

KOMPARASI EFEKTIFITAS ALGORITMA C4.5 DAN NAÏVE BAYES UNTUK MENENTUKAN KELAYAKAN PENERIMA MANFAAT PROGRAM KELUARGA HARAPAN (STUDI KASUS : KECAMATAN CICALENGKA KABUPATEN BANDUNG)

Adi Nur Abdul Rohim¹, Ade Irma Purnamasari², Irfan Ali³

^{1,2} Teknik Informatika, STMIK IKMI Cirebon

³ Teknik Rekayasa Perangkat Lunak, STMIK IKMI Cirebon

Jl. Perjuangan No.10B Majasem, Cirebon 4513, Indonesia

irma2974@yahoo.com

ABSTRAK

Program keluarga harapan PKH adalah sebuah program bantuan sosial yang disalurkan oleh pemerintah Indonesia melalui kementerian sosial. Bantuan yang diberikan merupakan bantuan keuangan kepada rumah tangga sangat miskin yang telah memenuhi kriteria tertentu. Namun program pemberdayaan pemerintah ini, seringkali timbul masalah dimana bantuan PKH tidak tepat sasaran. Oleh karena itu, penelitian ini bertujuan untuk menganalisis program keluarga harapan agar penerima bantuan benar benar layak menerima bantuan. Digunakan komparasi metode klasifikasi data mining untuk mengetahui algoritma mana yang baik untuk mengkalasifikasi kelayakan bantuan, dengan menggunakan dua algoritma, algoritma C4.5 dan Naïve Bayes. Algoritma C4.5 adalah salah satu algoritma *decision tree* yang digunakan untuk melakukan klasifikasi atau segmentasi yang bersifat prediktif. Kelebihan algoritma C4.5 yaitu menghasilkan model pohon keputusan yang mudah diinterpretasikan, memiliki tingkat akurasi yang tinggi, efisien dalam menangani atribut bertipe diskrit dan numerik. Algoritma Naïve Bayes adalah algoritma yang menggunakan cabang matematika dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. Naïve Bayes adalah tehnik yang diterapkan untuk menentukan kelas dari tiap masalah, yang sudah dibagi berdasarkan tiap-tiap masalah. Perhitungan numerik berdasarkan pada pendekatan grup. Proses pengujian dimulai dengan persiapan data awal. Selanjutnya, sampel data dianalisis menggunakan metode CRISP-DM, diuji dengan algoritma seperti C4.5 dan Naïve Bayes untuk memperoleh validasi yang akurat, setelah itu, data sampel data kembali diuji dengan validasi untuk mendapatkan *Confusion matriks* dan nilai ROC untuk C4.5 dan Naïve Bayes dengan tujuan mencapai akurasi maksimum. Hasil evaluasi perbandingan dan validasi menunjukkan bahwa algoritma Naive Bayes memiliki nilai akurasi sebesar 99,87% dan AUC sebesar 1,000, yang merupakan yang tertinggi di antara metode lainnya. Sementara itu, Algoritma C4.5 memiliki akurasi sebesar 99,61% dan AUC sebesar 0,743. Dari hasil algoritma Naive Bayes dalam data mining ini, tingkat AUC mengindikasikan klasifikasi yang sangat baik (Excellent classification). Sebaliknya, Algoritma C4.5 memiliki tingkat AUC yang diagnostik sebagai Fair (Adil). Dengan demikian, dapat disimpulkan bahwa algoritma Naive Bayes merupakan metode yang cukup baik dalam memprediksi kelayakan warga untuk menerima bantuan dari Program Keluarga Harapan (PKH).

Kata kunci : Kelayakan PKH, Data Mining, C4.5, Naïve Bayes,

1. PENDAHULUAN

Dalam era modern ini, pesatnya perkembangan bidang Informatika telah mengubah lanskap teknologi dan membawa dampak signifikan pada berbagai aspek kehidupan. Kemajuan teknologi informasi telah memengaruhi cara kita berinteraksi, bekerja, belajar, dan bahkan membuat keputusan dalam berbagai aspek kehidupan. Salah satu dampak yang sangat relevan adalah penggunaan teknologi informasi dalam bidang kebijakan sosial, terutama dalam program-program pemerintah seperti Program Keluarga Harapan (PKH). Program Keluarga Harapan (PKH) adalah program pemberian bantuan sosial bersyarat kepada Rumah Tangga Sangat Miskin (RTSM) sejak tahun 2007. Program ini diberikan melalui bantuan tunai kepada keluarga sangat miskin berdasarkan persyaratan dan ketentuan yang telah ditetapkan. Program Perlindungan Sosial ini diberikan melalui konsep

Conditional Cash Transfers (CCT) dan dianggap cukup berhasil dalam menanggulangi kemiskinan yang dihadapi berbagai negara, terutama masalah kemiskinan kronis [1]

Perbandingan efektivitas algoritma C4.5 dan Naïve Bayes dalam menentukan kelayakan penerima manfaat Program Keluarga Harapan (PKH) merupakan isu yang sangat relevan dan bernilai untuk dipecahkan dalam konteks informatika. Program Keluarga Harapan (PKH) yang dibuat oleh pemerintah tersebut bertujuan untuk menanggulangi kemiskinan [2]. Permasalahan sering terjadi pada program pemberdayaan pemerintah ini, bahwa bantuan Program Keluarga Harapan (PKH) tidak tepat sasaran. Munculnya masalah tersebut, diakibatkan verifikasi data yang belum efektif terhadap pemilihan warga yang layak menerima bantuan PKH (Fitriani, 2020) Terbatasnya sumber daya bantuan Program Keluarga

Harapan (PKH) juga menurut penelitian yang telah dilakukan oleh [3]. Penting untuk memastikan bahwa manfaat program tersebut tepat sasaran. Oleh karena itu, dari kasus ini mendorong peneliti untuk melakukan penelitian dan memverifikasi kembali penerimaan bantuan program keluarga harapan (PKH) di wilayah kecamatan cicalengka dengan teknik data mining. Penggunaan algoritma Naive Bayes dalam menentukan kelayakan Program Keluarga Harapan (PKH) yang dilakukan oleh [4] hanya menggunakan satu algoritma, *Naive Bayes* merupakan metode yang tidak memiliki aturan. *Naive Bayes* menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi, dengan cara melihat frekuensi tiap klasifikasi pada data training. *Naive Bayes* adalah tehnik yang diterapkan untuk menentukan kelas dari tiap masalah, yang sudah dibagi berdasarkan tiap-tiap masalah. Perhitungan numerik berdasarkan pada pendekatan grup [5]. Peneliti menggunakan algoritma C4.5 sebagai komparasi atau pembandingan yang mana Algoritma C4.5 adalah salah satu algoritma decision tree yang digunakan untuk melakukan klasifikasi atau segmentasi yang bersifat prediktif. Kelebihan algoritma C4.5 yaitu menghasilkan model pohon keputusan yang mudah diinterpretasikan, memiliki tingkat akurasi yang tinggi, efisien dalam menangani atribut bertipe diskrit dan numerik [6]. Penggunaan dua algoritma pada penelitian ini merupakan pengembangan dari penelitian sebelumnya, dengan menggunakan dua algoritma ini dapat diketahui nilai akurasi mana yang paling baik dari kedua metode tersebut sehingga dari salah satu metode dengan nilai akurasi yang baik ini dapat diterapkannya data maining dalam megkalsifikasi bantuan program keluarga harapan (PKH) dan membantu pemerintah dalam memperbaiki verifikasi data yang menimbulkan bantuan pkh tidak tepat sasaran serta mendukung upaya meningkatkan efektivitas penerimaan bantuan Program Keluarga Harapan (PKH) dan program-program serupa yang dapat meningkatkan kesejahteraan masyarakat secara menyeluruh.

Penelitian terdahulu yang dilakukan oleh [4] terkait Penerapan Algoritma *Naive Bayes* dalam Prediksi Penerimaan Bantuan Program Keluarga Harapan (PKH). Berdasarkan eksperimen yang telah dilakukan, algoritma *Naive Bayes* menghasilkan akurasi sebesar 96%, sensitivitas sebesar 1,000, dan spesifisitas sebesar 0,9259 dengan menggunakan satu algoritma. Sementara penelitian yang dilakukan oleh [7] membahas tentang prediksi penerimaan bantuan sosial, Prediksi ini dibuat menggunakan aplikasi Rapid Miner dengan menerapkan lima model operator yang melibatkan tahapan *retrieve*, *cross-validation*, penggunaan model, dan evaluasi kinerja. Hasil dari penggunaan algoritma *Naive Bayes* dalam model, menunjukkan tingkat akurasi sebesar 90,10%, presisi kelas sebesar 91,85%, dan *recall* kelas sebesar 91,85% dengan menggunakan satu algoritma. Penelitian

terdahulu yang dilakukan oleh [8] menghasilkan 6 (Enam) rules dan tingkat akurasi yang dihasilkan oleh metode tersebut adalah 96,67%. Menggunakan 1 algoritma, Kesimpulannya adalah penulis ingin mengembangkan sistem yang lebih baik. Yaitu memanfaatkan lebih dari satu algoritma dengan tujuan untuk meningkatkan kualitas klasifikasi dan analisis, dengan membandingkan dua metode algoritma.

Penelitian ini bertujuan untuk mengisi kesenjangan pengetahuan dalam bidang Informatika dengan melakukan eksperimen terhadap penerapan masing masing algoritma, yaitu algoritma C4.5 dan algoritma *Naive Bayes*. Penelitian ini juga mencakup perbandingan antara algoritma C4.5 dan *Naive Bayes* untuk mengklasifikasikan kelayakan penerima bantuan PKH, dengan tujuan mengidentifikasi algoritma yang memberikan nilai akurasi tertinggi. Penelitian ini diharapkan dapat memberikan wawasan yang lebih komprehensif terkait kelebihan dan kelemahan masing-masing algoritma dalam konteks aplikatif yang spesifik. Hasil penelitian diharapkan dapat memberikan kontribusi positif pada pengembangan sistem verifikasi kelayakan penerima bantuan PKH, membuka ruang untuk pemahaman lebih mendalam tentang efektivitas dan keunggulan relatif dari algoritma C4.5 dan *Naive Bayes* dalam domain ini. Dengan demikian, penelitian ini diarahkan untuk memberikan kontribusi penting pada perkembangan dan pemahaman dalam bidang Informatika.

Dalam penelitian ini, dilakukan penerapan dari masing masing algoritma C4.5 dan algoritma *Naive Bayes*. Penerapan menggunakan algoritma C4.5 dalam mengklasifikasi bantuan program keluarga harapan (PKH), dilakukan dengan menghitung masing masing entropy dan information gain dari tiap-tiap atribut dataset yang sudah di olah. Sehingga diperoleh node atau simpul akar, setelah simpul akar didapatkan maka akan terbentuk pohon keputusan selanjutnya adalah menentukan rule atau aturan, rule atau aturan inilah yang digunakan dalam mengklasifikasi bantuan program keluarga harapan (PKH). Setelah diketahui rule atau aturan dalam mengklasifikasi maka dibuatkan model k-fold cross validation, dan confusion matrix yang mana akan diketahui dari data awal klasifikasi yang diperoleh dari kecamatan, hingga telah diproses menggunakan algoritma C4.5. Penerapan pada algoritma *Naive Bayes* dengan menentukan probablitas hipotesis untuk masing-masing kelas, $P(H)$. Terdapat dua hipotesis yang ada, yakni warga layak untuk menerima bantuan Program Keluarga Harapan (PKH) dan warga yang tidak layak untuk menerima bantuan (PKH). Data training yang digunakan serupa dengan pengujian algoritma C4.5, yang terdiri dari total 774 data, dengan 477 data warga yang layak untuk menerima bantuan program keluarga harapan (PKH) dan 297 data yang tidak layak untuk menerima bantuan program keluarga harapan (PKH), Setelah menentukan probablitas untuk setiap hipotesis, langkah berikutnya adalah menghitung probablitas

kondisi tertentu (probabilitas X) berdasarkan probabilitas setiap hipotesis (probabilitas H), yang juga dikenal sebagai probabilitas prior. Setelah diketahui probabilitas prior maka dibuatkan model *k-fold cross validation*, dan *confusion matrix* yang mana akan diketahui dari data awal klasifikasi yang diperoleh dari kecamatan, hingga telah diproses menggunakan algoritma *naïve bayes*. dilakukan pengujian model dengan membandingkan dua metode data mining, yaitu algoritma C4.5 dan *Naïve Bayes*, menggunakan data warga dari salah satu kecamatan di kabupaten Bandung yaitu kecamatan cicalengka. Data tersebut terdiri dari warga yang layak untuk menerima bantuan PKH dan warga yang tidak layak menerima bantuan Program Keluarga Harapan, dengan total data sebanyak 774 data warga. Pengujian data akan dilakukan menggunakan perangkat *Rapid Miner*, dan setiap algoritma akan menghasilkan nilai-nilai seperti akurasi, presisi, recall, dan AUC. Hasil evaluasi dan validasi akan menunjukkan bahwa algoritma C4.5 memiliki tingkat akurasi yang akan dilakukan dan akan diketahui nilai AUC, nilai AUC ini adalah nilai yang tertinggi dibandingkan dengan metode lainnya. Sementara itu, metode *Naive Bayes* akan diketahui nilai akurasinya tentunya nilai AUC juga akan diketahui. Dari kedua metode data mining tersebut, diperkirakan nilai AUC akan mengindikasikan klasifikasi yang sangat baik. Oleh karena itu, dapat diperkirakan bahwa algoritma C4.5 merupakan metode yang efektif dalam memprediksi kelayakan warga untuk menerima bantuan Program Keluarga Harapan (PKH).

Perbandingan antara algoritma Naive Bayes dan C4.5 dalam prediksi penerimaan bantuan PKH memberikan wawasan penting bagi peneliti dan praktisi informatika, membantu dalam pemilihan algoritma yang tepat. Peningkatan akurasi dalam penyaluran bantuan dapat mengurangi penyimpangan dan frustrasi di kalangan penerima manfaat. Kontribusi pada bidang informatika juga merangsang pembahasan tentang peran teknologi informasi dalam administrasi program sosial, dengan potensi dampak positif pada perkembangan teknologi dan efisiensi sektor-sektor lainnya, memberikan manfaat yang lebih luas bagi masyarakat.

2. TINJAUAN PUSTAKA

2.1. Program Keluarga Harapan (PKH)

Program Keluarga Harapan (PKH) Merupakan suatu inisiatif bantuan sosial yang diterapkan sejak tahun 2007, Program Perlindungan Sosial ini ditujukan khusus kepada Rumah Tangga Sangat Miskin (RTSM). Bantuan ini disalurkan dalam bentuk tunai kepada keluarga yang memenuhi persyaratan dan ketentuan yang telah ditetapkan. Program Perlindungan Sosial ini mengadopsi konsep Conditional Cash Transfers (CCT) dan telah terbukti relatif berhasil dalam mengatasi masalah kemiskinan, terutama dalam mengatasi masalah kemiskinan kronis di berbagai negara (<https://pkh.kemsos.go.id>, 2019).

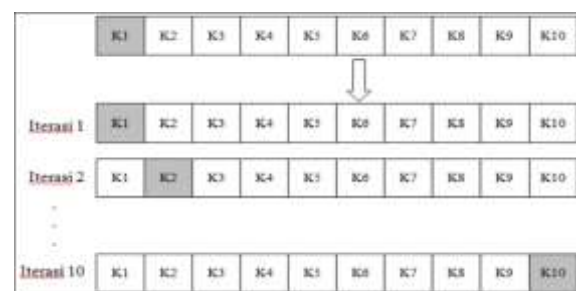
Temuan ini didukung oleh hasil penelitian yang telah dilakukan oleh[9]

2.2. Algoritma C4.5

Algoritma C4.5 menurut penelitian yang telah dilakukan oleh [10] Algoritma C4.5 adalah suatu metode dalam bidang data mining yang dirancang untuk konstruksi model pohon keputusan. Dikembangkan sebagai penyempurnaan dari Algoritma ID3 yang dipelopori oleh J. Ross Quinlan pada tahun 1996, algoritma C4.5 mendasarkan prinsip kerjanya pada algoritma ID3. Meskipun demikian, C4.5 memiliki keunggulan dan perbaikan tertentu dibandingkan dengan pendahulunya. Algoritma C4.5 termasuk dalam kelompok algoritma decision tree, yang berfokus pada penggunaan data latihan yang telah disediakan untuk membangun model pohon keputusan.[11] Algoritma C4.5, yang termasuk dalam kategori decision tree, telah dikenal luas dan diakui karena berbagai keunggulan yang dimilikinya. Salah satu keunggulan utama dari algoritma C4.5 adalah kemampuannya menghasilkan model pohon keputusan yang dapat diinterpretasikan dengan mudah. Selain itu, algoritma ini juga dikenal memiliki tingkat akurasi yang tinggi dan efisiensi yang baik dalam menangani atribut baik yang bersifat diskrit maupun numerik.

2.3. K-Fold Cross Validation

K-fold cross validation adalah metode statistika yang digunakan untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen, dimana yang satu digunakan menjadi *learning* atau *training* model dan lainnya digunakan untuk memvalidasi model. Cara kerja metode ini adalah set pelatihan dan validasi data harus disilangkan secara berturut-turut sehingga masing-masing titik memiliki kesempatan untuk divalidasi Pada proses *k-fold cross validation*, data dibagi sejumlah k dan data ke- k digunakan sebagai data testing, sedangkan data selain ke- k sebagai data pembentukan pola atau data training. Pada proses fold ke- k , parameter nilai k digunakan untuk menguji validasi data terhadap data *testing* menggunakan algoritma C4.5. Berdasarkan penelitian yang dilakukan oleh [10]



Gambar 1. K-fold Cross Validation

Gambar 1 Mengindikasikan bahwa pada tahap iterasi pertama, gambar yang teridentifikasi di dalam

kolom pertama dijadikan sebagai data uji (*testing*), sementara sembilan kolom lainnya dijadikan sebagai data latih (*training*), dan pola ini berulang hingga mencapai iterasi kesepuluh.

2.4. Evaluasi dan validasi model Menggunakan Confusion Matrix

Kemiskinan merupakan fenomena dan permasalahan sosial[12], Setiap algoritma akan dinilai berdasarkan akurasi, presisi, recall, dan nilai f-measure. Evaluasi klasifikasi dilakukan dengan menguji objek yang benar dan objek yang salah. Validasi digunakan untuk menentukan jenis model terbaik melalui confusion matriks sebagai informasi tentang hasil klasifikasi aktual yang dapat diprediksi oleh suatu sistem, yang diukur melalui nilai akurasi, presisi, dan recall dengan menggunakan persamaan sebagai berikut.

$$Akurasi = \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

$$(Presisi/Specificity) = \frac{TP}{(TP+FP)}$$

$$Recall/Sensitivity = \frac{TP}{TP+FN}$$

$$f - \text{measure} = \frac{(2 \times (\text{Presisi} \times \text{recall}))}{(\text{presisi} + \text{recall})}$$

2.5. Evaluasi dan validasi menggunakan Kurva ROC

Kurva ROC merupakan suatu alat pengukur yang digunakan untuk mengevaluasi kemampuan klasifikasi yang memanfaatkan dua kelas keputusan, di mana setiap subjek dikategorikan ke dalam salah satu kelompok pendukung, positif atau negatif. Pada kurva ROC, tingkat *True Positive* (TP) direpresentasikan pada sumbu Y, sedangkan tingkat *False Positive* (FP) direpresentasikan pada sumbu X. Dalam konteks pengklasifikasian data *mining*, nilai *Area Under the Curve* (AUC) dapat dibagi menjadi beberapa kategori, seperti *Excellent Classification*, *Good Classification*, *Fair Classification*, *Poor Classification*, dan *Failure*. [12]

Performance	Penggolongan
0.90-1.00	Penggolongan Sempurna
0.80-0.90	Penggolongan Baik
0.70-0.80	Penggolongan Adil
0.60-0.70	Penggolongan Cukup
0.50-0.60	Penggolongan Tidak Baik

Gambar 2. Performance ketepatan AUC

2.6. RapidMiner

Menurut penelitian yang telah dilakukan oleh [13] *RapidMiner* merupakan sebuah perangkat lunak dengan karakteristik open source yang menyediakan solusi untuk melakukan analisis pada data mining, text mining, dan analisis prediksi. Dengan menggunakan beragam teknik deskriptif dan prediksi, *RapidMiner*

bertujuan memberikan wawasan kepada pengguna sehingga mereka dapat membuat keputusan yang optimal.

3. METODE PENELITIAN

Langkah-langkah yang diterapkan dalam penelitian ini untuk menentukan pengumpulan data hingga pengujian data warga dalam klasifikasi penerima program bantuan Program Keluarga Harapan (PKH) melibatkan proses sebagai berikut:

3.1. Pengumpulan Data dan pengujian data

a. Jenis Penelitian

Dalam konteks penelitian ini, dilakukan implementasi perbandingan antara dua algoritma klasifikasi data mining, yakni algoritma C4.5 dan Naive Bayes. Dalam tahap pengujian kedua metode, akan dipilih metode yang menunjukkan tingkat akurasi terbaik.

b. Metode Pemilihan populasi dan sampel

Populasi dalam konteks penelitian ini mengacu kepada data masyarakat di wilayah kecamatan cicalengka yang telah diidentifikasi sebagai kandidat yang layak untuk menerima manfaat dari program keluarga harapan (PKH), serta individu yang tidak layak untuk menerima manfaat tersebut selama periode tahun 2022. Data populasi ini diperoleh dari sumber data internal yang tersedia di salah satu kecamatan yang terletak di wilayah kecamatan cicalengka kabupaten bandung. Sampel dalam kerangka penelitian ini mencakup sub kelompok tertentu yang terdiri dari individu yang telah memenuhi syarat untuk menerima manfaat dari Program Keluarga Harapan (PKH) dan individu yang tidak memenuhi syarat tersebut. Data sampel ini memiliki sifat internal dan belum pernah diungkapkan secara publik oleh pihak kecamatan. Informasi lebih rinci mengenai jumlah sampel yang akan digunakan dapat ditemukan dalam contoh Tabel

Tabel 1. Populasi Sampel

Layak	Tidak Layak	Total Sampel
477	297	774

3.2. Teknik Pengumpulan Data

Metode pengumpulan data dalam penelitian ini melibatkan pelaksanaan wawancara dengan kepala seksi kesejahteraan sosial di kecamatan cicalengka dan kepala sub kepegawaian kecamatan cicalengka untuk mendapatkan surat rujukan dari kecamatan, ke badan kesatuan bangsa dan politik kabupaten bandung, dilanjutkan dengan permintaan izin dari dinas sosial sebagai persetujuan untuk mengambil data dari kecamatan di ikuti dengan penghimpunan data berdasarkan informasi yang tersedia di kantor kecamatan. Data yang dianalisis terdiri dari data masyarakat yang layak dan data masyarakat yang tidak layak mendapatkan bantuan program keluarga harapan (PKH) Kecamatan Cicalengka Kabupaten Bandung.

3.3. Teknik Analisis Data

Teknik analisis data yang digunakan dalam Penelitian ini mengadopsi metodologi yang digunakan dalam data mining, yaitu Model Cross-Industry Standard Process for Data Mining (CRISP-DM) yang terdiri dari enam tahap proses, yakni: pemahaman bisnis (Business Understanding), pemahaman data (Data Understanding), persiapan data (Data Preparation), pemodelan (Modeling), evaluasi (Evaluation), dan penyebaran (Deployment).[10]

- Tahap Pengolahan Data (*Business Understanding*)**
Penelitian ini dimotivasi oleh adanya indikasi ketidakakuratan dalam penetapan penerima PKH. Penelitian ini bertujuan untuk mengembangkan komparasi antara algoritma klasifikasi C4.5 dan Naïve Bayes dengan tujuan untuk menilai kelayakan penerima PKH dan meningkatkan akurasi dari algoritma klasifikasi tersebut
- Tahap Pemahaman Data (*Data Understanding*)**
Pada tahap ini, data dikumpulkan, dianalisis secara eksploratif untuk memahami lebih lanjut karakteristik data, dan dilakukan evaluasi terhadap kualitas data. Sumber data utama yang digunakan adalah data populasi di salah satu kecamatan di kabupaten bandung dengan 17 atribut. Analisis meliputi pemilihan data yang akan digunakan dan identifikasi kriteria kelayakan.
- Tahap Pengolahan Data (*Data Preparation*)**
Pada tahap ini, dari total 774 data yang mencakup 477 warga yang layak dan 297 tidak layak menerima PKH dengan 17 atribut, dilakukan pemilihan data yang relevan untuk penelitian. Dataset lengkap dapat dilihat pada: https://drive.google.com/drive/folders/1mPGXCuRXymPDKMkjc6SCkvDGHnAI4ax7?usp=drive_link

Tabel 2. Datset Validasi I

No	Nama	Pekerjaan	Status Bangunan	Lantai	Luas lantai	Dinding	Usia Dini	Ibu Hamil	Lansia	Disabilitas	Anak SD	Anak SMP	Anak SMA	Sumber Air Minum	Jenis Bahan Bakar Memasak	Sumber Penerangan	Aset	Kelayakan
1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Tabel 3. Dataset Validasi II

No	Nama	Pekerjaan	Status Bangunan	Lantai	Luas lantai	Dinding	Usia Dini	Ibu Hamil	Lansia	Disabilitas	Anak SD	Anak SMP	Anak SMA	Sumber Air Minum	Jenis Bahan Bakar Memasak	Sumber Penerangan	Aset	Kelayakan
1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

d. *Selecting Data*

Tahap seleksi data digunakan untuk memilih data dan atribut yang relevan dan sesuai dengan tujuan data mining. Setelah mengumpulkan semua data yang diperlukan, langkah berikutnya adalah melakukan proses penyaringan untuk mencari kesesuaian data penyesuaian dilakukan terhadap data validasi I penerima PKH dan validasi II

penerima PKH. Proses pencarian kesesuaian data dilakukan dengan meminta data tambahan kepada kepala pendamping koordinator cabang Kecamatan Cicalengka data tambahan berupa data data rekapan pendataan kemiskinan wilayah Kecamatan Cicalengka, data yang berhasil ditemukan kemudian disimpan dalam sebuah tabel baru yang diberi nama "Proses_Akhir" sebagaimana terlihat pada tabel 4 berikut.

Tabel 4. Proses_Akhir

No	Nama	Pekerjaan	Status Bangunan	Lantai	Luas lantai	Dinding	Usia Dini	Ibu Hamil	Lansia	Disabilitas	Anak SD	Anak SMP	Anak SMA	Sumber Air Minum	Jenis Bahan Bakar Memasak	Sumber Penerangan	Aset	Kelayakan
1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Atribut yang dihasilkan pada tahap ini sejumlah 17 atribut yang dipilih berdasarkan rujukan dari beberapa hasil penelitian seperti yang termuat dalam literature review. Atribut yang dipilih yaitu nama, pekerjaan, status bangunan, lantai, luas lantai, dinding, anak usia dini, ibu hamil, lansia, disabilitas, anak SD, SMP, SMA, sumber air minum, pendidikan orang tua, jumlah tanggungan, status bangunan, luas lantai, jenis lantai, jenis atap, jenis dinding, sumber air minum, jenis bahan bakar masak, sumber penerangan, aset, dan kelayakan.

e. *Cleaning data*

Pada tahap ini akan dilakukan cleaning data pada tabel data yang terdapat noise. Proses pembersihan data pada tabel menggunakan menu filter untuk menyaring record yang kosong. Pada tahap ini terdapat record kosong pada atribut yang tidak dipakai, total record pada tabel proses yang digunakan sebanyak 774 dan 24 atribut tahap, proses cleaning data menghasilkan 774 record dan 17 attribute. Berikut daftar atribut yang digunakan:

Tabel 5. atribut yang digunakan

Atribut
Nama
Pekerjaan
Status Bangunan
Lantai
Luas lantai
Dinding
Usia Dini
Ibu Hamil
Lansia
Disabilitas
Anak SD
Anak SMP
Anak SMA
Sumber Air Minum
Jenis Bahan Bakar Memasak
Sumber Penerangan
Aset
Kelayakan

f. *Constructing new data*

Pada langkah ini informasi yang terdapat dalam tabel proses akhir di simpan untuk kemudian

dimanfaatkan dalam tahap pemodelan data. Jumlah total data yang dihasilkan mencapai 774 record, yang terdiri dari 477 layak dan 229 tidak layak. Jumlah keseluruhan atribut yang digunakan adalah sebanyak 17 atribut. Seperti yang terlihat pada tabel 6.

Tabel 6. Dataset telah melalui *preprocessing*

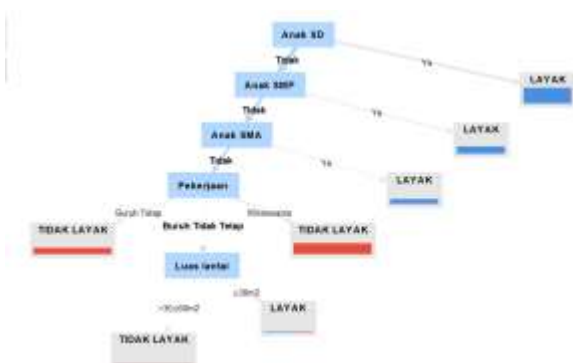
[illegible]

- g. Tahap Pemodelan (*Modeling*)
Dilakukan proses pelatihan data untuk menghasilkan aturan-aturan yang digunakan dalam pembentukan pohon keputusan. Penelitian ini membandingkan dua algoritma klasifikasi, yaitu C4.5 dan *Naïve Bayes*
- h. Tahap Evaluasi (*Evaluation*)
Tahap Evaluasi (*Evaluation*) pada penelitian ini mencakup hasil evaluasi dari percobaan operator Cross-Validation pengujian K-fold, klasifikasi algoritma C4.5 dan *Naïve Bayes* menggunakan (confusion matrix) dan kurva Receiver Operating Characteristic (ROC).
- i. Tahap penyebaran (*Deployment*)
Tahap Penyebaran (*Deployment*) Hasil penelitian ini akan digunakan sebagai dasar untuk menentukan kelayakan penerima bantuan Program Keluarga Harapan (PKH). Sehingga dapat menghasilkan hasil yang sesuai dengan tujuan awal proses CRISP-DM, dengan output akhir dari tahap evaluasi adalah penyusunan laporan.

4. HASIL DAN PEMBAHASAN

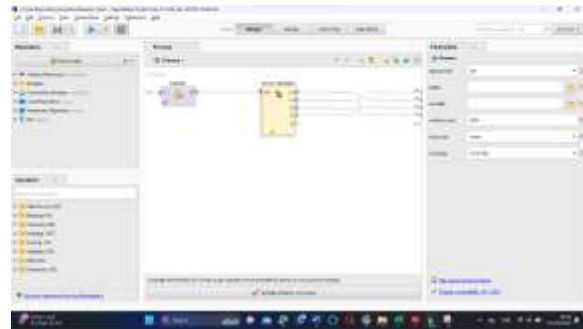
Hasil pengujian data warga, yang dijalankan melalui *tools* RapidMiner menggunakan algoritma C4.5 dan *Naïve Bayes*, masing masing menghasilkan metrik evaluasi seperti *K-Fold Cross Validation*, akurasi (*accuracy*), matriks kebingungan (*confusion matrix*), dan kurva ROC (*Receiver Operating Characteristic*).

4.1. Hasil Experimen algoritma C4.5

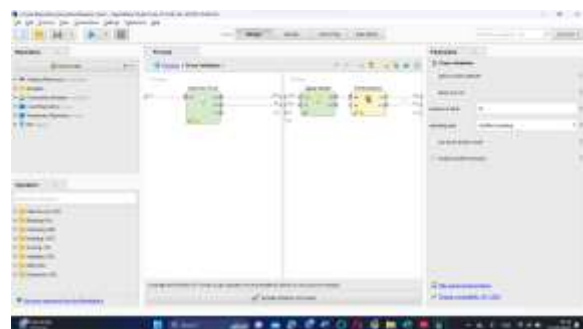


Gambar 3. Pohon Keputusan

Dapat disimpulkan bahwa dalam klasifikasi penerima bantuan PKH, atribut anak SD berperan sebagai root atau akar dari pohon keputusan.



Gambar 4. Proses Cross Validation



Gambar 5. Susunan Proses Cross Validation

Hasil optimal dalam pengklasifikasian data warga penerima Program Keluarga Harapan (PKH) dengan menggunakan Algoritma C4.5 dapat ditemukan pada gambar 6 berikut ini:



Gambar 6. Hasil Pengujian Nilai Terbesar

Hasil pengujian terbesar di ambil dari nilai *K-fold* terbesar. Didapatkan nilai akurasi diketahui dari 774 data, 477 diklasifikasikan Layak sesuai dengan prediksi yang dilakukan dengan metode Algoritma C4.5, lalu 5 data diprediksi Layak tetapi ternyata hasilnya Tidak, 292 data *class* Tidak diprediksi sesuai, dan 1 data diprediksi Tidak ternyata Layak. Berdasarkan Tabel 4.12 tersebut menunjukkan bahwa, tingkat akurasi dengan menggunakan algoritma C4.5 adalah sebesar 99.61%



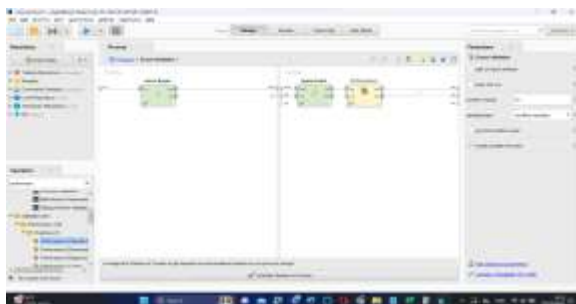
Gambar 7. Grafik ROC Algoritma C4.5

Terdapat grafik ROC dengan nilai AUC (Area Under Curve) sebesar 1.000 dimana hasilnya dapat dinyatakan sebagai *Excellent Classification* karena *Performance* keakurasian AUC 0.930 ada pada rentang 0.900 – 1.000 termasuk *performace Excellent Clasification*.

4.2. Hasil Experimen Algoritma Naïve Bayes



Gambar 8. Proses Cross Validation



Gambar 9. Proses Susunan Cross Validation



Gambar 10. Hasil Pengujian Nilai Terbesar

Diketahui dari 774 data, di mana 477 diidentifikasi sebagai kelas Layak sesuai dengan prediksi yang dilakukan menggunakan metode Algoritma *Naïve Bayes*. Selanjutnya, terdapat satu data

yang awalnya diprediksi sebagai kelas Layak, namun hasilnya menunjukkan kelas Tidak. Sebanyak 296 data diprediksi dengan benar sesuai dengan kelasnya, sedangkan tidak ada data yang diprediksi sebagai kelas Tidak yang pada akhirnya menunjukkan kelas Layak.



Gambar 11. Grafik ROC Algoritma Naïve Bayes

Dari Gambar 11 terdapat grafik ROC dengan nilai AUC (Area Under Curve) sebesar 1.000 dimana hasilnya dapat dinyatakan sebagai *Excellent Classification* karena *Performance* keakurasian AUC 0.930 ada pada rentang 0.900 – 1.000 termasuk *performace Excellent Clasification*.

5. KESIMPULAN DAN SARAN

Diketahui dari 774 data, 477 diklasifikasikan Layak sesuai dengan prediksi yang dilakukan dengan metode Algoritma C4.5, lalu 5 data diprediksi Layak tetapi ternyata hasilnya Tidak, 292 data class Tidak diprediksi sesuai, dan 1 data diprediksi Tidak ternyata Layak. tingkat akurasi dengan menggunakan algoritma C4.5 adalah sebesar 99,61%. Pada pengujian cross validation dapat disimpulkan bahwa pengujian dengan menggunakan 9 kali pengujian (fold 10), diperoleh nilai rata-rata akurasi sebesar 99,26%. Rata-rata presisi mencapai 99,09%, dan rata-rata recall mencapai 99,13%. tingkat akurasi tertinggi dengan menggunakan Algoritma C4.5 mencapai 99,61%. Dari pengujian K-Fold Teringgi yaitu Fold Ke 8 Dihasilkan nilai Accuracy sebesar 99,61%, Sensitivity 99,58%, Specitivity 99,66 %, PPV 99,79 %, NPV 99,33 %. terdapat grafik ROC dengan nilai AUC (Area Under Curve) sebesar 0,743 dimana hasilnya dapat dinyatakan sebagai kategori fair (adil). Terdapat satu data yang awalnya diprediksi sebagai kelas Layak, namun hasilnya menunjukkan kelas Tidak. Sebanyak 296 data diprediksi dengan benar sesuai dengan kelasnya, sedangkan tidak ada data yang diprediksi sebagai kelas Tidak, yang pada akhirnya menunjukkan kelas Layak. Pada hasil pengujian Cross Validation Dari pengujian K-Fold Teringgi yaitu Fold Ke-8 Dihasilkan nilai Accuracy sebesar 99,87%, Sensitivity 99,79%, Specitivity 100,00 %, PPV 100,00 %, NPV 99,66%. Pengujian dengan menggunakan 9 kali pengujian (fold 10), diperoleh nilai akurasi sebesar 99,87%. Sementara itu, rata-rata presisi mencapai 99,79%, dan rata-rata recall mencapai 100,00%. Tingkat akurasi dengan menggunakan Algoritma Naive Bayes mencapai 99,87%. Nilai AUC (Area Under Curve) sebesar 1.000 dimana hasilnya dapat dinyatakan sebagai *Excellent Classification* Hasil evaluasi Perbandingan dan validasi menunjukkan

bahwa algoritma Naive Bayes memiliki nilai akurasi sebesar 99,87% dan AUC sebesar 1,000, yang merupakan yang tertinggi di antara metode lainnya. Sementara itu, Algoritma C4.5 memiliki akurasi sebesar 99,61% dan AUC sebesar 0,743. Dari hasil algoritma Naive Bayes dalam data mining ini, tingkat AUC mengindikasikan klasifikasi yang sangat baik (Excellent classification). Sebaliknya, Algoritma C4.5 memiliki tingkat AUC yang diagnostik sebagai Fair (Adil). Dengan demikian, dapat disimpulkan bahwa algoritma Naive Bayes merupakan metode yang cukup baik dalam memprediksi kelayakan warga untuk menerima bantuan dari Program Keluarga Harapan (PKH).

DAFTAR PUSTAKA

- [1] A. Damuri, U. Riyanto, dan ..., "Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako," *JURIKOM (Jurnal ...)*, 2021, [Daring]. Tersedia pada: <http://www.ejurnal.stmik-budidarma.ac.id/index.php/jurikom/article/view/3655>
- [2] D. A. Virgoreta dan R. Nur Pratiwi, "Implementasi Program Keluarga Harapan (Pkh) Dalam Upaya Meningkatkan Kesejahteraan Masyarakat (Studi Pada Desa Beji Kecamatan Jenu, Kabupaten Tuban)," *Jap*, vol. 2, hal. 1–6, 2018.
- [3] A. S. Nurchotimah, Maftuh, Bunyamin, Malihah, Elly, dan Y. Harmawati, "Peran Pendamping Sosial Dalam Penanganan Konflik Keluarga Penerima Manfaat Program Keluarga Harapan," *J. Sociol. Pendidik. Humanis*, vol. 5, no. 1, hal. 61, 2020.
- [4] A. W. S. Sriwibowo, "PENERAPAN ALGORITMA NAÏVE BAYES DALAM PREDIKSI PENERIMAAN BANTUAN PROGRAM KELUARGA HARAPAN (PKH) Studi Kasus: Ds. Selur Kec ...," eprints.umpo.ac.id, 2020. [Daring]. Tersedia pada: <http://eprints.umpo.ac.id/6018/>
- [5] E. Fitriani, "Perbandingan Algoritma C4. 5 Dan Naïve Bayes Untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan," *Sist. J. Sist. Inf.*, 2020, [Daring]. Tersedia pada: <http://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/596>
- [6] T. Novika, P. Poningsih, H. Okprana, A. P. Windarto, dan H. Siahaan, "Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Pelajaran Matematika," *J. Media Inform. Budidarma*, vol. 5, no. 1, hal. 9, 2021, doi: 10.30865/mib.v5i1.2498.
- [7] A. Surahman, U. Hayati, B. Sosial, dan R. Miner, "Implementasi Algoritma Naïve Bayes Untuk Prediksi Penerima Bantuan Sosial," vol. 7, no. 1, hal. 347–352, 2023.
- [8] M. A. Tanjung, P. P. dan H. Qurniawan, "Analisa Kelayakan Penerima Program Keluarga Harapan (PKH) Menggunakan Algoritma C4.5," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.)*, vol. 6, no. 1, hal. 217, 2021, doi: 10.30645/jurasik.v6i1.286.
- [9] A. Sofianto, "Implementasi Program Keluarga Harapan (PKH) di Provinsi Jawa Tengah," *Sosio Konsepsia*, vol. 10, no. 1, 2020, doi: 10.33007/ska.v10i1.2091.
- [10] Elisyah, "Klasifikasi penerima bantuan program keluarga harapan (pkh) menggunakan algoritma c4.5 (studi kasus: kecamatan kupang timur)," 2022.
- [11] N. I. Nella, N. Y. Setiawan, dan D. E. Ratnawati, "Klasifikasi Penerima Bantuan Program Keluarga Harapan menggunakan Algoritme Decision Tree C4. 5 (Studi Kasus: Desa Mlirip Kabupaten Mojokerto)," ... *Tek. Inf. dan Ilmu* ..., 2022, [Daring]. Tersedia pada: <https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/10801>
- [12] Y. A. Singgalen, "Analisis Sentimen Konsumen terhadap Food, Services, and Value di Restoran dan Rumah Makan Populer Kota Makassar Berdasarkan Rekomendasi Tripadvisor Menggunakan Metode CRISP-DM dan SERVQUAL," *Build. Informatics, Technol. Sci.*, vol. 4, no. 4, hal. 1899–1914, 2023, doi: 10.47065/bits.v4i4.3231.
- [13] J. P. Gultom dan A. Rikki, "Implementasi Data Mining menggunakan Algoritma C-45 pada Data Masyarakat Kecamatan Garoga untuk Menentukan Pola perimaan Beras Raskin," vol. 02, no. 01, hal. 11–19, 2020.