KLASIFIKASI JENIS BENIH KACANG MENGGUNAKAN SMOTE DAN DECISION TREE C4.5

J.M. Havidz Yasaf Al-Afghoni, Wahyudi Setiawan, Yudha Dwi Putra Negara
Sistem Informasi, Universitas Trunojoyo Madura
Jalan Raya Telang Kamal Bang kalan, Indonesia
Yasaf.hafidz@gmail.com

ABSTRAK

Di era yang berubah semakin modern ini semua hal dapat diidentifikasikan berdasarkan kriteria atau atribut yang di miliki, mulai dari benda, hewan, tanaman, dan semua hal yang memiliki materi. Penelitian ini bertujuan untuk mengklasifikasikan jenis kacang, ada banyak kacang yang belum memiliki label yang sesuai sehingga sering salah dalam menentukan jenisnya, penelitian ini bertujuan untuk mengelompokkan data kacang baru agar sesuai dengan jenis kacang yang dibutuhkan, metode decision tree C4.5. Metode ini digunakan untuk membuat model pengelompokan data yang dapat menggolongkan kacang ke dalam class yang berbeda berdasarkan atribut-atributnya. Data yang dipergunakan pada penelitian ini data yang berasal Murat Koklu, dengan data yang terbagi menjadi 7 class yaitu Seker 2027 data, Barbunya 1322 data, Bombay 522 data, Cali 1630 data, Dermason 3546 data, Horoz 1928 data dan Sira 2636 data dengan total data 13611dan memiliki atribut data berjumlah 16 jenis. Data yang diperoleh di normalisasi lalu dilakukan penyeimbangan data karena data tidak seimbang tiap Class dengan teknik SMOTE (Synthetic Minority Over-sampling Technique), berikutnya dilakukan pembagian data menggunakan K-Fold Cross Validation. Klasifikasi yang digunakan pada penelitian ini adalah decision tree C4.5. dan dari penelitian ini menghasilkan tingkat akurasi sebesar 94% presisi sebesar 94%, recallnya 94%, dan f1-scorenya adalah 94%, hasil ini dievaluasi menggunakan Confusion Matrix.

Kata kunci : *klasifikasi, Decision tree C4.5, Kacang, SMOTE, Confusion Matrix*

1. PENDAHULUAN

Di Indonesia pertanian sering kali menjadi sumber kebutuhan pokok bagi sebagian besar penduduk di daerah pinggiran atau desa. Petani dan pekerja di sektor pertanian menggantungkan hidup mereka dari hasil produksi pertanian. Sehingga pertanian memiliki peran penting dalam pengurangan kemiskinan, penciptaan lapangan kerja, dan kesejahteraan ekonomi bagi komunitas lokal. Dari Badan Pusat Statistik (BPS) nilai produksi hasil tani tahun 2020 meningkat 5,12 poin dibandingkan tahun 2019, yaitu dari 162,43 (angka konstan) pada tahun 2019 menjadi 167,55 (angka sementara) pada tahun 2020. Hal ini disebabkan oleh meningkatnya indeks produksi ternak hortikultura, perkebunan, dan peternakan[1]

Kacang merupakan sumber pangan pokok yang penting di dunia. Kacang dapat digunakan sebagai bahan utama untuk berbagai makanan, seperti bubur, pasta, roti, serta kudapan. Selain itu, kacang juga mengandung nutrisi yang banyak dan baik untuk kesehatan, salah satunya protein, serat, vitamin, dan mineral, jenis kacang-kacangan yang akan di bahas memiliki berbagai jenis, seperti Seker. Barbunya, Bombay, Cali, Horoz, Sira. dan Dermason. Kacang kering (Phaseolus vulgaris L.) merupakan tumbuhan berjenis pulse atau tanaman menghasilkan kulit (Fabaceae - Leguminosae), kacang juga menjadi tanaman terpenting dan memiliki produksi terbanyak di dunia

Klasifikasi merupakan pengelompokan data bisa juga berupa objek baru ke kelas atau label berdasarkan atribut yang dimiliki [2], dengan klasifikasi ini dapat pula digunakan untuk mengklasifikasikan tanaman kacang dengan berbagai metode yang ada di dalamnya, klasifikasi biasanya dilakukan dengan menggunakan algoritma pembelajaran mesin. Algoritma ini akan mempelajari pola atau hubungan antara fitur-fitur yang dimiliki oleh objek-objek dalam data latih atau training data, dan kemudian membuat model atau aturan-aturan untuk memprediksi label atau kelas dari objek-objek yang belum di identifikasi kelas nya dalam data uji atau testing data

Salah satu metode yang cukup umum digunakan dalam klasifikasi yaitu decision tree, Decision Tree adalah model yang merepresentasikan proses pengambilan keputusan dalam bentuk struktur pohon. Setiap titik percabangan pada pohon ini mewakili suatu atribut data yang digunakan untuk membagi data menjadi kelompok-kelompok yang lebih kecil. Proses klasifikasi dimulai dari akar pohon dan mengikuti cabang-cabang hingga mencapai daun, yang menunjukkan kelas atau kategori dari data tersebut.

Dan juga dapat mengatasi data yang hilang atau yang kosong sehingga perhitungan dapat selesai dengan akurat dan tepat, dari penelitian yang dikemukakan pada tahun 2022 yang ditulis oleh Aulia Rahmayanti, Lili Rusdiana, dan Suratno, penelitian ini menggunakan data nilai mahasiswa dengan data berjumlah 7 atribut. Penelitian ini mengulik tentang perbandingan antara metode decision tree dan naïve bayes, penelitian ini mengukur tingkat Dalam memprediksi kelulusan tepat waktu mahasiswa STMIK Palangkaraya, algoritma C4.5 menunjukkan kinerja yang lebih baik dibandingkan algoritma Naïve Bayes. Algoritma C4.5 mencapai akurasi sebesar 90%,

recall 92%, dan precision 94%, sedangkan Naïve Bayes hanya mencapai akurasi 85%, recall 86%, dan precision 93%. Hasil ini mengindikasikan bahwa C4.5 lebih handal dalam mengklasifikasikan mahasiswa berdasarkan kemungkinan kelulusan tepat waktu.[3]

2. TINJAUAN PUSTAKA

2.1. Kacang

Arachis hypogaea, atau kacang tanah, diperkenalkan ke wilayah Nusantara diperkirakan antara tahun 1521 hingga 1529. Meskipun demikian, budidaya kacang tanah secara intensif baru dimulai pada awal abad ke-18. Varietas yang pertama kali dibudidayakan adalah tipe kacang tanah menjalar [4], Karena kandungan gizi nya yang baik dan fleksibilitas dalam pengolahan, kacang menjadi salah satu bahan makanan yang banyak disukai.

Kacang tanah (Arachis hypogaea L.) merupakan komoditas strategis dalam sektor pertanian Indonesia, setelah kedelai. Kandungan lemak, protein, dan karbohidrat yang tinggi menjadikan kacang tanah sebagai sumber nutrisi yang bernilai. Fleksibilitas kacang tanah dalam pengolahan, baik sebagai bahan pangan langsung maupun bahan baku industri, telah mendorong peningkatan permintaan secara signifikan seiring dengan pertumbuhan penduduk. [4]

2.2. Data Mining

Data mining adalah proses penggalian informasi berharga dari kumpulan data yang besar. Proses ini melibatkan beberapa teknik untuk menemukan pola, hubungan, dan tren yang tersembunyi dalam data. Berdasarkan jenis tugasnya, data mining dapat dikategorikan menjadi lima jenis utama, yaitu Exploratory Data Analysis (EDA) atau Memahami karakteristik data secara mendalam., Descriptive juga disebut Mendeskripsikan data dengan menggunakan model statistik, Modelling atau Membangun model untuk memprediksi nilai atau kelas suatu variabel, Predictive Modelling, Penemuan Pola serta Aturan yang digunakan, dan juga Pemanggilan Konten[5].

Data mining adalah proses ekstrak data untuk mendapat informasi atau pola-pola yang tersembunyi dalam data besar yang dapat digunakan untuk pengambilan keputusan. Data mining adalah proses membangun sebuah model berdasarkan data yang ada. Model ini kemudian digunakan untuk menganalisis data baru dan menemukan pola-pola yang tersembunyi. Dengan kata lain, kita melatih sebuah sistem untuk "belajar" dari data yang kita miliki, sehingga sistem tersebut dapat membuat prediksi atau klasifikasi pada data yang belum pernah dilihat sebelumnya.[6]. Data mining digunakan untuk mengidentifikasi hubungan dan keterkaitan antar variabel yang mungkin tidak terlihat secara langsung, sehingga dapat membantu dalam menemukan pola dan informasi baru dalam data. Pemrosesan data mining umumnya melibatkan beberapa tahapan, mulai dari pemilihan data yang relevan, pemrosesan data, pengolahan data, dan analisis data. Proses ini dilakukan dengan menggunakan teknik-teknik data mining seperti klasifikasi, clustering, regresi, asosiasi, dan lain-lain.

2.3. SMOTE

SMOTE adalah metode yang digunakan untuk membuat model pembelajaran mesin lebih akurat. Dengan cara meningkatkan jumlah data pada kelas minoritas, SMOTE membantu model untuk belajar lebih baik dan menghindari bias terhadap kelas mayoritas. [7]. Ini merupakan salah satu teknik dari yang umum digunakan untuk pemrosesan data untuk menangani dalam ketidakseimbangan data dalam suatu kelas, khususnya dalam konteks masalah klasifikasi, Ketidakseimbangan kelas terjadi apabila jumlah sampel untuk setiap kelas dalam dataset tidak seimbang. Artinya, ada satu atau beberapa kelas yang memiliki jumlah sampel yang jauh lebih sedikit daripada kelas lain. Masalah ini dapat menyebabkan model machine learning menjadi cenderung memihak kelas mayoritas dan kurang mampu memprediksi kelas minoritas. SMOTE bekerja dengan cara menciptakan sampel sintetis (artificial samples) untuk kelas minoritas. Sampel sintetis ini dibuat dengan menggabungkan atribut dari sampel minoritas yang ada.

Secara umum untuk menyeimbangkan data menggunakan smote yang pertama adalah identifikasi Kelas Minoritas, Cari kelas dengan jumlah data lebih sedikit dibanding kelas lainnya. Lalu menemukan Tetangga Terdekat, Untuk setiap sampel data di kelas minoritas, temukan sejumlah tetangga terdekat menggunakan algoritma seperti k-Nearest Neighbors (k-NN). Berikutnya pembuatan Data Sintetis, data baru dibuat dengan mengambil selisih antara sampel data asli dan tetangganya, kemudian mengalikan dengan nilai random (0-1) dan menambahkannya ke sampel asli. Selanjutnya penambahan Data Sintetis, data hasil sintesis ditambahkan ke dataset, sehingga jumlah sampel pada kelas minoritas menjadi seimbang dengan kelas mayoritas.

2.4. Klasifikasi

Klasifikasi adalah metode pembelajaran mesin yang menggunakan data berlabel untuk membuat model. Model ini kemudian digunakan untuk memprediksi kelas atau kategori dari data baru. Dengan kata lain, kita melatih komputer untuk menggolongkan data ke dalam kelompok-kelompok tertentu berdasarkan ciri-ciri yang telah dipelajari, Klasifikasi adalah proses mengelompokkan data ke dalam kategori yang sudah ditentukan. Ada banyak algoritma yang bisa digunakan untuk melakukan klasifikasi, seperti pohon keputusan, aturan berbasis, jaringan saraf tiruan, dan Naive Bayes. Algoritma-algoritma ini belajar dari data untuk menemukan pola yang menghubungkan antara ciri-ciri data dengan kategori yang sesuai[9]. Tujuan klasifikasi adalah untuk memprediksi kelas atau kategori dari suatu data. Misalnya, kita ingin memprediksi apakah suatu email

adalah spam atau bukan, atau mengidentifikasi jenis penyakit berdasarkan gejala-gejala yang muncul. Dengan membangun model klasifikasi, kita bisa membuat keputusan yang lebih baik berdasarkan data yang ada[5].

2.5. Algoritma *C4.5*

Algoritma C4.5 digunakan untuk membuat sebuah model yang disebut pohon keputusan. Pohon keputusan ini seperti diagram alur yang membantu kita mengambil keputusan, Untuk menggunakan C4.5, kita perlu menyiapkan data terlebih dahulu. Kemudian, kita akan mencari atribut terbaik untuk menjadi akar pohon keputusan. Selanjutnya, kita akan menghitung nilai gain untuk setiap atribut dan memilih atribut dengan nilai gain tertinggi. Proses ini akan diulang untuk setiap cabang pada pohon sampai kita tidak bisa lagi membagi data[10]. J. Ross Quinlan adalah orang yang pertama kali mengembangkan algoritma pohon keputusan pada tahun 1970-an. Algoritma yang pertama kali dia buat adalah ID3. Kemudian, Quinlan melakukan perbaikan dan pengembangan pada algoritma ID3 sehingga menghasilkan algoritma C4.5. Algoritma C4.5 merupakan pengembangan lebih lanjut dari algoritma ID3.

Tahap proses algoritma C4.5:

a. Mempersiapkan data training, data ini umumnya diambil dari peneliti yang telah dikelompokkan menjadi kelas tertentu
b. Menghitung *root*, atau dalam bahasa Indonesia memiliki arti akar pohon, nama akar dirujuk dari atribut tertentu dengan cara menghitung nilai gain dari setiap atribut yang dimiliki, nilai gain yang tertinggi dijadikan akar paling awal, namun sebelum mencari nilai gain harus mencari nilai *entropy* dahulu
c. Mencari nilai Entropy

$$Entropy(S) = \sum_{i=1}^{n} -pi * Log2pi \quad (1)$$

Keterangan:
S :Kumpulan Data
n : Jumlah Bagian dari S
pi : proporsi Si terhadap S

d. Berikutnya untuk mencari nilai gain

$$Gain(S, A) = Entropy(S) - \sum_{i-1}^{n} \frac{|si|}{|s|} * Entropy(Si) \quad (2)$$

keterangan:
S : Fumpulan Data
A : Fitur
n : Jumlah Partisi Atribut A
|Si| : Jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S [11]

2.6. *Confusion matrix*

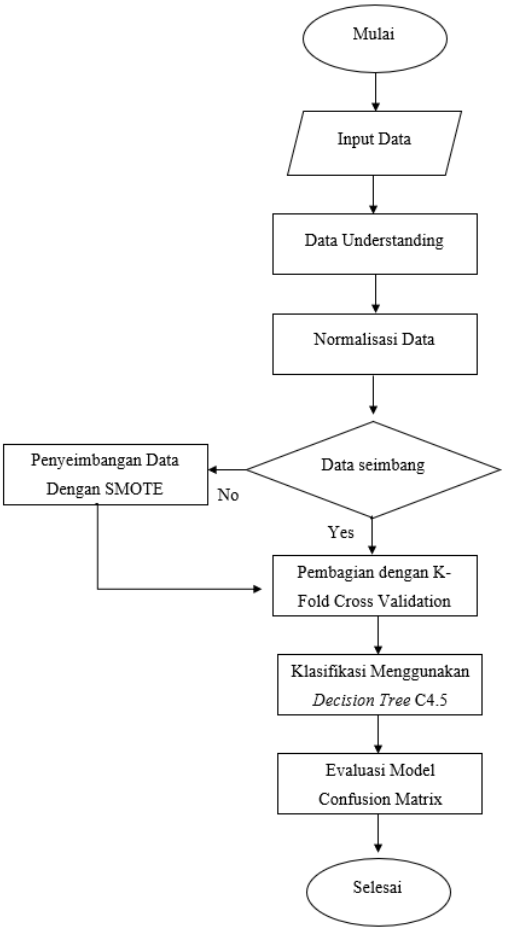
Confusion matrix juga dikenal sebagai tabel kontingensi, adalah sebuah tabel yang digunakan untuk mengevaluasi performa model klasifikasi, *Confusion matrix* menggambarkan hasil prediksi model klasifikasi dengan membandingkan prediksi yang dibuat oleh model dengan nilai sebenarnya dari data yang diamati. Confusion matrix memberikan informasi yang berguna untuk mengevaluasi performa model decision tree.

Prediksi Model yang telah dihasilkan dari proses untuk memprediksi kelas pada dataset akan dibandingkan hasil prediksi model dengan nilai aktual (label asli). Setiap prediksi dikategorikan ke dalam salah satu dari empat kategori True Positive (TP): Prediksi benar, sesuai dengan kelas positif, True Negative (TN): Prediksi benar, sesuai dengan kelas negatif, False Positive (FP): Prediksi salah, model memprediksi positif, tetapi sebenarnya negatif, False Negative (FN): Prediksi salah, model memprediksi negatif, tetapi sebenarnya positif. Berikutnya dilakukan Menyusun Matriks dengan Tampilkan keempat kategori tersebut dalam bentuk matriks untuk melihat distribusi hasil, dan dari tabel yang dihasilkan digunakan untuk menghitung metrik evaluasi dari matriks tersebut, hitung metrik evaluasi seperti Akurasi: Proporsi prediksi yang benar, Presisi: Proporsi prediksi positif yang benar, Recall (Sensitivitas): Proporsi data positif yang berhasil diprediksi dengan benar., dan F1-Score: Harmonic mean dari presisi dan recall.

2.7. Penelitian Terdahulu

romli, ikhsan dan ahmad turmudi, pada 2020 meneliti tentang Teknologi informasi dan ilmu informatika sangat di butuhkan untuk menunjang kinerja. Perusahaan besar maupun kecil juga membutuhkan informasi yang cepat dan akurat untuk memudahkan dalam mengambil suatu keputusan. Penelitian ini menggunakan metode *Decision tree C4.5*, *RapidMiner*, *Confusion Matrix*. Dan Hasil pengujian dengan algoritma *C4.5* memiliki nilai *accuracy*, precision, dan *recall* yang bagus yaitu dengan *accuracy* 91%, precision 86.05% dan *recall* 92.5%. pada tahun 2021 Yuda Irawan meneliti tentang Seleksi pendonor darah yang selama ini dilakukan secara manual untuk menentukan apakah calon pendonor bisa mendonorkan darah atau tidak. Metode ini menggunakan *Decision tree C4.5*, Black Box, White Box, dan hasilnya Darah, Hemoglobin dengan nilai gain tertinggi (0.861212618) merupakan variabel yang paling menentukan terhadap keberhasilan melakukan donor darah, pada tahun 2020 hana dan fida maisa melakukan penelitian pengidap penyakit diabetes semakin bertambah. Merujuk dari sumber data Federasi Diabetes Internasional, dengan menggunakan meode *Decision tree C4.5*, *RapidMiner*, *Confusion Matrix*, hasil Pengujian ini menghasilkan akurasi yang sangat besar yaitu 97,12 % Precision sebesar 93,02% %, dan *Recall* sebesar 100,00%

3. METODE PENELITIAN



Gambar 1. Flowchart alur penelitian

3.1. Input Data

Dari data yang diperoleh, data berupa dataset tabel yang tersimpan pada file excel hal ini cukup mempermudah karena data telah tertata sebagai tabel dengan kolom sebagai atribut kacang dan baris sebagai record atau jumlah data yang digunakan. Data berasal dari website Kaggle, data ini diterbitkan pada tahun 2020 dan yang menerbitkannya adalah Murat Koklu dan Ilker Ali Ozkan. Data yang digunakan disini adalah data jenis kacang diperoleh dari data Kaggle yang memiliki 16 atribut dan terbagi menjadi 7 *class*, dari seluruh hasil *class* tersebut akan di klasifikasi berdasarkan atribut yang dimiliki, data tersebut memiliki jumlah 13611 record,

Keterangan dari data:

1. Area (A): Luas permukaan kacang dalam satuan piksel². Contoh: A = 28395 piksel.
2. Perimeter (P): Keliling kacang dalam satuan piksel. Contoh: P = 610,291.
3. Major Axis Length (Lmaj): Panjang diameter terpanjang kacang dalam piksel. Contoh: Lmaj = 208,1781 piksel
4. Minor Axis Length (Lmin): Panjang diameter terpendek kacang dalam piksel. Contoh: Lmin = 173,8887 piksel
5. Aspect Ratio (AR): Rasio antara panjang dan lebar kacang dihitung dengan Lmaj dibagi Lmin. Contoh: AR = 1,197191
6. Eccentricity (Ecc): Keovalan kacang, diukur antara 0–1. Contoh: Ecc = 0,549812
7. Convex Area (Aconvex): Luas bentuk cembung yang mengelilingi kacang dalam piksel². Contoh: Aconvex =28715 piksel
8. Equivalent Diameter (Deq): Diameter lingkaran dengan luas setara kacang dalam piksel. Contoh: Deq = 190,1411 piksel
9. Extent (Ext): Rasio antara luas kacang dan bounding box dihitung dengan cara membaginya. Contoh: Ext = 0,763923
10. Solidity (S): Rasio antara luas kacang dan luas bentuk cembungnya atau Aconvex. Contoh: S = 0,988856
11. Roundness (R): Ukuran kebulatan kacang dengan perbandingan 4 area dibagi phi perimeter2. Contoh: R = 0,958027
12. Compactness (C): Perbandingan luas area dengan keliling atau perimeter yang dipangkatkan dua. Contoh: C = 0,913358
13. Shape Factor 1 (SF_1): Faktor bentuk berdasarkan perimeter dibagi dengan akar dari area. Contoh: $SF1$ = 0,007332
14. Shape Factor 2 (SF_2): Faktor bentuk berdasarkan perimeter² dan area. Contoh: $SF2$ = 0,003147
15. Shape Factor 3 (SF_3): Faktor bentuk tambahan terkait perimeter dibagi Lmaj. Contoh: $SF3$ = 0,834222
16. Shape Factor 4 (SF_4): Indikator bentuk yang mengukur keunikan kacang dengan perhitungan area dibagi Lmaj dikali Lmin. Contoh: $SF4$ = 0,998724
17. Kolom Class dari kacang yang akan di klasifikasi.

3.2. Data Understanding

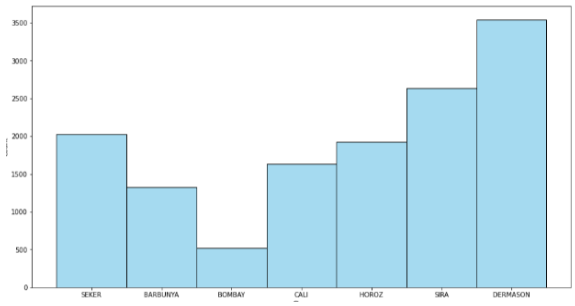
Data understanding (pemahaman data) merupakan tahap awal dalam proses analisis data. Tujuan dari tahap ini adalah untuk mengumpulkan, memahami, dan menganalisis data yang akan digunakan dalam analisis lebih lanjut. Poin ini juga dapat mendeskripsikan data, seperti Informasi umum tentang data, seperti ukuran, format, jenis data (numerik, kategori, teks), serta variabel apa yang terdapat dalam data, dalam hal ini dapat dicari Dimensi data, tipe dataset, jumlah atribut unik, hitung data tiap class, data yang tidak lengkap.

3.3. Normalisasi Data

Normalisasi adalah proses mengubah skala atau rentang fitur dalam dataset sehingga fitur-fitur tersebut memiliki nilai yang sebanding atau serupa. Tujuan dari normalisasi adalah untuk menghindari atribut dengan skala besar mendominasi proses pembelajaran mesin dan untuk meningkatkan kinerja algoritma.

3.4. Penyeimbangan Data Menggunakan SMOTE

Pada tahap ini data yang telah dibagi akan dilakukan penyeimbangan data, data yang di seimbangkan adalah data train penyeimbangan dilakukan dengan menggunakan metode teknik SMOTE, Tujuan dari langkah ini adalah untuk mengatasi ketidakseimbangan kelas dalam data pelatihan sehingga model dapat belajar dengan baik dari setiap kelas



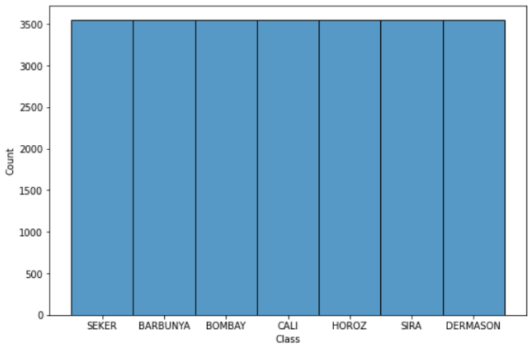
Gambar 2. Data sebelum di seimbangkan

Dapat dilihat dari gambar 2 terdapat data yang memiliki total yang sangat sedikit yaitu Bombay dengan total data 522 sebaliknya data pada Class Dermason memiliki data yang paling banyak yaitu 3546 hal ini memiliki selisih yang sangat banyak sehingga mempengaruhi hasil dari keakuratan klasifikasi. Langkah dalam Perhitungan Smote

a. Periksa distribusi data untuk mengidentifikasi kelas minoritas (kelas dengan jumlah sampel lebih sedikit dibandingkan kelas lainnya).
   Contoh: Kelas Bombay memiliki 522 data, sementara kelas Dermason memiliki 3546 data. Kelas Bombay adalah kelas minoritas.
b. Pilih Sampel Minoritas
   Pilih sampel dari kelas minoritas sebagai basis untuk menghasilkan data baru dalam kasus ini kelas minoritas adalah Bombay.
c. Generate Data Sintetis
   Pilih salah satu tetangga terdekat secara acak, Buat data sintetis dengan interpolasi, dengan persamaan

$$data\ sintesis = x_{min} + \lambda\,(x_{tetangga} - x_{min}) \quad (3)$$

   Keterangan
   x_{min} = Sampel minoritas asli.
   $x_{tetangga}$ = Sampel tetangga terdekat.
   λ = bilangan acak antara 0 dan 1
d. Memasukan ke data sets
   Data sintetis yang dihasilkan ditambahkan ke kelas minoritas sehingga jumlahnya mendekati jumlah kelas mayoritas.

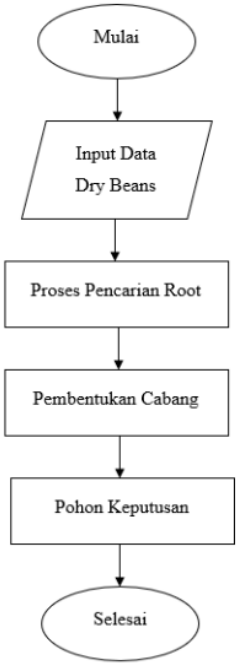


Gambar 3. Data setelah di seimbangkan

Data yang telah di seimbangkan menghasilkan data yang seimbang pada tiap kelas nya namun hasil tersebut merubah total seluruh data, dari hasil penyeimbangan ini menambahkan banyak data yang sekarang total data berjumlah 24822

3.5. Pembagian Data Training dan Data Testing

Pada tahap ini data yang diperoleh dibagi menjadi dua data yaitu data training dan data testing, data dibagi umumnya menghasilkan 80% data training dan 20% data testing, dari pembagian itu harus acak karena acak dan proporsional agar hasil evaluasi dapat dianggap valid. Namun selain itu data lain yang dapat dibagi dari proses selanjutnya yaitu data validasi, data ini digunakan pada proses selanjutnya untuk mengukur tingkat ketepatan pada proses klasifikasi ini.



Gambar 4. Flowchart decision tree C4.5

3.6. Model *C4.5 Classification*

3.6.1. Input Data

Pada proses awal poin ini memasukan data yang telah di proses atau data yang telah dibagi menjadi data training dan testing sehingga data telah menjadi data yang bagus dan tidak ada kecacatan data.

3.6.2. Proses pencarian Root

Pada proses pencarian root adalah langkah awal dalam membangun struktur decision tree. Root node merupakan titik awal dari decision tree, yang digunakan untuk membagi data berdasarkan atribut yang memiliki pengaruh paling signifikan dalam melakukan prediksi atau klasifikasi, mencari root dilakukan dengan 3 langkah.

a. Menghitung nilai entropy menggunakan rumus mencari entropy pada persamaan 2.1, entropy digunakan untuk mengukur keberagaman kelas target dalam himpunan data. Himpunan data dapat terdiri dari beberapa kelas yang berbeda, dan entropy digunakan untuk menentukan atribut mana yang paling baik untuk digunakan sebagai pemisah atau pembagi data pada setiap langkah dalam membangun decision tree.
b. Berikutnya menghitung nilai Gain, gain berfungsi untuk memilih atribut yang paling relevan atau memberikan kontribusi terbesar dalam memisahkan data dan membuat keputusan. Gain mengukur seberapa banyak informasi yang diperoleh atau ketidakpastian yang dikurangi ketika data dibagi berdasarkan atribut tertentu. Untuk menghitung gain dapat menggunakan persamaan 2.2 perhitungan nilai gain.
c. Apabila nilai gain telah ditetapkan selanjutnya menetapkan root, proses ini ditentukan dengan atribut yang mempunyai nilai gain tertinggi[12].

3.6.3. Pembentukan Cabang

Setelah menentukan akar dari pohon keputusan, kita akan melihat semua kemungkinan cabang yang bisa dibuat dari akar tersebut. Cabang yang paling memberikan informasi baru (nilai gain tertinggi) akan dipilih menjadi cabang utama. Proses ini akan diulang untuk setiap cabang baru sampai kita tidak bisa membuat cabang lagi. Apabila cabang terlalu banyak atau terlalu kompleks maka diperlukan pruning, pruning adalah pemangkasan cabang yang terlalu banyak, namun dalam pruning harus memperhatikan cabang yang akan di pangkas.

3.6.4. Keputusan

Pohon keputusan ditentukan dengan melalui serangkaian langkah yang melibatkan pemilihan atribut pembagi, pemisahan data, dan pembentukan cabang atau node anak, pohon keputusan terus berlanjut hingga semua atribut telah digunakan atau tidak ada atribut lain yang relevan untuk dipilih sebagai pembagi. Dan memutuskan sebuah data tersebut termasuk ke dalam kategori atau kelas yang memenuhi.

3.7. Evaluasi Confusion Matrix

Confusion matrix juga dikenal sebagai tabel kontingensi, adalah sebuah tabel yang digunakan untuk mengevaluasi performa model klasifikasi, Confusion matrix menggambarkan hasil prediksi model klasifikasi dengan membandingkan prediksi yang dibuat oleh model dengan nilai sebenarnya dari data yang diamati. Confusion matrix memberikan informasi yang berguna untuk mengevaluasi performa model decision tree. Dari confusion matrix, berbagai metrik evaluasi dapat dihitung, termasuk:

a. Akurasi: Rasio prediksi benar secara keseluruhan terhadap jumlah total contoh.
b. Presisi: Rasio prediksi benar positif terhadap total prediksi positif (TP / (TP + FP)).
c. Recall: Rasio prediksi benar positif terhadap total kelas positif sebenarnya (TP / (TP + FN)).

Confusion matrix memberikan informasi detail tentang keakuratan dan kesalahan model decision tree dalam melakukan prediksi kelas target. Hal ini membantu dalam evaluasi dan pemahaman performa model serta memungkinkan penyesuaian atau perbaikan yang diperlukan.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pengujian Model Training Data

Hasil dari proses pengujian dari data yang telah di olah atau di seimbangkan, dapat dilakukan metode decision tree C4.5 dengan menggunakan data yang telah diperoleh dapat dilakukan langkah-langkah utama yang dilakukan dalam pengujian ini mencakup perhitungan entropy, gain, dan hasil akhir dari pembentukan pohon keputusan.

4.1.1. Perhitungan Entropy

Entropy merupakan ukuran ketidakpastian dalam suatu dataset. Dalam konteks ini, entropy digunakan untuk mengevaluasi sejauh mana informasi yang dapat diperoleh dari atribut tertentu.

Tabel 1.. Contoh nilai entropy tiap Node

Node 0	Entropy = 2.8074
Node 1	Entropy = 1.9349
Node 2	Entropy = 1.6421
Node 3	Entropy = 0.7043
Node 4	Entropy = 0.1442
Node 5	Entropy = 0.0072
Node 6	Entropy = -0.0000
Node 7	Entropy = 0.1687
Node 8	Entropy = -0.0000
Node 9	Entropy = -0.0000

4.1.2. Perhitungan Gain

Setelah menghitung entropy pada parent node, langkah berikutnya adalah menghitung gain untuk setiap atribut. Information Gain mengukur seberapa besar suatu atribut dapat membantu dalam mengurangi ketidakpastian atau entropy dari data. Dalam proses decision tree C4.5, atribut yang memiliki nilai

information gain tertinggi akan dipilih sebagai split atau pembagi pada node tersebut.

Tabel 2.. Nilai Gain tiap atribut

Area	25.324
Perimeter	26.375
MajorAxisLength	26.466
MinorAxisLength	26.466
AspectRation	26.466
Eccentricity	26.466
ConvexArea	25.350
EquivDiameter	25.324
Extent	26.460
Solidity	26.448

4.1.3. Hasil Perhitungan dan Pembentukan Pohon Keputusan

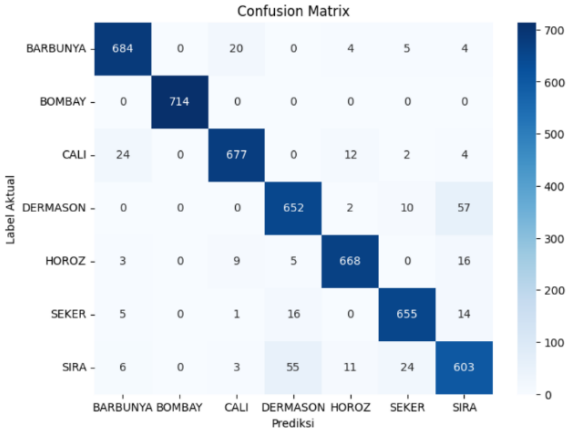
Berdasarkan perhitungan entropy dan gain, kami menemukan bahwa atribut dengan gain tertinggi adalah atribut yang akan digunakan untuk membagi dataset. Dalam hal ini, atribut yang dipilih menghasilkan pembagian yang optimal dengan entropy subset yang lebih rendah dibandingkan dengan entropy parent.

Tabel 3.. Contoh hasil Prediksi dan data aktual

Label	Aktual	Prediksi
2862	BARBUNYA	BARBUNYA
21074	HOROZ	SIRA
9400	SIRA	SIRA
7486	SIRA	DERMASON
18029	BOMBAY	BOMBAY
18029	BOMBAY	BOMBAY
12988	DERMASON	DERMASON
14000	BARBUNYA	BARBUNYA
1009	SEKER	SEKER
7488	SIRA	SIRA
4774	CALI	CALI

4.2. Pengujian Evaluasi Model

Pada eksperimen ini, model Decision Tree C4.5 diterapkan pada data pengujian untuk memprediksi kelas dari data. Berikut adalah hasil Confusion Matrix yang telah dilakukan:



Gambar 5. Hasil dari Perhitungan Confusing Matrix

Pada gambar 5 adalah hasil dari perhitungan confusing matrix yang telah dilakukan pada data yang telah di seimbangkan dengan menggunakan SMOTE dapat disimpulkan dari gambar menghasilkan nilai true positif, true negatif, false positif, dan false negatif. Untuk mencari nilai dari true positif adalah melihat titik dimana dua class yang sama bertemu, dan untuk mencari nilai true negatif adalah semua data selain dari data true positif, dan false positif adalah dengan menjumlahkan total dari nilai prediksi selain true positif, dan false negatif adalah menjumlahkan semua nilai dari label aktual selain true positif, dan hasilnya sebagai berikut:

- True Positive (TP): 668
- True Negative (TN): 4235
- False Positive (FP): 29
- False Negative (FN): 33

Dari perhitungan membandingkan data prediksi dan data aktual pada class kedua atau Horoz sehingga dapat dilakukan perhitungan untuk mencari akurasi, presisi, recall, dan f1-score

4.2.1. Akurasi

Untuk menghitung nilai akurasi adalah sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Jadi dari data dapat diperhitungkan

$$Accuracy = \frac{668 + 4235}{668 + 4235 + 29 + 33}$$

Maka hasil dari perhitungan tersebut adalah 0.98 atau tingkat akurasinya 98%.

4.2.2. Presisi

Selanjutnya untuk menghitung nilai presisi dapat menggunakan persamaan berikut

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Apabila data dimasukan maka hasilnya seperti ini:

$$Precision = \frac{668}{668 + 29}$$

Dari perhitungan nilai presisinya adalah 0.9583 atau 96%.

4.2.3. Recall

Agar dapat menghitung nilai recall dapat menggunakan rumus berikut

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

Dan hasil perhitungan seperti dibawah:

$$Recall = \frac{668}{668 + 33}$$

Dan hasilnya adalah 0.9534 atau dapat dijadikan menjadi persentase adalah 95%.

4.2.4. F1-Score

Dan yang terakhir yaitu menghitung nilai F1-score dengan menggunakan nilai dari presisi dan recall

$$F1 - Score = 2\ x \frac{precision\ x\ recall}{recall + precision}\ x\ 100\% \quad (6)$$

Apabila di hitung dengan menggunakan nilai yang diperoleh akan menjadi seperti ini:

$$F1 - Score = 2\ x \frac{0.9583\ \ x\ 0.9534}{0.9534 + 0.9583}\ x\ 100\%$$

Dan hasil dari perhitungan f1-score adalah 96%.

Yang paling akhir adalah nilai keseluruhan dari tiap class atau rata-rata dari akurasi, presisi, recall, dan f1-score dan hasilnya adalah Akurasi: 0.94 Presisi: 0.94 Recall: 0.94 F1-Score: 0.94, atau semua nilainya adalah 94%.

5. KESIMPULAN DAN SARAN

Berdasarkan pada seluruh penelitian yang telah dilakukan salah satu model untuk menyeimbangkan data adalah SMOTE metode ini melakukan penyeimbangan data dengan menggunakan data sintetis yang sesuai dengan data yang ada. berikutnya melakukan analisis dengan menggunakan metode Decision Tree C4.5 untuk membagi data secara optimal dengan mempertimbangkan nilai entropy dan gain. Perhitungan entropy pada masing-masing node menunjukkan pengurangan ketidakpastian secara bertahap pada setiap langkah pemisahan, yang diukur melalui nilai gain dari setiap atribut. Dan hasil dari evaluasi menggunakan confusion matrix adalah Akurasi: 94% Presisi: 94% Recall: 94% F1-Score: 94%. Diharapkan untuk penelitian berikutnya dapat dikembangkan dengan menggunakan metode yang lain dan juga dapat menambahkan fitur lain agar mendapat nilai yang lebih optimal, namun apabila tidak dilakukan penyeimbangan data menggunakan SMOTE maka hasilnya akan lebih kecil yaitu Akurasi: 0.90 Presisi: 0.90 Recall: 0.90 F1-Score: 0.90, atau semua nilainya adalah 90%.

DAFTAR PUSTAKA

[1] A. S. Wibowo, "Indikator Pertanian 2020," *05100.2107*, vol. 4, no. 1, p. xvi+134, 2020.

[2] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[3] A. Rahmayanti, L. Rusdiana, and S. Suratno, "Perbandingan Metode Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Kelulusan Mahasiswa," *Walisongo J. Inf. Technol.*, vol. 4, no. 1, pp. 11–22, 2022, doi: 10.21580/wjit.2022.4.1.9654.

[4] G. Sianipar, A. Indrawati, and A. Rahman, "Respon pertumbuhan dan produksi tanaman kacang tanah (arachis hypogaea l.) Terhadap pemberian kompos batang jagung dan pupuk organik cair limbah ampas tebu," *J. Ilm. Pertan. (JIPERTA)*, vol. 2, no. 1, pp. 11–22, 2020, doi: 10.31289/jiperta.v2i1.81.

[5] S. Supangat, A. R. Amna, and T. Rahmawati, "Implementasi Decision Tree C4.5 Untuk Menentukan Status Berat Badan dan Kebutuhan Energi Pada Anak Usia 7-12 Tahun," *Teknika*, vol. 7, no. 2, pp. 73–78, 2018, doi: 10.34148/teknika.v7i2.90.

[6] D. P. Utomo and B. Purba, "Penerapan Datamining pada Data Gempa Bumi Terhadap Potensi Tsunami di Indonesia," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 846, 2019, doi: 10.30645/senaris.v1i0.91.

[7] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *J. Edukasi dan Penelit. Inform.*, vol. 6, no. 3, p. 379, 2020, doi: 10.26418/jp.v6i3.42896.

[8] A. Peryanto, A. Yudhana, and R. Umar, "Klasifikasi Citra Menggunakan Convolutional Neural Network dan K Fold Cross Validation," *J. Appl. Informatics Comput.*, vol. 4, no. 1, pp. 45–51, 2020, doi: 10.30871/jaic.v4i1.2017.

[9] P. B. N. Setio, D. R. S. Saputro, and Bowo Winarno, "Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5," *Prism. Pros. Semin. Nas. Mat.*, vol. 3, pp. 64–71, 2020.

[10] S. Febriani and H. Sulistiani, "Analisis Data Hasil Diagnosa Untuk Klasifikasi Gangguan Kepribadian Menggunakan Algoritma C4.5," *89Jurnal Teknol. dan Sist. Inf.*, vol. 2, no. 4, pp. 89–95, 2021.

[11] J. Brier and lia dwi jayanti, "Algoritma C4.5 Untuk Kalsifikasi Calon Peserta Lomba Cerdas Cermat Siswa SMP Dengan Menggunakan Aplikasi Rapid Miner," vol. 21, no. 1, pp. 1–9, 2020, [Online]. Available: http://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203

[12] R. Haqmanullah Pambudi, B. Darma Setiawan, and Indriati, "Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 7, pp. 2637–2643, 2018, [Online]. Available: http://j-ptiik.ub.ac.id