya	Tidak	Ya	Jarang	Ada	Ada	Ya
17	Tua	Wanita	Pasif	Ya	Ya	Tidak	Sering	Ada	Ada	Ya
18	Tua	Wanita	Aktif	Tidak	Ya	Tidak	Jarang	Ada	Tidak	Tidak
19	Muda	Pria	Aktif	Tidak	Ya	Ya	Jarang	Ada	Tidak	Tidak
20	Tua	Wanita	Aktif	Ya	Tidak	Ya	Jarang	Ada	Ada	Tidak
…	…	…	…	…	…	…	…	…	…	…
10000	Tua	Wanita	Pasif	Tidak	Ya	Tidak	Sering	Tidak	Tidak	Ya

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

3.2 Penerapan Algoritma Naive Bayes

Setelah didapatkan dataset pada penelitian maka proses selanjutnya adalah dilakukan proses pengujian kinerja terhadap algoritma data minig. Proses pertama yang dilakukan pada penelitian yaitu dengan melakukan pengukuran kinerja terhadap algoritma Naive Bayes. Adapun proses yang dilakukan pada proses pengujian dapat dilihat pada gambar berikut:



Gambar 2. Proses Pengujian Algoritma Naive Bayes

Pada Gambar 2 diatas merupakan proses operator pengujian yang dilakukan dengan menggunakan tools rapid miner. Dari proses tersebut nantinya didapatkan hasil pengukuran kinerja. Adapun hasil kinerja dari algoritma Naive Bayes dapat dilihat berikut:

accuracy: 87.01%

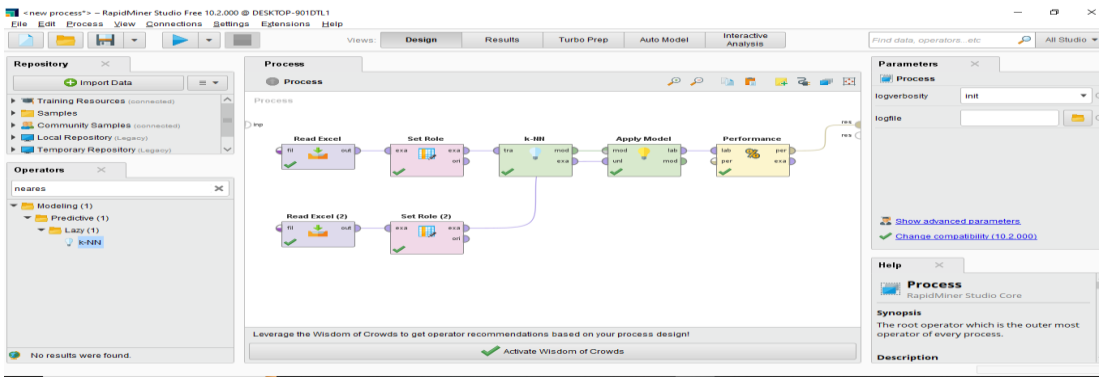
	true Ya	true Tidak	class precision
pred. Ya	4125	651	86.37%
pred. Tidak	648	4576	87.60%
class recall	86.42%	87.55%	

Gambar 3. Hasil Kinerja Algoritma Naive Bayes

Dari Gambar 3 diatas dapat dilihat bahwasannya hasil kinerja ataupun tingkat akurasi dari algoritma Naive Bayes sebesar 87,01% dengan Class_Recall_Ya 86,42%, Class_Recall_Tidak 87,55%, Class_Precision_Ya 86,37% dan Class_Precission_Tidak 87,60%.

3.3 Penerapan Algoritma C4.5

Setelah dilakukan proses pengujian dengan menggunakan algoritma Naive Bayes serta didapatkan hasil kinerja maka proses selanjutnya adalah dilakukan proses pengujian kinerja terhadap algoritma C4.5. Proses kedua yang dilakukan pada penelitian yaitu dengan melakukan pengukuran kinerja terhadap algoritma C4.5. Adapun proses yang dilakukan pada proses pengujian dapat dilihat pada gambar berikut:



Gambar 4. Proses Pengujian Algoritma Naive Bayes

Pada Gambar 4 diatas merupakan proses operator pengujian yang dilakukan dengan menggunakan tools rapid miner. Dari proses tersebut nantinya didapatkan hasil pengukuran kinerja. Adapun hasil kinerja dari algoritma C4.5 dapat dilihat berikut:

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

accuracy: 94.72%

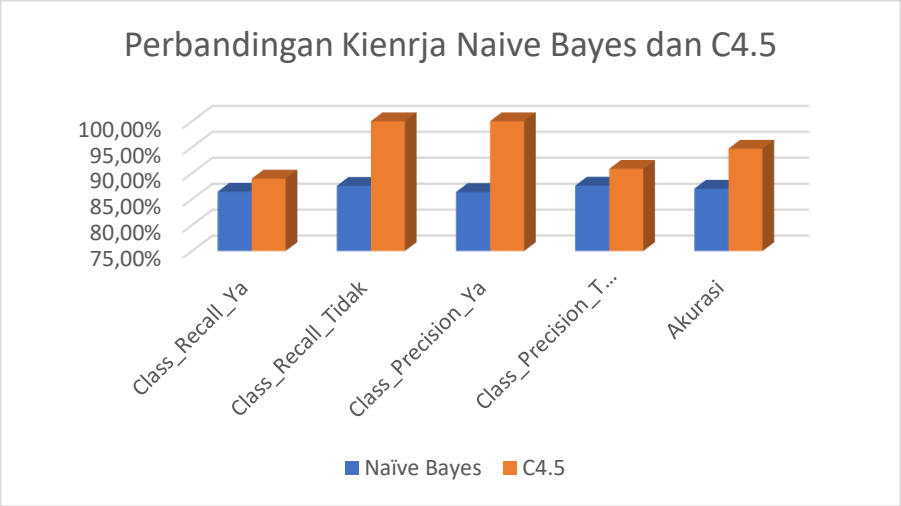
	true Ya	true Tidak	class precision
pred. Ya	4245	0	100.00%
pred. Tidak	528	5227	90.83%
class recall	88.94%	100.00%	

Gambar 5. Hasil Kinerja Algoritma C4.5

Dari Gambar 5 diatas dapat dilihat bahwasannya hasil kinerja ataupun tingkat akurasi dari algoritma Naive Bayes sebesar 94,72% dengan Class_Recall_Ya 88,94%, Class_Recall_Tidak 100%, Class_Precision_Ya 100% dan Class_Precission_Tidak 90,83%.

3.4 Pembahasan

Setelah dilakukan proses pengujian terhadap algoritma Naive Bayes dan C4.5 serta didapatkan hasil kinerja dari setiap algoritma maka dapat dilakukan proses pembahasan dalam penentuan hasil akhir dari penelitian. Hasil kinerja dari algoritma dapat dilihat berikut:



Gambar 6. Hasil Perbandingan Kinerja Algoritma Naive Bayes dan C4.5

Pada Gambar 6 diatas dapat dilihat hasil dari perbandingan kinerja dari algoritma Naive Bayes dan C4.5. Pada gambar tersebut dapat dilihat bahwasannya untuk nilai akurasi dari algoritma Naive Bayes sebesar 87,01% dan nilai akurasi dari algoritma C4.5 sebesar 94,72%. Dari proses yang telah dilakukan dapat dikatakan bahwasannya algoritma C4.5 merupakan algoritma dengan kinerja lebih baik dibandingkan algoritma Naive Bayes. Dengan demikian model yang dipergunakan dalam proses diagnosa penyakit hipertensi yaitu model dari algoritma C4.5

4. KESIMPULAN

Proses akhir dari penelitian merupakan penarikan sebuah kesimpulan. Adapun kesimpulan yang didapatkan dari penelitian bahwasannya dengan menggunakan teknologi informasi dapat mempercepat proses untuk diagnosa terhadap penyakit darah tinggi. Penerapan data mining dalam proses pengolahan data kiranya dapat membantu mempermudah dalam mendapatkan informasi dalam dataset yang memiliki jumlah data yang besar. Dari proses yang dilakukan tingkat akurasi algoritma Naive Bayes sebesar 87,01% dengan Class_Recall_Ya 86,42%, Class_Recall_Tidak 87,55%, Class_Precision_Ya 86,37% dan Class_Precission_Tidak 87,60%. Sedangkan tingkat akurasi dari algoritma Naive Bayes sebesar 94,72% dengan Class_Recall_Ya 88,94%, Class_Recall_Tidak 100%, Class_Precision_Ya 100% dan Class_Precission_Tidak 90,83%. Dari proses yang telah dilakukan dapat dikatakan bahwasannya algoritma C4.5 merupakan algoritma dengan kinerja lebih baik dibandingkan algoritma Naive Bayes. Dengan demikian model yang dipergunakan dalam proses diagnosa penyakit hipertensi yaitu model dari algoritma C4.5

REFERENCES

[1] A. Az, R. Septian, M. A. Saktiawan, and R. A. Saputra, "PREDIKSI PENYAKIT HIPERTENSI MENGGUNAKAN MACHINE LEARNING DENGAN ALGORITMA REGRESI LOGISTIK," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 6, pp. 12062–12068, 2024.

[2] A. S. Novari and U. K. Nisak S, "Prediksi Faktor yang Mempengaruhi Hipertensi dengan Metode Data Mining untuk meningkatkan Pelayanan Kesehatan di UPT Puskesmas Ngoro," *Phys. Sci. Life Sci. and Engineering*, vol. 1, no. 2, p. 16,

Building of Informatics, Technology and Science (BITS)
Volume 6, No 3, Desember 2024 Page: 1987−1994
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i3.6477

2024, doi: 10.47134/pslse.v1i2.201.
[3] R. Sahila, T. Widiharih, and I. T. Utami, "ANALISIS KLASIFIKASI MENGGUNAKAN REGRESI LOGISTIK BINER DAN ALGORITMA NAÏVE BAYES CLASSIFIER PADA PENYAKIT HIPERTENSI," *J. GAUSSIAN*, vol. 13, no. 2007, pp. 319–327, 2024, doi: 10.14710/j.gauss.13.2.319-327.
[4] R. A. Anggraini, Ratningsih, Y. Apriyani, M. W. Pertiwi, M. Kusmira, and S. Bahri, "Klasifikasi Jenis Kismis Menggunakan Teknik Data Mining," *J. Kaji. Ilm.*, vol. 24, no. 1, pp. 45–56, 2024, doi: 10.31599/ryvqk945.
[5] Wartumi, R. Kurniawan, and A. Y. Wijaya, "Analisis Data Sentimen Ulasan Pengguna Aplikasi Shopee di Google Play Store dengan Klasifikasi Algoritma Naïve Bayes," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 164–170, 2024.
[6] W. Bagaskara, N. N. Pusparini, and Irwansyah, "KLASIFIKASI PENJADWALAN KERJA PERAWATAN AIR CONDITIONER (AC) MENGGUNAKAN ALGORITMA DECISION TREE (C4.5) PADA PT XYZ," *INFOTECH J. Technol. Inf.*, vol. 10, no. 1, pp. 67–76, 2024.
[7] H. T. Santoso, F. A. Felmidi, A. Nur, A. Ristyawan, and E. Daniati, "Analisis Kinerja Algoritma Data Mining pada Klasifikasi Tingkat Obesitas dengan K-Fold Cross Validation dan AUC," *INOTEK*, vol. 8, pp. 113–122, 2024.
[8] N. Anthira and Suendri, "Penerapan Data Mining Pada Klasifikasi Gangguan Jiwa Menggunakan Algoritma C5.0 Di RSJ. Mahoni Kota Medan," *J. Tek.*, vol. 18, no. x, pp. 571–582, 2024.
[9] R. Hamonangan, R. K. Sari, S. Anwar, and T. Hartati, "Klasifikasi Algoritma KNN dalam menentukan Penerima Bantuan Langsung Tunai," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, pp. 198–204, 2024.
[10] F. M. Siddiq, R. D. Aditya, and M. N. Hamidah, "Klasifikasi Penerima Zakat Fitrah Menggunakan Metode Naïve Bayes," *J. Electr. Eng. Comput. Sci.*, vol. 6, no. August, p. 128, 2024, [Online]. Available: http://eprints.ubhara.ac.id/908/.
[11] A. Setiawan, Z. H. Nst, Z. Khairi, and L. Efrizoni, "KLASIFIKASI TINGKAT RISIKO DIABETES MENGGUNAKAN ALGORITMA," *JIRE (Jurnal Inform. Rekayasa Elektron.*, vol. 7, no. 2, pp. 263–271, 2024.
[12] B. Susilo, N. A. Ramdhan, and O. S. Bachri, "Application of the K-Nearest Neighbor Algorithm for Predicting Digital Product Sales Penerapan Algoritma K-Nearest Neighbor untuk Prediksi Penjualan Produk Digital," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. October, pp. 1466–1476, 2024.
[13] S. Dini Widianti *et al.*, "Jurnal Informatika dan Rekayasa Perangkat Lunak Menentukan Nilai Gizi pada Balita Menggunakan Algoritma Support Vektor Machine (SVM) di Posyandu Kelurahan Ciherang," *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 6, no. 1, 2024.
[14] A. R. Raharja, Jayadi, A. Pramudianto, and Y. Muchsam, "Penerapan Algoritma Decision Tree dalam Klasifikasi Data 'Framingham' Untuk Menunjukkan Risiko Seseorang Terkena Penyakit Jantung dalam 10 Tahun Mendatang," *Technol. J.*, vol. 1, no. 1, 2024, doi: 10.62872/cwgzp962.
[15] W. A. Ridho, T. Wuryandari, and A. R. Hakim, "Perbandingan Kinerja Metode Klasifikasi K-Nearest Neighbor Dan Support Vector Machines Pada Dataset Parkinson," *J. Gaussian*, vol. 12, no. 3, pp. 372–381, 2024, doi: 10.14710/j.gauss.12.3.372-381.
[16] A. Masruriyah, H. Novita, C. Sukmawati, A. Ramadhan, S. Arif, and B. Dermawan, "Pengukuran Kinerja Model Klasifikasi dengan Data Oversampling pada Algoritma Supervised Learning untuk Penyakit Jantung," *Comput. Sci.*, vol. 4, no. 1, pp. 62–70, 2024, doi: 10.31294/coscience.v4i1.2389.
[17] N. I. Yaman, A. R. Juwita, S. Arum, P. Lestari, and S. Faisal, "Perbandingan Kinerja Algoritma Decision Tree dan Random Forest untuk Klasifikasi Nutrisi pada Makanan Cepat Saji," *J. Algoritm.*, vol. 21, no. 2, pp. 184–195, 2024, doi: 10.33364/algoritma/v.21-2.1649.
[18] M. R. Fanani and D. S. Sintia, "Klasifikasi Kesiapan Anak Masuk Sekolah Dasar menggunakan Algoritma Naïve Bayes dan Algoritma C4.5," *Innov. J. Soc. Sci. Res.*, vol. 4, no. 3, pp. 10547–10555, 2024, doi: 10.31004/innovative.v4i3.10425.
[19] Sriani, I. Rusydi, and S. R. Nur Aisyiyah, "Impelementasi Data Mining terhadap Evaluasi Kinerja Guru dalam Mengajar Menggunakan Metode Naive Bayes Classifier," *VISA J. Vis. Ideas*, vol. 4, no. 1, pp. 117–127, 2024, doi: 10.47467/visa.v4i1.1274.
[20] M. Fansyuri and D. Yunita, "Analisa Citra Wajah Untuk Identifikasi Klasifikasi Jenis Kelamin Menggunakan Algoritma Naive Bayes," *Log. J. Ilmu Komput. dan Pendidik.*, vol. 2, no. 3, pp. 594–606, 2024, [Online]. Available: https://journal.mediapublikasi.id/index.php/logic.
[21] N. Fidiyanto and A. N. Izzati, "Penerapan Data Mining Klasifikasi Lahan Tanam Buah Alpukat dengan Algoritma Naïve Bayes," *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 5, no. 2, pp. 95–103, 2024.
[22] M. Julkarnain and M. Yustiardin, "Penerapan Algoritma Naive Bayes Dalam Memprediksi Lulus Tepat Waktu Mahasiswa," *Digit. Transform. Technol.*, vol. 4, no. 2, pp. 848–858, 2024.
[23] F. Ramadhan and H. Bhakti Dwi, "Klasifikasi Penilaian Kinerja Karyawan Menggunakan Algoritme Naïve Bayes (Studi Kasus Pt. As Sabar Sukses Berkah)," *Kohesi J. Multidisiplin Saintek*, vol. 4, no. 2, 2024, [Online]. Available: https://ejournal.warunayama.org/index.php/kohesi/article/view/4707.
[24] Z. D. R. Sari, Jasmir, and Y. Arvita, "Penerapan Data Mining Untuk Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5," *J. Inform. Dan Rekayasa Komput.*, vol. 4, no. April, pp. 827–834, 2024.
[25] N. P. Panjaitan, S. Z. Harahap, and R. M. Ah, "Analisis Minat Masyarakat Menggunakan Media Sosial Menggunakan Algoritma C4.5 dan Metode Naïve Bayes," *Informatika*, vol. 12, no. 3, pp. 1–23, 2024.
[26] W. A. Cahyadi, S. Munafis, Y. E. Muda, and P. S. Informatika, "Penerapan Data Mining Dalam Penentuan Jurusan Siswa Dengan Metode Klasifikasi Algoritma C4 . 5 Studi Kasus SMAN 1 Leuwisadeng," in *Prosiding SENANTIAS: Seminar Nasional Hasil Penelitian dan PkM*, 2024, vol. 5, no. 1, pp. 80–86.
[27] N. Ndruru and A. Sinda, "Penerapan Data Mining Klasifikasi Kepuasan Pelanggan Transportasi Online Menggunakan Algoritma C4.5," *Katera J. Sains dan Teknol.*, vol. 1, no. 1, pp. 1–7, 2023, doi: 10.54367/means.v8i1.2569.
[28] R. Xsanal Hakim, F. Putrawansyah, and R. Syahri, "Penerapan Algoritma C4.5 Untuk Prediksi Anak Stunting Di Kota Pagar Alam," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 2469–2478, 2024, doi: 10.36040/jati.v8i2.9301.
[29] D. M. Musa *et al.*, "Penerapan Data Mining Untuk Klasifikasi Data Penjualan Pakan Ternak Terlaris Dengan Algoritma C4.5," *J. Teknlogi Inform. dan Komput. MH. Thamrin*, vol. 10, no. 1, pp. 168–182, 2024.