SKRIPSI

PERBANDINGAN TINGKAT AKURASI METODE *WEIGHTED NAÏVE BAYES* DENGAN *RANDOM FOREST* DALAM MENGKLASIFIKASI PENERIMA PROGRAM KELUARGA HARAPAN (PKH)

NILAWATI
E0220305

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SULAWESI BARAT
TAHUN 2024

HALAMAN PENGESAHAN

Skripsi ini diajukan oleh:

Nama : Nilawati
NIM : E0220305
Judul : Perbandingan Tingkat Akurasi Metode *Weighted Naïve Bayes* dengan *Random Forest* dalam Mengklasifikasi Penerima Program Keluarga Harapan (PKH)

Telah berhasil dipertahankan di depan Tim Penguji (SK Nomor: 47/UN55.7/HK.04/2024, tanggal 10 Juli 2024) dan diterima sebagai bagian persyaratan memperoleh gelar Sarjana Statistika pada Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sulawesi Barat.

Disahkan oleh:

Dekan FMIPA
Universitas Sulawesi Barat

Musafira, S.Si., M.Sc.
NIP. 197709112006042002

Tim Penguji:

Ketua Penguji : Musafira, S.Si., M.Sc. (.....)
Sekretaris : Muh. Hijrah, S.Pd., M.Si. (.....)
Pembimbing 1 : Darma Ekawati, S.Pd., M.Sc. (.....)
Pembimbing 2 : Muhammad Hidayatullah, S.Pd., M.Kom. (.....)
Penguji 1 : Hikmah, S.Pd., M.Sc. (.....)
Penguji 2 : Retno Mayapada, S.Si., M.Si. (.....)
Penguji 3 : Putri Indi Rahayu, S.Si., M.Stat. (.....)

ABSTRAK

Proses penentuan penerima Program Keluarga Harapan (PKH) di Kelurahan Mosso, Kecamatan Sendana, Kabupaten Majene masih dilakukan secara manual, yang sering kali menyebabkan ketidaktepatan dalam penyaluran bantuan. Penelitian ini bertujuan untuk melihat tingkat akurasi metode *Weighted Naïve Bayes* dan *Random Forest* dalam mengklasifikasi penerima PKH serta membandingan hasil kedua metode tersebut. Data yang digunakan mencakup 73 responden dari kuisioner dan 100 data simulasi dan diproses melalui pra-pemrosesan, penanganan keseimbangan data menggunakan *Synthetic Minority Oversampling Technique* (SMOTE), dan klasifikasi dengan kombinasi data latih dan data uji sebesar 70%:30%, 80%:20%, dan 90%:10%, serta divalidasi menggunakan 5-*fold cross-validation.* Hasil penelitian ini menunjukkan bahwa tingkat akurasi metode *Weighted Naïve Bayes* pada data riil uji model sebesar 72,73% pada pembagian data 70%:30%, meningkat menjadi 87,50% pada pembagian 90%:10%, dengan akurasi *cross-validation* antara 81,44% hingga 85,32%. Sedangkan pada data simulasi, akurasi uji model mencapai 86,67% pada pembagian 70%:30% dan stabil pada sekitar 80% pada pembagian lainnya, dengan akurasi *cross-validation* antara 85,61% hingga 86,78%. Metode *Random Forest* menunjukkan performa yang lebih baik, dengan akurasi uji model pada data asli mencapai 82% hingga 100%, dan *cross-validation* antara 78,89% hingga 88,53%. Pada data simulasi, akurasi *Random Forest* berkisar antara 80% hingga 90% pada uji model dan 86,48% hingga 87,74% pada *cross-validation*. Berdasarkan hasil penelitian, *Random Forest* menunjukkan performa yang lebih stabil dan andal baik pada data riil maupun simulasi dalam mengklasifikasi penerima PKH dibandingkan *Weighted Naïve Bayes*.

Kata kunci: klasifikasi, Program Keluarga Harapan (PKH), *Random Forest, Weighted Naïve Bayes*

ABSTRACT

The process of determining recipients for the Program Keluarga Harapan (PKH) in Kelurahan Mosso, Kecamatan Sendana, Kabupaten Majene is still done manually, which often leads to inaccuracies in the distribution of aid. This study aims to assess the accuracy of the Weighted Naïve Bayes and Random Forest methods in classifying PKH recipients and to compare the results of these two methods. The data used includes responses from 73 questionnaires and 100 simulated data points, processed through preprocessing, data balancing using Synthetic Minority Oversampling Technique (SMOTE), and classification with training and test data splits of 70%:30%, 80%:20%, and 90%:10%, and validated using 5-fold cross-validation. The results show that the accuracy of the Weighted Naïve Bayes method on the real data was 72,73% for the 70%:30% split, increasing to 87,50% for the 90%:10% split, with cross-validation accuracy ranging from 81,44% to 85,32%. For simulated data, the model accuracy reached 86,67% for the 70%:30% split and remained around 80% for other splits, with cross-validation accuracy between 85,61% and 86,78%. The Random Forest method demonstrated better performance, with model accuracy on original data ranging from 82% to 100%, and cross-validation ranging from 78,89% to 88,53%. On simulated data, Random Forest accuracy ranged from 80% to 90% for the model and 86,48% to 87,74% for cross-validation. Based on the results, Random Forest shows more stable and reliable performance in classifying PKH recipients compared to Weighted Naïve Bayes, both for real and simulated data.

Keywords: classification, Program Keluarga Harapan (PKH), Random Forest, Weighted Naïve Bayes

BAB I
PENDAHULUAN

1.1. Latar Belakang

Kemiskinan di Indonesia adalah masalah kompleks yang melibatkan faktor alamiah dan kebijakan pembangunan (Annisya & Novira, 2023). Dalam mengatasi masalah kemiskinan, pemerintah mengadakan berbagai program sosial, salah satunya adalah Program Keluarga Harapan (PKH). PKH adalah salah satu program jaminan sosial yang diperuntukkan oleh Rumah Tangga Sangat Miskin (RTSM) melalui beberapa ketentuan yang berlaku (Nuraeni dkk., 2024). Program ini dikenal dengan *Conditional Cash Transfers* (CCT) yang dikembangkan oleh Bank Dunia untuk negara-negara berkembang. Program tersebut kemudian menjangkau negara Indonesia yang mulai dilaksanakan pada tahun 2007. Penyelenggaraan PKH diatur melalui Peraturan Menteri Sosial No. 1/2018 tentang Program Keluarga Harapan (Annisya & Novira, 2023). Tujuan dari program ini adalah untuk mengurangi kemiskinan dan meningkatkan sumber daya manusia melalui pembayaran tunai secara berkala (Amanda dkk., 2022).

Proses penerimaan bantuan PKH di Kelurahan Mosso, Kecamatan Sendana, Kabupaten Majene masih dilakukan secara manual. Namun kenyataannya, terdapat beberapa keluhan yang menyatakan bahwa penyaluran bantuan PKH di Kelurahan Mosso tidak tepat sasaran sehingga mengurangi efektivitas bantuan. Oleh karena itu, diperlukan solusi untuk meningkatkan akurasi dalam menentukan penerima PKH.

Salah satu solusi untuk menyelesaikan ketidaksesuaian penerima bantuan PKH adalah dengan melakukan klasifikasi. Klasifikasi merupakan proses pengelompokkan item ke dalam kategori yang berbeda (Aldiyansyah dkk., 2024). Klasifikasi kelayakan di Kelurahan Mosso dapat dilakukan dengan menggunakan teknik *data mining.*

Data mining adalah metode yang berguna untuk mengekstraksi informasi berharga dari sekumpulan data, yang dilakukan dengan menggunakan

pengetahuan seperti statistik, matematika, dan pengenalan pola (Damuri dkk., 2021).

Weighted Naïve Bayes adalah pengembangan dari metode *Naïve Bayes* yang dasar akurasi klasifikasi tidak hanya pada probabilitas setiap fitur, tetapi juga bobot pada setiap fitur (Panharesi & Anugrah, 2022). Pada dasarnya, *Naïve Bayes* adalah metode klasifikasi probabilistik dasar yang menghitung nilai probabilitas dengan menambahkan frekuensi dan jumlah nilai dari kumpulan data tertentu. Untuk membuat nilai fitur, metode ini menggunakan *teorema Bayes* yang mana mengasumsikan fitur-fitur tersebut *independent* (Fauzan & Hikmah, 2022). *Random Forest* merupakan metode klasifikasi yang terdiri dari kombinasi pohon klasifikasi *independent* (CART). Proses *voting* pohon klasifikasi yang dibuat diperoleh prediksi klasifikasi. *Random Forest* merupakan evolusi dari metode *ensemble* yang digunakan untuk meningkatkan akurasi klasifikasi (Fachruddin, 2015).

Penelitian sebelumnya menunjukkan efektivitas tinggi dari kedua metode ini. Beberapa diantaranya adalah penelitian yang dilakukan oleh Nurjannah dkk. (2023) yang meneliti tentang kelayakan penerima bantuan Program Keluarga Harapan (PKH) di Desa Cinta Rakyat menggunakan Metode *Weighted Naïve Bayes* dengan *Laplace Smoothing* dengan akurasi yang dihasilkan 95,65%. Penelitian yang dilakukan Muti & Pamuji (2023) yang meneliti tentang perbandingan metode *Naïve Bayes* dan *KNN* dalam klasifikasi kelayakan keluarga terdaftar DTKS penerimaan bantuan Sosial di Desa Dubesi dengan menghasilkan nilai akurasi pada *Naïve Bayes* 82% dan KNN 71%. Dan penelitian yang lain yang dilakukan oleh Haidar dkk. (2023) tentang penerapan *Deep Learning* model *Random Forest* untuk prediksi penerima bantuan Program Keluarga Harapan (PKH) dengan menghasilkan akurasi 96%, 98% dan 99%.

Berdasarkan latar belakang di atas, pada penelitian ini bertujuan untuk mengetahui tingkat akurasi metode *Weighted Naïve Bayes* dan *Random Forest* dalam mengklasifikasi penerima PKH, serta membandingkan hasil klasifikasi

penerima PKH pada metode *Weighted Naïve Bayes* dan *Random Forest*. Penelitian ini diharapkan dapat memberikan solusi yang lebih akurat dan efektif dalam menentukan penerima bantuan PKH, sehungga dapat membantu perangkat kelurahan dalam menyalurkan bantuan tepat sasaran.

1.2. Rumusan Masalah

Penentuan penerima PKH di Kelurahan Mosso yang masih dilakukan secara manual dan kurang efisien. Perlu dirumuskan metode yang membantu perangkat kelurahan untuk menentukan penerima bantuan PKH secara cepat dan efisien. Dua metode statistika yang sedang berkembang saat ini yakni *Weighted Naïve Bayes* dan *Random Forest* bisa digunakan untuk menjawab masalah tersebut. Sehingga dalam penelitian ini dirumuskan masalah yang akan dibahas, yakni:

1. Bagaimana tingkat akurasi metode *Weighted Naïve Bayes* dalam mengklasifikasi penerima PKH?
2. Bagaimana tingkat akurasi metode *Random Forest* dalam mengklasifikasi penerima PKH?
3. Bagaimana perbandingan hasil klasifikasi penerima PKH dengan metode *Weighted Naïve Bayes* dan *Random Forest*?

1.3. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk:

1. Mengetahui tingkat akurasi metode *Weighted Naïve Bayes* dalam mengklasifikasi penerima PKH.
2. Mengetahui tingkat akurasi metode *Random Forest* dalam mengklasifikasi penerima PKH.
3. Mengetahui perbandingan hasil klasifikasi penerima PKH dengan metode *Weighted Naïve Bayes* dan *Random Forest.*

1.4. Manfaat Penelitian

Dari penelitian ini diharapkan memberi memberikan manfaat sebagai berikut:

1. Memberikan informasi mengenai metode *Weighted Naïve Bayes* dan *Random Forest* yang lebih efektif dalam mengklasifikasi penerima PKH.
2. Membantu instansi terkait dalam memilih metode yang tepat dalam mengklasifikasi penerima PKH agar penerima tepat sasaran.

1.5. Batasan Masalah

Pada penelitian ini, terdapat beberapa batasan masalah agar penelitian ini tidak keluar dari pokok permasalahan sehingga penelitian ini terarah. Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Data yang digunakan dalam penelitian ini terbatas pada data penerima PKH yang ada di Kelurahan Mosso yang diambil melalui pengisian kuisioner.
2. Penelitian ini hanya membandingkan dua metode klasifikasi, yaitu *Weighted Naïve Bayes* dan *Random Forest.*
3. Penelitian ini hanya fokus pada perbandingkan hasil metode klasifikasi *Weighted Naïve Baye* dan *Random Forest* dengan menggunakan kombinasi perbandingan data *training* dan data *testing* adalah sebesar 70%:30%, 80%:20% dan 90%:10%.

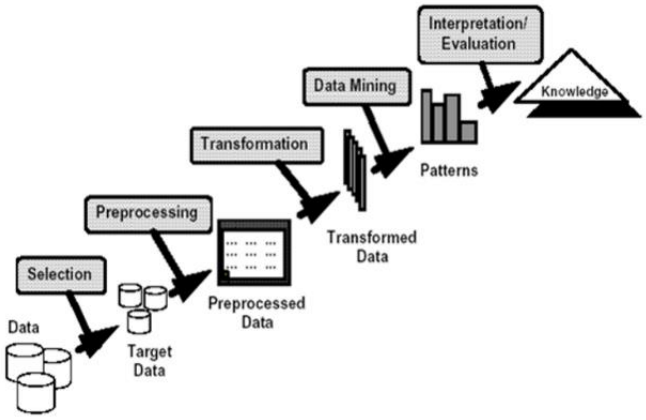
BAB II
KAJIAN PUSTAKA

2.1. Landasan Teori

Landasan teori ini menjelaskan konsep dan metode yang mendukung analisis data dan klasifikasi penerima PKH, seperti *data mining*, teknik klasifikasi dan pendekatan evaluasi model yang relevan.

2.1.1. Data Mining

Data mining dapat diartikan sebagai serangkaian proses yang mendapatkan pengetahuan atau pola dari kumpulan data. Tujuan *data mining* adalah untuk melakukan klasifikasi, klastering, menemukan pola asosiasi dan melakukan prediksi (Almuqorobin, 2021). Istilah *data mining* lebih dikenal dengan sebutan *Knowledge Discovery from Database* (KDD). KDD adalah proses analisis terstruktur untuk memperoleh informasi baru yang akurat dan menemukan pola pada data yang besar dan kompleks. Proses *text mining* menggunakan *Knowledge Discovery from Database* (KDD) sama dengan *data mining* (Aldiyansyah dkk., 2024).



Gambar 2.1. KDD *process* (sumber: (Novita dkk., 2022)**)**

Adapun tahapan KDD sebagai berikut:

a. *Data Selection*

   Tahap ini diperlukan sebelum tahap ekstraksi atau pengambilan informasi. Penyimpanan hasil pada tahap ini dilakukan guna melakukan proses *data mining* pada file selain database produksi.

b. Pre-processing

KDD fokus pada tahap *cleaning* yaitu tahap pembersihan data sehingga sebanyak data dapat diolah melalui *data mining.* Selain itu, pada tahap ini dilakukan *transformation*, pada tahap ini dilakukan tahap pengkodean agar data sesuai untuk proses *data mining*. Pada tahap ini, pengkodean dilakukan dalam proses kreatif dan harus sesuai dengan pola atau jenis informasi yang diperlukan dalam database.

c. *Data Mining*

Data mining adalah proses penggunaan metode atau teknik tertentu untuk mencari pola dalam data dan mengekstrak informasi yang berguna dan menarik.

d. *Interpretation/Evaluation*

Evaluasi merupakan tahap untuk menverifikasi apakah informasi yang diterima relevan dengan perkiraan awal sebelumnya.

2.1.2. Klasifikasi

Klasifikasi adalah teknik *data mining* yang merupakan suatu bentuk analisis data yang mengekstraksi model yang menggambarkan kelas-kelas data yang penting (Ekasatria, 2019). Proses klasifikasi bekerja dengan menemukan pola yang menjelaskan atau mendefinisikan kelas dalam data yang bertujuan untuk memprediksi kelas yang labelnya tidak diketahui (Utami & Devi, 2022). Langkah klasifikasi yaitu data dimasukkan, yang dikenal dengan data *training* terdiri dari banyak sampel masing-masing dengan beberapa atribut. Setiap sampel kemudian diberi label kelas khusus. Tujuannya untuk menganalisis data masukan dan menggunakan fitur data untuk mengembangkan deskripsi atau model yang akurat untuk setiap kelas. Deskripsi kelas ini digunakan untuk mengklasifikasikan data uji lainnya yang label kelasnya tidak diketahui. Deskripsi tersebut juga dapat digunakan untuk memahami setiap data (Almuqorobin, 2021).

2.1.3. Synthetic Minority Oversampling Technique (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) adalah salah satu dari metode untuk mengatasi masalah ketidakseimbangan data. Ketidakseimbangan data terjadi Ketika suatu kelas memiliki lebih banyak anggota dibandingkan kelas lainnya, keakuratan klasifikasi cenderung jauh lebih tinggi untuk kelas mayoritas dibandingkan kelas minoritas. SMOTE menangani data yang tidak seimbang dengan menghasilkan data buatan atau sintetik untuk kelas minoritas sehingga proporsi datanya lebih seimbang. Data buatan dihasilkan menggunakan konsep tetangga terdekat (Hidayatulloh, 2022).

2.1.4. Teorema Bayes

Teorema Bayes adalah metode yang efektif dalam pembelajaran mesin yang berbasis data pelatihan, memanfaatkan probabilitas bersyarat sebagai dasarnya. *Teorema Bayes* juga digunakan untuk menghasilkan estimasi parameter dengan menggabungkan informasi dari sampel dengan informasi yang sudah ada sebelumnya. Keunggulan dari penggunaan *Teorema Bayes* adalah penyederhanaannya dibandingkan dengan metode klasik yang kompleks dan penuh perhitungan integral untuk mendapatkan model marginal (Ratnasari & Hasibuan, 2017).

Dalam teori probabilitas klasik, untuk menghitung probabilitas kejadian A dengan syarat B, menggunakan rumus:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ untuk } P(B) \neq 0 \qquad (2.1)$$

Dimana $P(A|B)$ merupakan probabilitas terjadinya A ketika B telah terjadi, $P(A \cap B)$ adalah probabilitas terjadinya kedua kejadian A dan B secara bersamaan $P(B)$ adalah probabilitas terjadinya B.

Jadi, probabilitas bersyarat didefiniskan dalam konteks kejadian bersama, dimana perbandingan antara probabilitas kejadian bersama A dan B dengan

probabilitas B menghasilkan probabilitas bersayarat (Ratnasari & Hasibuan, 2017). Hal ini dapat dilihat dari persamaan di atas diperoleh:

$$P(A \cap B) = P(B|A)P(A) \tag{2.2}$$

Keterangan:

$P(A)$ = *Prior probability* dari A

$P(A|B)$ = *Posterior probability*

2.1.5. Naïve Bayes

Naïve Bayes adalah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data tertentu. Algoritma menggunakan *teorema Bayes* dan mengasumsikan bahwa semua fitur independen berdasarkan nilai label kelasnya (Muti & Pamuji, 2023). Persamaan *Naive Bayes* sebagai berikut:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \tag{2.3}$$

Keterangan:

$P(H|X)$ = Probabilitas hipotesis H berdasarkan kondisi X

H = Hipotesis data merupakan suatu kelas spesifik

X = Data dengan kelas yang belum diketahui

$P(X|H)$ = Probabilitas X berdasarkan kondisi pada hipotesis H

$P(H)$ = Probabilitas hipotesis H

$P(X)$ = Probabilitas X

2.1.6. Weighted Naïve Bayes

Dasar akurasi klasifikasi tidak hanya probabilitas tetapi juga bobot setiap fitur dalam suatu kelas yang dapat ditentukan dengan memperhatikan bobot fitur dalam kelas tersebut. Menambahkan bobot w_i ke setiap fitur akan menghitung bobot *Naïve Bayes* (Nurjannah dkk., 2023). Sehingga didapatkan persamaan berikut:

$$P(y|x) = P(y)\prod_{i=1}^{\infty} P(x_i|y)^{w_i} \quad (2.4)$$

Keterangan:

$P(y|x)$ = Probabilitas pada kelas y dalam data x

$P(y)$ = Probabilitas pada label kelas y

$P(x_i|y)$ = Probabilitas pada fitur x_i dengan label kelas y

w_i = Bobot atribut

Klasifikasi dapat ditentukan dengan melakukan perhitungan menggunakan metode *Weighted Naïve Bayes* dengan *Laplace Smoothing* (Utami & Devi, 2022). *Smoothing* berupaya mengurangi kemungkinan hasil yang diamati sekaligus meningkatkan kemungkinan hasil yang tidak teramati (Nurjannah dkk., 2023).

Berikut langkah perhitungan yang dilakukan dengan menggunakan metode *Weighted Naïve Bayes* dengan *Laplace Smoothing* (Utami & Devi, 2022)*:*

1. Perhitungan nilai probabilitas masing-masing kelas, dirumuskan sebagai berikut:

$$P(y) = \frac{\sum y}{n} \quad (2.5)$$

Keterangan:

$P(y)$ = Probabilitas label pada kelas y

$\sum y$ = Jumlah data dengan label pada kelas y

n = Jumlah total pada data latih

2. Penentuan nilai probabilitas untuk setiap karakteristik. Untuk menghindari nilai probabilitas menjadi 0, gunakan rumus *Laplace Smoothing* untuk menghitung jumlah data dan probabilitas nilai K = 1. Persamaannya sebagai berikut

$$P(x_i|y) = \frac{\sum x_i|y + K}{\sum y + K|x|} \quad (2.6)$$

Keterangan:

$P(x_i|y)$ = Probabilitas fitur x_i dengan label pada kelas y

$\sum x_i|y$ = Jumlah data fitur x_i dengan label pada kelas y

$\sum y$ = Jumlah data dengan label pada kelas y

K = Parameter *smoothing*, $K = 1$

$|x|$ = Jumlah kelas pada sampel

3. Pada persamaan (2.4) digunakan untuk menghitung distribusi klasifikasi data *training* menggunakan metode *Weighted Naïve Bayes* dengan *Laplace Smoothing.*

Bobot prioritas karakteristik data dapat diperoleh dengan menggunakan teknik *Rank Order Centroid (ROC)* (Nurjannah dkk., 2023). Sehingga persamaannya sebagai berikut:

$$W_m = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{i}\right) \tag{2.7}$$

Dengan keterangan W_m adalah bobot kriteria, m adalah jumlah kriteria dan i adalah kriteria.

2.1.7. Classification and Regression Trees (CART)

Classification and Regression Trees (CART) adalah salah satu metode atau algoritma dari teknik pohon keputusan. Metode yang dikembangkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen dan Charles J. Stone ini adalah teknik klasifikasi yang menggunakan algoritma penyekatan biner secara rekursif (*binary recursive partitioning*) (Sumartini & Purnami, 2015).

CART membuat pohon klasifikasi jika variabel respon memiliki skala kategorikal dan pohon regresi jika variabel respon berupa data kontinu. Tujuan utama CART adalah mendapatkan kumpulan data yang akurat sebagai fitur untuk klasifikasi (Siahaan dkk., 2016). Ada 3 proses dalam pembentukan algoritma CART yaitu pembentukan pohon klasifikasi dan penentuan pohon klasifikasi yang

optimal (Hana dkk., 2023). Tahapan pembentukan algoritma CART adalah sebagai berikut:

1. Pembentukan Pohon Klasifikasi

   Tahap awal dengan menentukan variabel dan *threshold* untuk dijadikan pemilah tiap simpul. Tahap pembentukan pohon klasifikasi terdiri dari:

   a. Pemilihan Pemilah

   Data yang digunakan adalah data latih sampel. Subset yang dihasilkan dari proses penyortiran harus lebih seragam dibandingkan penyortiran sebelumnya. Fungsi heterogenitas yang digunakan adalah indeks Gini karena akan selalu memisahkan kelas berdasarkan anggota terbesar/kelas terpenting dalam node (Sumartini & Purnami, 2015). Fungsi indeks Gini ditunjukkan pada persamaan berikut:

   $$i(t) = \sum_{i,j=1} p(j|t)\, p(i|t), i \neq j \qquad (2.8)$$

   Dengan $p(j|t)$ adalah proporsi kelas j pada simpul t dan $p(i|t)$ adalah proporsi kelas i pada simpul t.

   Pemilihan yang dipilih membentuk himpunan kelas yang disebut node. Node ini melakukan pengurutan rekursif hingga node akhir diperoleh. Langkah selanjutnya adalah menentukan kebaikan kriteria *goodness of split* untuk mengevaluasi pemilah dari s pada sampul t dengan rumus (Hana dkk., 2023):

   $$\varphi(s,t) = \triangle\, i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \qquad (2.9)$$

   Di mana $\varphi(s,t)$ adalah fungsi kebaikan atau *gooodness of split* untuk mengevaluasi pemilah dari s pada sampul t, $i(t)$adalah indeks Gini pada waktu t, p_L dan p_R adalah koefisien untuk node kiri t_L dan kanan t_R, $i(t_L)$ dan $i(t_R)$ adalah indeks Gini pada node kiri dan kanan.

   Pemilah yang menghasilkan $\varphi(s,t)$ lebih tinggi merupakan pemilah terbaik karena mampu mereduksi heterogenitas lebih tinggi.

   b. Penentuan Simpul Terminal

Jika sebuah node memiliki observasi dan jumlah totalnya kurang atau sama dengan 5 $(n \leq 5)$ pohon berhenti berkembang. Selain itu, proses pembentukan pohon terhenti meskipun batas jumlah level yang ditemukan atau level kedalaman pohon klasifikasi maksimum tercapai (Sumartini & Purnami, 2015).

c. Penandaan Label Kelas

Penentuan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak, yaitu jika

$$p\left(j_0|t\right) = \max_j \frac{N_j(t)}{N(t)} \tag{2.10}$$

Label kelas untuk simpul terminal t adalah j_0 yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul t yang paling kecil sebesar $r(t) = 1 - \max_j p\left(j|t\right)$.

Dimana $r(t)$ adalah nilai kemurnian simpul t, $max_j\, p\,(j|t)$ adalah proporsi terbesar dari kelas j di simpul terminal t.

2. Pemangkasan Pohon Klasifikasi

Pemberhentian pohon didasarkan pada jumlah observasi atau derajat keseragaman pada simpul akhir, pohon yang dibentuk menggunakan aturan pengurutan dan kriteria *goodness of split* bisa berukuran sangat besar. Ukuran pohon yang besar dapat menyebabkan *overfitting* (Sumartini & Purnami, 2015). Namun, jika pengamatan pohon dibatasi pada tingkat ketelitian tertentu, maka dapat terjadi *underfitting*. Pohon dengan ukuran optimal dapat dicapai dengan memangkas pohon berdasarkan ukuran kompleksitas biaya minimum.

$$R_\propto(T) = R(T) + \propto \left|\tilde{T}\right| \tag{2.11}$$

$R_\propto(t)$ merupakan kombinasi linear biaya dan kompleksitas pohon yang dibentuk dengan menambahkan *cost penalty* bagi kompleksitas terhadap biaya kesalahan klasifikasi pohon. Setelah itu, dilakukan pencarian pohon bagian $T(\propto) < T\max$ yang meminimumkan $R_\propto(t)$ yaitu:

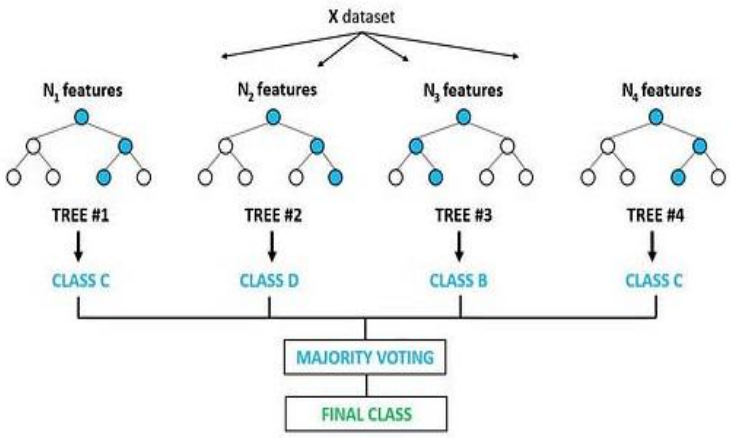
$$R_{\propto}\left(T\left(\propto\right)\right)=\min_{T<T_{\max}} R_{\propto}\left(T\right) \quad (2.12)$$

3. Penentuan Pohon Klasifikasi Optimum

   Pohon klasifikasi yang terbentuk bisa sangat kompleks. Oleh karena itu, optimasi awal dianggap perlu sebelum digunakan untuk mengklasifikasi data baru. Pengoptimalan pohon memilih ukuran pohon yang benar dan memangkas node yang tidak penting (Hana dkk., 2023).

2.1.8. Random Forest

Random Forest merupakan teknik CART (*Classification and Regression Trees*) dalam *data mining* yang tidak memerlukan asumsi. Metode ini menggunakan konsep pohon keputusan karena model ini terbentuk dari banyak pohon seperti membentuk kumpulan pohon mirip hutan dengan menerapkan teknik *bootstrap aggregating (bagging)* dan pemilihan fitur acak (Almuqorobin, 2021).



Gambar 2.2. Klasifikasi *Random Forest* (sumber: (Siregar dkk., 2023)**)**

Saat membentuk pohon klasifikasi, *Random Forest* dilakukan pada *cluster* data yang terdiri dari *n* observasi dan *p* variabel penjelas (Darwanto dkk., 2021). Tahap *Random Forest* adalah sebagai berikut (Purwa, 2019):

1. Menentukan jumlah pohon (k) yang akan dibentuk.

2. Mengambil sampel secara acak dengan pergantuan n observasi sebanyak mungkin dalam kumpulan data berukuran n untuk setiap pohon.
3. Untuk setiap pohon subset prediktor sebanyak m juga dipilih secara acak. Dimana $m < p$, dengan p merupakan jumlah variabel prediktor.
4. Ulangi proses kedua dan ketiga hingga memiliki pohon (k) sebanyak mungkin.
5. Untuk kasus prediksi, hasil prediksi *Random Forest* adalah nilai prediksi rata-rata sebanyak k pohon. Sedangkan pada kasus klasifikasi diperoleh hasil prediksi *Random Forest* berdasarkan hasil klasifikasi jumlah pohon yang sama.

Penerapan *Random Forest* dalam model klasifikasi telah terbukti meningkatkan kemampuan model dalam mendeteksi masalah. Analisis komparatif ini memberikan pemahaman yang jelas tentang kelebihan dan keterbatasan algoritma *Random Forest* dibandingkan dengan *Weighted Naïve Bayes* (Aldiyansyah dkk., 2024). Hasil penelitian diharapkan dapat memberikan panduan praktis kepada peneliti dalam memilih algoritma yang sesuai untuk setiap kasus.

2.1.9. Randomized Search Cross-Validation

Pada metode *machine learning*, ada beberapa nilai parameter yang disebut *hyperparameter* yang diperkirakan dapat meningkatkan performa model. Teknik *hyperparameter* yang digunakan adalah *Randomized Search CV*. Metode ini fungsinya sama dengan *Grid Search CV*. Namun, perbedaannya adalah *Randomized Search CV* dapat menemukan parameter optimal lebih cepat (Hidayati dkk., 2023).

2.1.10. Perbandingan Metode Klasifikasi

Dalam membandingkan metode klasifikasi, evaluasi dilakukan untuk menilai seberapa akurat model yang dibuat dapat menghasilkan prediksi atau klasifikasi berdasarkan data yang belum pernah dilihat sebelumnya (Sidabutar dkk., 2023). Proses pengujian ini menggunakan *confusion matrix* yang akan

menghitung nilai *precision, recall, dan accuracy*. *Confusion matrix* terdiri dari *true positive, false positive, true negative* dan *false negative* unuk menghitung presisi, *recall,* dan akurasi. Presisi merupakan derajat ketepatan antara informasi yang diminta oleh pengguna dengan respon yang diberikan sistem. *Recall* adalah tingkat keberhasilan sistem mengambil informasi. Akurasi adalah tingkat kesesuaian antara nilai prediksi dan nilai aktual. Untuk menghitung presisi, *recall,* dan akurasi dapat menggunakan persamaan berikut (Damuri dkk., 2021):

$$Precision = \frac{TP}{TP + FP} \tag{2.13}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.14}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.15}$$

$$F1 - Score = 2 \times \frac{\left(Recall \times Precision\right)}{\left(Recall + Precision\right)} \tag{2.16}$$

Tabel 2.1. *Confusion matrix*

Prediksi	Aktual	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Keterangan:

TP = *True positive* yang didapatkan dari jumlah data positif yang diprediksi benar

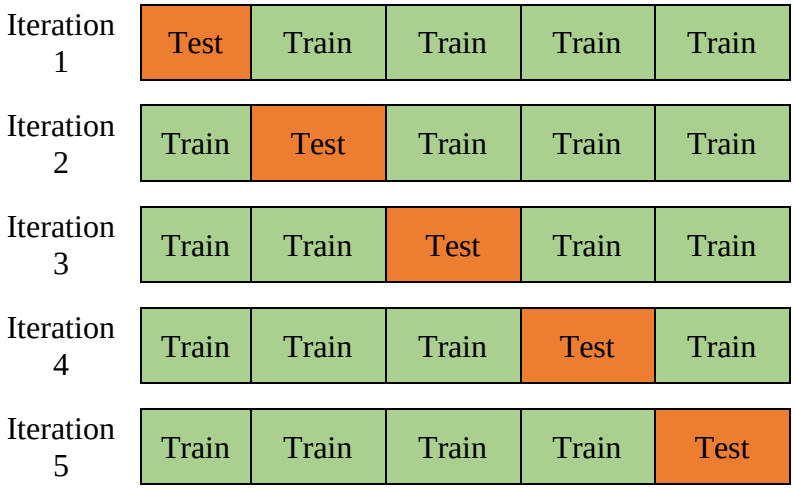
TN = *True negative* didapatkan dari jumlah data negatif yang diprediksi benar

FP = *False positive* didapatkan dari jumlah data negatif namun diprediksi sebagai data positif

FN = *False negative* yang didapatkan dari jumlah data positif namun diprediksdi sebagai data negatif

2.1.11. Cross-Validation

Cross-validation atau biasa disebut estimasi rotasi merupakan teknik validasi model untuk mengevaluasi bagaimana hasil analisis statistik diterapkan pada kumpilan data independent. Teknik ini terutama digunakan untuk melakukan prediksi model dan memperkirakan seberapa akurat model prediksi sebenarnya. Salah satu teknik validasi silang adalah k-*fold cross-validation*, yang membagi data menjadi K bagian kumpulan data dengan ukuran yang sama. Penggunaan 5-*fold cross-validation* untuk menghilangkan bias data. Pelatihan dan pengujian dilakukan sebanyak K (Azis dkk., 2020). Alur kerja *cross-validation* ditunjukkan di bawah ini.



Gambar 2.3. Simulasi *cross-validation* (sumber: (Azis dkk., 2020))

2.2. Teori Pendukung

Teori pendukung membantu konteks penelitian, termasuk isu kemiskinan dan program bantuan sosial.

2.2.1. Kemiskinan

Kemiskinan merupakan keadaan dimana seseorang atau sekelompok orang, baik laki-laki maupun perempuan tidak terpenuhi hak-hak dasar mereka dan menjalani kehidupan yang bermartabat, sebagaimana tercantum dalam Keputusan Presiden Republik Indonesia Nomor 7 Tahun 2005 tentang Rencana

Pembangunan Jangka Menengah Nasional (RPJMN) untuk periode 2004-2009 (Almuqorobin, 2021). Sedangkan menurut BPS, kemiskinan dapat diartikan sebagai terbatasnya akses terhadap kecukupan asupan kalori makanan dan non makanan (Nuraeni dkk., 2024).

2.2.2. Program Keluarga Harapan

Program Keluarga Harapan atau yang dikenal dengan PKH merupakan program bantuan sosial yang memberikan bantuan keuangan kepada Rumah Tangga Sangat miskin (RTSM) berdasarkan syarat dan ketentuan yang ditetapkan melalui pemenuhan komitmen (Nuraeni dkk., 2024). Program ini menyasar keluarga yang memenuhi kriteria tertentu untuk membatasi perilaku buruk (Nurjannah dkk., 2023). Untuk memastikan efektivitas dan keberlanjutan program PKH, beberapa variabel penting yang menjadi indikator dalam evaluasi program ini meliputi (Kemensos, 2024) dan (Budiono, 2021):

1. Usia penerima, memperhatikan rentang usia penerima bantuan untuk memastikan sasaran program. Penerima bantuan umumnya berusia $\geq$ 60 tahun.
2. Pekerjaan penerima, menilai jenis pekerjaan penerima bantuan untuk memahami kondisi ekonomi.
3. Penghasilan penerima, mengukur pendapatan rumah tangga untuk menentukan tingkat kesejahteraan. Penerima bantuan umumnya memiliki penghasilan $\leq$ 1.000.000 rupiah perbulan.
4. Jumlah anggota keluarga, menilai jumlah anggota keluarga dalam rumah tangga penerima bantuan. Rumah tangga dengan $\geq$ 4 anggota lebih diutamakan.
5. Jumlah tanggungan anak, memperhatikan jumlah anak yang menjadi tanggungan penerima bantuan. Keluarga dengan $\geq$ 2 anak lebih diutamakan.
6. Status rumah, menilai kepemilikan dan kondisi rumah sebagai indikator kesejahteraan. Penerima bantuan umumnya tinggal di rumah sendiri atau rumah orang tua.

7. Jenis lantai rumah, mengukur jenis lantai rumah untuk memahami kondisi kehidupan penerima bantuan. Jenis lantai yang diperhatikan adalah kayu, semen.
8. Jenis dinding rumah, menilai jenis dinding rumah untuk mendapatkan gambaran tentang kondisi tempat tinggal penerima bantuan. Jenis dinding yang diperhatikan adalah kayu, batu bata.

DAFTAR PUSTAKA

Aldiyansyah, A., dkk, 2024, Perbandingan Tingkat Akurasi Algoritma Decision Tree dan Random Forest dalam Mengklasifikasi Penerima Bantuan Sosial BPNT Di Desa Slangit. *Jurnal Mahasiswa Teknik Informatika*, No.1, Vol.8, 127-132, : https://ejournal.itn.ac.id/index.php/jati/article/view/8290.

Ali, M. M., dkk, 2022, Metodologi Penelitian Kuantitatif Dan Penerapan Nya. *JPIB : Jurnal Penelitian Ibnu Rusyd*, No.2, Vol.1, 1–5, : https://ojs.stai-ibnurusyd.ac.id/index.php/jpib/article/view/86.

Almuqorobin, A. R., 2021, Klasifikasi Kelayakan penerima Bantuan Program Keluarga Harapan Menggunakan Random Forest (Studi Kasus: Kelurahan Desa Balerejo), *Skripsi*, Fakultas Teknik, Univ. Muhammadiyah Magelang, Magelang.

Amanda, A., dkk., 2022, Komunikasi dan Sumber Daya dalam Implementasi Program E-Warong Kube PKH. *Economics, Social and Humanities Journal (Esochum)*, No.2, Vol.1, 2798-6926, : https://jurnal.unupurwokerto.ac.id/index.php/esochum.

Annisya, N. M. O., & Novira, A., 2023, Implementasi Program Keluarga Harapan (PKH) di Kelurahan Kampung Seraya Kecamatan Batu Ampar Kota Batam. *Jurnal Wacana Kinerja: Kajian Praktis-Akademis Kinerja dan Administrasi Pelayanan Publik*, No.1, Vol.26, 29-50, : https://doi.org/10.31845/jwk.v26i1.810.

Azis, H., dkk., 2020, Performa Klasifikasi K-NN dan Cross Validation Pada Data Pasien Pengidap Penyakit Jantung. *ILKOM Jurnal Ilmiah*, No.2, Vol.12, 81–86, : https://doi.org/10.33096/ilkom.v12i2.507.81-86.

Budiono, E., 2021, Kemensos Tetapkan Sembilan Kriteria Kemiskinan. *InfoPublik Portal Berita Info Publik*, https://infopublik.id/kategori/nasional-sosial-budaya/582705/kemensos-terapkan-sembilan-kriteria-kemiskinan, diakses pada tgl 3 Agustus 2024.

Damuri, A., dkk., 2021, Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako. *JURIKOM (Jurnal Riset Komputer)*, No.6, Vol.8, 219-225, : https://doi.org/10.30865/jurikom.v8i6.3655.

Darwanto, A. R. S., dkk., 2021, Analisis Regresi Logistik Binomial dan Algoritma Random Forest pada Proses Pengklasifikasian Penyakit Ginjal Kronis. *Jurnal Statistika dan Aplikasinya*, No.1, Vol.5, 1-14, : https://doi.org/10.21009/JSA.05101.

Ekasatria, H., 2019, Optimasi Klasifikasi Random Forest dengan Algoritma Genetika untuk Identifikasi Mutu Beras, *Tesis*, Sekolah Pascasarjana, Institut Pertanian Bogor, Bogor.

Fachruddin, M. I, 2015, Perbandingan Metode Random Forest Classification dan Support Vector Machine untuk Deteksi Epilepsi Menggunakan Data Rekaman Electroencephalograph (EEG), *Skripsi*, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember, Surabaya.

Fauzan, A. C., & Hikmah, K., 2022, Implementasi Algoritma Naive Bayes dalam Analisis Polarisasi Opini Masyarakat Terkait Vaksin Covid-19. *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, No.2, Vol.7, 122–128, : https://doi.org/10.36341/rabit.v7i2.2403.

Haidar, D., dkk., 2023, Penerapan Deep Learning Model Random Forest untuk Prediksi Penerima Bantuan Program Keluarga Harapan (PKH). *Jurnal Mahasiswa Teknik Informatika* No.6, Vol.7, 3564-3571, : https://ejournal.itn.ac.id/index.php/jati/article/view/8250/4866.

Hana, F. M., dkk., 2023, Implementasi Algoritma CART dalam Klasifikasi Penyakit Diabetes. Dalam *Jurnal Ilmu Komputer dan Matemtika*, 1-8.

Hidayati, A. R., dkk., 2023, Analisa Sentimen Pemilu 2019 Pada Judul Berita Online Menggunakan Metode Logistic Regression, *KESATRIA:Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, No.2, Vol.4, 298-305, : https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/164.

Hidayatulloh, N. G. T., 2022, Perbandingan Kinerja Random Forest dan Double Random Forest untuk Klasifikasi Status Kemiskinan di Level Kabupaten/Kota, *Skripsi*, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor.

Kemensos, 2024, Program Keluarga Harapan (PKH), *Kementerian Sosial Republik Indonesia*, https://kemensos.go.id/program-keluarga-harapan-pkh, diakses pada tgl 3 Agustus 2024.

Muti, D., & Pamuji, F. Y., 2023, Analisis Perbandingan Metode Naïve Bayes dan K-NN dalam Klasifikasi Kelayakan Keluarga Terdaftar DTKS Penerimaan Bantuan Sosial di Desa Dubesi, *Seminar Nasional Sistem Informasi*, UNMER Malang, 7 September.

Novita, A., dkk., 2022, Klasterisasi Provinsi Di Indonesia Berdasarkan Produktivitas Komoditas Pangan Menggunakan Algoritma K-Means. *Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)*, Jakarta-Indonesia, 20 Agustus.

Nuraeni, S., dkk., 2024, Analisis Akurasi Naïve Bayes dan KNN dalam Penentuan Penerima PKH Di Lombok Utara. *Journal of Information System Management (JOISM),* No.2, Vol.5, 121-126, : https://doi.org/10.24076/joism.2024v5i2.1205.

Nurjannah, dkk., 2023, Eligibility Classification of Aid Recipients Hope Family Program in Cinta Rakyat Village Using the Method Weighted Naive Bayes with Laplace Smoothing. *Jurnal Matematika, Statistika dan Komputasi*, No.2, Vol.20, 440–454, : https://doi.org/10.20956/j.v20i2.32069.

Panharesi, G. A., & Anugrah, I. G., 2022, Menggunakan Metode Weighted Naive Bayes (Studi Kasus: Program Studi Teknik Informatika Universitas Muhammadiyah Gresik). *INDEXIA: Informatic and Computational Intelligent Journal,* No.1, Vol.4, 33–46, : https://journal.umg.ac.id/index.php/indexia/article/view/3589.

Purwa, T., 2019, Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017). *Jurnal Matematika, Statistika dan Komputasi*, No.1, Vol.16, 58-73, : https://doi.org/10.20956/jmsk.v16i1.6494

Ratnasari, D., & Hasibuan, N. A., 2017, Penerapan Teorema Bayes dalam Memprediksi Bayi Terlahit Cacat, *Jurnal Media Informatika Budidarma*, No.3, Vol.1, 62-66, : https://ejurnal.stmik-budidarma.ac.id/index.php/mib/article/view/523/0.

Siahaan, D., dkk., 2016, Aplikasi Classification and Regression Tree (CART) dan Regresi Logistik Ordinal dalam Bidang Pendididikan (Studi Kasus: Predikat Kelulusan Mahasiswa S1 Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman). *Jurnal EKSPONENSIAL*, No.1, Vol.7, 95-104, : https://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/46.

Sidabutar, A. F., dkk., 2023, Perbandingan Metode Klasifikasi untuk Pengelompokkan Risiko Magang Mahasiswa. *Jurnal Mahasiswa Teknik Informatika*, No.3, Vol.7, : https://ejournal.itn.ac.id/index.php/jati/article/view/7026.

Siregar, A. P., dkk., 2023, Implementasi Algoritma Random Forest Dalam Klasifikasi Diagnosis Penyakit Stroke. *Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)*, No.4, Vol.2, 155–164, : https://doi.org/10.55606/juprit.v2i4.3039.

Sumartini, S. H., & Purnami, S. W., 2015, Penggunaan Metode Classification and Regression Trees (CART) untuk Klasifikasi Rekurensi Pasien Kanker Serviks di RSUD Dr. Soetomo Surabaya. *Jurnal Sains dan Seni ITS*, No.2,

Vol.4, 2337–3520, : https://ejurnal.its.ac.id/index.php/sains_seni/article/view/10673.

Utami, D., & Devi, P. A. R., 2022, Klasifikasi Kelayakan Penerima Bantuan Program Keluarga Harapan (PKH) Menggunakan Metode Weighted Naïve Bayes dengan Laplace Smoothing. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, No.04, Vol.07, 1373-1384, : https://jurnal.stkippgritulungagung.ac.id/index.php/jipi/article/view/3592.