E-ISSN : 2540-9719
SISTEMASI : Jurnal Sistem Informasi, Volume 9 No 1, Januari 2020 : 103 – 115 ISSN : 2302-8149

PERBANDINGAN ALGORITMA *C4.5* DAN *NAÏVE BAYES* UNTUK MENENTUKAN KELAYAKAN PENERIMA BANTUAN PROGRAM KELUARGA HARAPAN

Eka Fitriani
Program Studi Sistem Informasi, Fakultas Teknologi Informasi,
Universitas Bina Sarana Informatika
Jl. Kamal Raya No.18, Cengkareng Barat, Kota Jakarta Barat
Email : eka.ean@bsi.ac.id

(Diterima: 16 September 2019, direvisi: 3 Desember 2019, disetujui: 8 Desember 2019)

ABSTRACT

PKH is a program that provides conditional cash transfer to very poor households where they have been determined as PKH participants with certain conditions. Sometimes, this assistance is not in its target. Therefore, it is becoming a problem in this Government empowerment program. This problem arises due to ineffective data verification in the selection of citizens who are eligible to receive PKH assistance. Therefore, it is necessary to analyze PKH for determining the feasibility of a PKH Program problem. Through the results of the PKH analysis, it could be seen whether or not the citizens were feasible to get assistance. Based on this problem, comparative data mining classification methods were used to find out which algorithm is good for predicting citizenship feasibility. There were two algorithms used, namely the C4.5 algorithm and Naïve Bayes. After testing with the two algorithms using RapidMiner tools, the results showed that C4.5 algorithm produced an accuracy value of 91.25% and AUC value of 0.930 with a diagnosis level of Excellent Classification. While the Naïve Bayes algorithm produced an accuracy value of 87.11% and AUC value amounted to 0.923 with a diagnosis level of Excellent Classification. In conclusion, C4.5 algorithm is a good algorithm to be applied to the feasibility of PKH.

Keywords: *PKH Feasibility, Data Mining, C4.5, Naïve Bayes*

ABSTRAK

Program Keluarga Harapan (PKH) adalah program yang memberikan bantuan tunai bersyarat kepada Rumah Tangga Sangat Miskin yang telah ditetapkan sebagai peserta PKH dengan ketentuan tertentu. Permasalahan yang sering terjadi pada program pemberdayaan Pemerintah ini, salah satunya pada bantuan PKH adanya tidak tepat sasaran warga yang menerima bantuan PKH. Munculnya masalah tersebut, diakibatkan verifikasi data yang belum efektif terhadap pemilihan warga yang layak menerima bantuan PKH. Oleh karena itu perlu dilakukan analisis mengenai PKH sehingga dapat mengetahui kelayakan dari suatu permasalahan Program PKH. Melalui hasil analisis PKH, dapat diketahui apakah warga layak atau tidak dari permasalahan yang ada digunakan komparasi metode klasifikasi data mining untuk mengetahui algoritma mana yang baik untuk memprediksi kelayakan warga yaitu dengan menggunakan dua algoritma yaitu algoritma C4.5 dan Naïve Bayes. Setelah dilakukan pengujian dengan dua algoritma tersebut menggunakan tools *RapidMiner* didapatkan hasil yaitu algoritma C4.5 menghasilkan nilai akurasi sebesar 91,25% dan nilai AUC sebesar 0,930 dengan tingkat diagnosa *Excellent Classification* sedangkan algoritma *Naïve Bayes* menghasilkan nilai akurasi sebesar 87,11% dan nilai AUC sebesar 0,923 dengan tingkat diagnosa *Excellent Classification*. Sehingga didapat kesimpulan algoritma C4.5 merupakan algoritma yang baik untuk diterapkan pada kelayakan PKH.

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

1. PENDAHULUAN

Program pemberdayaan di Indonesia saat ini belum dapat mencerdaskan masyarakat untuk keluar dari kemiskinan karena program yang bersifat bantuan masih menjadi prioritas utama pemerintah. Program yang dijalankan seharusnya bersifat memberdayakan, sehingga dapat menciptakan masyarakat yang cerdas dalam menyelesaikan masalahnya sendiri, khususnya masalah kemiskinan. Masalah kemiskinan merupakan salah

satu persoalan mendasar yang menjadi pusat perhatian pemerintah di negara manapun[1]. Kemiskinan merupakan salah satu masalah yang dialami oleh beberapa Negara berkembang, termasuk indonesia. Banyak cara yang dilakukan untuk menanggulangi kemiskinan, diantaranya dengan program bantuan sosial untuk rakyat miskin [2]. Menyadari pentingnya permasalahan tersebut, pemerintah melakukan segala upaya untuk menanggulangi permasalahan yang terjadi akibat kemiskinan. Upaya yang dilakukan oleh pemerintah adalah mengeluarkan suatu kebijakan yang berkaitan dengan pemberdayaan keluarga miskin. Salah satu kebijakan pemerintah dalam hal ini diwujudkan melalui Program Keluarga Harapan (PKH).

Program Keluarga Harapan (PKH) adalah program yang memberikan bantuan tunai bersyarat kepada Rumah Tangga Sangat Miskin (RTSM/KSM) yang telah ditetapkan sebagai peserta PKH dengan ketentuan tertentu. Dengan program ini menjadi suatu program yang bagus untuk menanggulangi garis kemiskinan di Indonesia khususnya di Kota Karawang. Faktanya Program Keluarga Harapan (PKH) ini belum tepat sasaran dan belum tepat jumlah yang menerima bantuan PKH. Hal ini menyebabkan adanya kecemburuan sosial antar warga pada lingkungan masyarakat tersebut. Oleh karena itu, dibutuhkan penelitian untuk menentukan kelayakan masyarakat dalam menerima bantuan Program Keluarga Harapan (PKH). Untuk itu perlu dilakukan penelitian terkait kelayakan penerima bantuan Program Keluarga Harapan (PKH) yang telah dilakukan dengan beberapa metode *data mining. Data Mining* adalah proses menemukan korelasi baru yang bermakna, pola dan tren dengan memilah-milah sejumlah besar data yang tersimpan dalam repositori, menggunakan teknologi penalaran pola serta teknik-teknik statistik dan matematika [3].

Ada beberapa penelitian dan teknik analisa kelayakan PKH yang dibuat oleh beberapa peneliti seperti [4] menganalisa penerima bantuan program keluarga harapan (PKH) dengan metode AHP dan *Promethee*. Penelitian ini menentuan kelayakan PKH masih bersifat manual serta menggunakan data pada beberapa tahun yang lalu. Hal ini dikhawatirkan menimbulkan suatu kerancuan dan ketidaktepatan dalam menilai sehingga PKH tidak sampai kepada masyarakat kurang mampu yang benar-benar membutuhkan. Menganalisa perbandingan akurasi klasifikasi tingkat kemiskinan antara algoritma C4.5 dan *naïve bayes*. Pada penelitian ini menjelaskan perbandingan algotitma C4.5 dan *naïve bayes* untuk memprediksi tingkat kemiskinan[2].

Dalam penelitian ini, dilakukan analisis komparasi algoritma klasifikasi *data mining* yaitu C4.5 dan *naïve bayes*. Penulis mengusulkan untuk membuat sistem yang bisa membantu dalam klasifikasi Layak dan Tidak masyarakat dalam menerima bantuan PKH dengan berdasarkan kriteria kemiskinan yang ditentukan oleh pemerintah. Penelitian ini menggunakan komparasi algoritma C4.5 dan *naïve bayes* dalam membantu mengolah data.

2. TINJAUAN PUSTAKA

Dalam penulisan ini, digunakan berbagai referensi terkait dengan penelitian yang dilakukan. Sumber referensi tersebut terdiri dari buku, jurnal nasional dan internasional untuk menjelaskan *data mining*, algoritma klasifikasi terkait dengan penelitian yang dilakukan serta referensi juga berasal dari *Internet* mengenai Program Keluarga Harapan (PKH).

Ada beberapa penelitian dan teknik analisa kelayakan PKH yang dibuat oleh beberapa peneliti seperti:

a. Sistem Pendukung Keputusan Seleksi Calon Penerima Bantuan Program Keluarga Harapan (PKH) Dengan Metode AHP Dan Promethee [4]. Penelitian ini kombinasi dua metode sebagai sistem penunjang keputusan dalam seleksi kelayakan calon penerima Bantuan Program Keluarga Harapan dengan menggunakan AHP dan Promethee. Kriteria yang digunakan pada penelitian ini terdiri atas 4 kriteria (pekerjaan, pendapatan, tanggungan dan keadaan rumah) dan beberapa subkriteria yang telah di tentukan dari masing-masing kriteria pada metode yang digunakan serta dilakukannya pembobotan kriteria.
b. Perbandingan akurasi klasifikasi tingkat kemiskinan antara algoritma C4.5 dan Naïve Bayes Clasifier[2] membuat penelitian untuk membandingkan algoritma C4.5 dan *naïve bayes*. Model algoritma tersebut digunakan untuk menganalisis tingkat kemiskinan.
c. Analisa Komparasi Algoritma Naive Bayes Dan C4.5 Untuk Prediksi Penyakit Liver [5] penelitian ini dilakukan pembuatan model menggunakan *Naïve Bayes* dan C4.5 menggunakan data Pasien Penderita Liver. Data diperoleh dari UCI yang terdiri dari 583 dengan 11 bidang. Model yang dihasilkan, dikomparasikan untuk mengetahui algoritma yang paling baik dlaam penentuan identifikasi penyakit

liver. Untuk mengukur kinerja kedua algoritma tersebut digunakan metode pengujian *Cross Validation* dan Split Percentace, dan pengukurannya dengan menggunakan *Confusion Matri.*

d. Kajian Komparasi Algoritma C4.5, Naïve Bayes Dan Neural Network Dalam Pemilihan Penerima Beasiswa (Studi Kasus Pada Sma Muhammadiyah 4 Jakarta) [6]. Dalam penelitian ini menggunakan data siswa sebagai objek pengujian oleh tiga algoritma klasifikasi. Dari hasil pengujian dengan mengukur kinerja ketiga algoritma tersebut menggunakan metode pengujian *cross validation*, *confusion matrix* dan kurva ROC.
e. Random Forests for Poverty Classification [7]. Penelitian ini menerapkan metode yang relatif baru dalam *data mining* untuk mengatasi masalah kemiskinan. Algoritma *random forest* diterapkan pada data sensus untuk meningkatkan akurasi klasifikasi tingkat kemiskinan.

2.1. Kemiskinan

Kemiskinan merupakan fenomena dan masalah sosial yang terus menerus dikaji. Kemiskinan menjadi perhatian pemerintah pusat dan pemerintahan daerah. Salah satu faktor penyebab ketertinggalan dan penghambat dalam pembangunan suatu bangsa adalah tingginya angka kemiskinan. Kemiskinan juga bisa dikatakan suatu keadaan terjadi dengan ketidakmampuan untuk memenuhi kebutuhan dasar seperti makanan, pakaian, rumah untuk berlindung, pendidikan, dan kesehatan. Kemiskinan dapat juga disebabkan oleh kelangkaan alat pelengkap kebutuhan dasar, ataupun sulitnya akses terhadap pendidikan dan pekerjaan[8].

2.2. Program Keluarga Harapan (PKH)

Program Keluarga Harapan adalah program yang memberikan bantuan tunai kepada Rumah Tangga Sangat Miskin (RTSM). Sebagai imbalannya RTSM diwajibkan memenuhi persyaratan yang terkait dengan upaya peningkatan kualitas sumberdaya manusia (SDM), yaitu pendidikan dan kesehatan. Sebenarnya, PKH sendiri memiliki tujuan umum untuk meningkatkan aksesibilitas terhadap pelayanan pendidikan, kesehatan, dan kesejahteraan sosial dalam mendukung tercapainya kualitas hidup keluarga miskin. PKH diharapkan dapat mengurangi beban pengeluaran keluarga miskin dalam jangka pendek serta memutus rantai kemiskinan dalam jangka panjang. Sebab peningkatan kualitas kesehatan, pendidikan dan terpeliharanya tarap penghidupan masyarakat akan memberikan kesempatan pada masyarakat untuk mampu meningkatkan kualitas dirinya [9].

2.3. *Data Mining*

Data mining adalah proses menemukan korelasi baru yang bermakna, pola dan tren dengan memilah-milah sejumlah besar data yang tersimpan dalam repositori, menggunakan teknologi penalaran pola serta teknik-teknik statistik dan matematika [10].

2.4. Klasifikasi

Klasifikasi merupakan bagian dari prediksi, dimana nilai yang diprediksi berupa label. Klasifikasi menentukan class atau grup untuk tiap contoh data, input dari model klasifikasi adalah atribut dari contoh data (data sample) dan outputnya adalah *class* dari data samples itu sendiri, dalam *machine learning* untuk membangun model klasifikasi digunakan metode *supervised learning* [11].

Merupakan suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan.Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi datayang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlahaturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik inimenggunkan supervised induction yang memanfaatkan kumpulan pengujian dari data set yang terklasifikasi[12].

2.5. Algoritma C4.5

Algoritma C4.5 adalah bagian dari algoritma untuk klasifikasi dalam pembelajaran *machine learning* dan *data mining*. C4.5 merupakan algoritma yang cocok digunakan untuk masalah klasifikasi pada *machine learning* dan *data mining* [3].

Dalam klasifikasi pohon keputusan terdiri dari sebuah node yang membentuk akar. *Node* akar tidak memiliki input. *Node* lain yang bukan sebagai akar tetapi memiliki tepat satu input disebut *node* internal atau *test node*, sedangkan *node* lainnya dinamakan daun. Daun mewakili nilai target yang paling tepat dari salah satu *class* [13].

2.6. Algoritma *Naïve Bayes*

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 [14].

Naive bayes adalah tehnik yang diterapkan untuk menentuan kelas dari tiap masalah, yang sudah dibagi berdasarkan tiap-tiap masalah. perhitungan numerik berdasarkan pada pendekatan grup [15].

2.7. Pengujian *K-Fold Cross Validation*

Cross Validation adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi [11].

2.8. Evaluasi dan Validasi Metode

Evaluasi yang di lakukan pada penelitian ini menggunakan model *Confusion Matrix* dan ROC *curve*.

a. *Confusion Matrix*

Confusion matrix adalah alat yang sangat berguna untuk menganalisa seberapa baik pengklasifikasi bias mengenali tuple dari class yang berbeda [16]. Evaluasi dengan mengunakan fungsi *confusion matrix* akan mengahsilkan nilai *accuracy*, *precision*, dan *recall. Confusion matrix* merupakan tabel matrix yang terdiri dari dua kelas, yaitu kelas yang dianggap sebagai positif dan kelas yang dianggap sebagai negatif [17].

b. Kurva ROC

Fungsi Kurva ROC adalah untuk menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *Confusion Matrix*, ROC adalah grafik dua dimensi dengan *false positive* sebagai garis horizontal dan *true positive* sebagai garis *vertical* [10].

ROC (*Receiver Operating Characteristic*) merupakan cara yang digunakan untuk menggambarkan akurasi diskriminasi dari suatu pengujian diagnosis untuk menentukan apakah seseorang menderita suatu penyakit tertentu atau tidak.[11]

Performance keakurasian AUC dapat diklasifikasikan menjadi lima kelompok yaitu:

1. Akurasi bernilai 0.90 – 1.00 = *Excellent classification*
2. Akurasi bernilai 0.80 – 0.90 = *Good classification*
3. Akurasi bernilai 0.70 – 0.80 = *Fair classification*
4. Akurasi bernilai 0.60 – 0.70 = *Poor classification*
5. Akurasi bernilai 0.50 – 0.60 = *Failure*

2.9. *RapidMiner*

RapidMiner adalah sebuah software untuk pengolahan data mining. *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik [18].

3. METODE PENELITIAN

Langkah-langkah yang digunakan pada penelitian ini dalam penentuan pengumpulan data sampai pengujian data warga untuk klasifikasi penerima program bantuan PKH yaitu :

1. Jenis Penelitian

Dalam penelitian ini melakukan penerapan komparasi algoritma klasifikasi *data mining* yaitu algoritma C4.5 dan *naïve bayes* dalam hal pengujian kedua metode akan dipilih salah satu metode yang paling baik tingkat akurasinya.

2. Metode Pemilihan Populasi dan Sampel

a. Populasi

Populasi dalam penelitian ini merupakan warga yang dinyatakan layak mendapatkan PKH dan warga yang tidak layak mendapatkan PKH tahun 2016-2017 yang berasal dari data internal di salah satu kecamatan yang ada di kota Karawang.

b. Sampel

Sampel dari penelitian ini adalah data warga yang layak mendapatkan PKH dan warga yang tidak mendapatkan PKH. Data tersebut bersifat intern yang belum dipublikasikan oleh pihak kecamatan dapat dilihat pada Tabel 1 jumlah sampel yang diambil:

Tabel 1 Sampel Dataset

Layak	Tidak Layak	Total Sampel

599	510	1.109

3. Metode Pengumpulan Data

Metode pengumpulan data dibagi menjadi dua sumber data yaitu data primer dan data sekunder. Data primer yaitu data yang dikumpulkan pertama kali, dan untuk melihat apa yang sesungguhnya terjadi melalui observasi, *interview*, kuesioner, dll.

Dalam penelitian ini metode pengumpulan data untuk mendapatkan sumber data yang digunakan adalah metode pengumpulan data sekunder. Data utama diperoleh dari data warga yang layak mendapatkan PKH dan warga yang tidak layak mendapatkan PKH sedangkan data pendukung lainnya didapat dari buku, jurnal dan publikasi lainnya.

4. Metode Analisis dan Pengujian Data

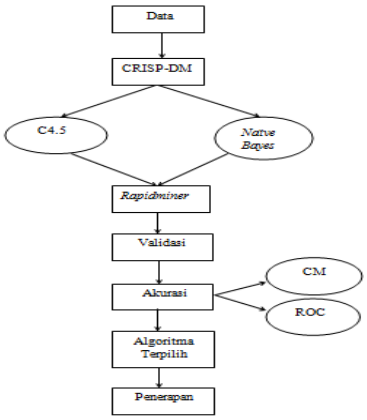
Teknik Analisis data menggunakan berupa matematika terhadap angka atau numerik dan nominal. Pada penelitian ini, analisis data dilakukan melalui data warga salah satu kecamatan yang ada di kota Karawang dengan nilai rata-rata warga yang mendapatkan PKH dan warga yang tidak layak mendapatkan PKH. Data diolah dan di uji dalam pengujian pada algoritma C4.5 dan *naïve bayes*. Kemudian pengujian *Rule* yang diperoleh C4.5 dan *naïve bayes* tersebut kemudian diuji dengan *confusion matrix* dan *kurva Receiver Operating Characteristic* (ROC) untuk mengukur tingkat akurasi yang akan dihasilkan dari metode tersebut.

Metode penelitian yang digunakan pada eksperimen ini menggunakan metodologi standar dalam penelitian *data mining* adalah model *Cross-Standard Industry for Data Mining* (CRISP-DM). Model CRISP-DM (Cross – Industry Standard Process for Data Mining) yang terdiri dari 6 tahap proses yaitu : *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment* [19].

a. *Business Understanding*
   Berdasarkan data masyarakat yang menerima bantuan Program Keluarga Harapan (PKH) menunjukan adanya tidak tepat sasaran penerima PKH tersebut. Pada penelitian ini dilakukan pengembangan dengan komparasi algoritma klasifikasi *C4.5* dan *naïve bayes* dengan tujuan untuk mengetahui kelayakan warga dalam penerima bantuan PKH dan meningkatkan akurasi dari perhitungan algoritma klasifikasi tersebut.
b. Data Understanding
   Pada tahap *Data Understanding*, dilakukan pengumpulan data, melakukan analisis penyelidikan data (data warga penerima PKH) untuk mengenali lebih lanjut data dan pencarian pengetahuan awal kemudian mengevaluasi kualitas dari data tersebut. Adapun sumber data utama yang digunakan dalam penelitian ini menggunakan data warga di salah satu kecamatan yang ada di kota karawang dengan 17 atribut tersebut. Data tersebut dianalisis misalnya jumlah data yang akan diambil, dan jumlah data dengan keterangan layak atau tidak layak.
c. Data Preparation
   Pada tahapan ini data sebanyak 1.109 data warga yang terdiri dari warga yang layak dan yang tidak layak, atribut terdapat 17 atribut, akan dilakukan beberapa penyeleksian untuk menghasilkan data yang dibutuhkan.
d. Modelling
   Dilakukan pemrosesan *data training* sehingga akan menghasilkan beberapa aturan dan akan membentuk sebuah pohon keputusan. Pada penelitian ini komparasi algoritma yang akan digunakan ada tiga yaitu algoritma klasifikasi C4.5 dan *naïve bayes*.
e. *Evaluation*
   melihat hasil akurasi pada proses klasifikasi komparasi algoritma C4.5 dan *naïve bayes* serta evaluasi dengan *confision matrix* dan ROC *curve*. Evaluasi bertujuan untuk menentukan nilai kegunaan dari model yang telah berhasil kita buat pada langkah sebelumnya. Penjelasan secara lengkap tentang membandingkan ketiga model tersebut terdapat pada bab empat.
f. Develoyment
   Berdasarkan penelitian yang dilakukan dengan penerapan model komparasi algoritma klasifikasi C4.5 dan *naïve bayes* untuk menentukan kelayakan penerima bantuan Program Keluarga Harapan (PKH).

5. Langkah-langkah Penelitian

Langkah-langkah penelitian dapat dilihat pada gambar sebagai berikut :



Gambar 1 Langkah-langkah penelitian

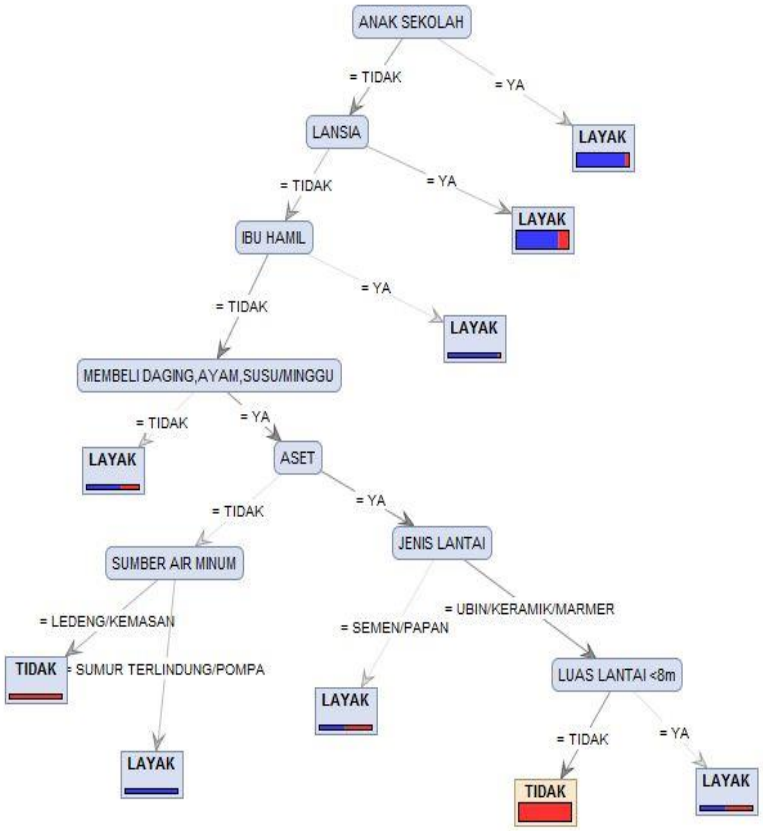
Pada langkah-langkah pemikiran diatas dapat dijelaskan sebagai berikut:

1. Menyiapkan data mentah untuk diolah.
2. Dengan menggunakan metode CRISP–DM data *traning* diuji dengan algoritma C4.5 dan *naïve bayes* mendapatkan validasi dari data tersebut.
3. Data *training* di uji dengan menggunakan validasi untuk mendapatkan nilai *confusion matrix* C4.5 dan *naïve bayes* dan nilai *ROC* C4.5 dan naïve *b*ayes, untuk mendapatkan akurasi yang terbaik.
4. Hasil dari perbandingan kedua algoritma tersebut diambil *confusion matrix* dan ROC yang terbesar.
5. Algoritma terpilih diterapkan pada pembuatan *Grapic User Interface (GUI)* untuk menguji keakuratan rule yang dihasilkan oleh algoritma terpilih.

4. HASIL DAN PEMBAHASAN

Hasil pengujian dari data warga yang diuji coba menggunakan *tools rapidminer* dengan menggunakan algoritma C4.5 dan Naïve Bayes dapat menghasilkan *K-Fold Cross Validation, accuracy,confusion matrix dan ROC.*

4.1. Hasil Eksperimen Algoritma C4.5

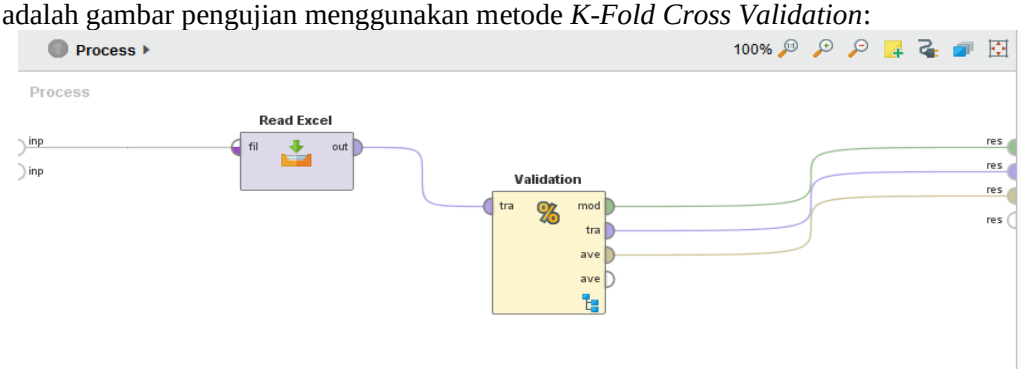


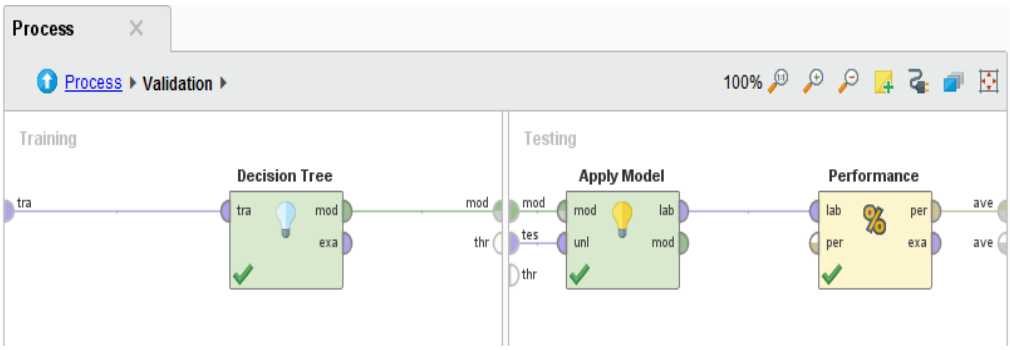
Gambar 2 Pohon keputusan klasifikasi penerima PKH algoritma C4.5

Berdasarkan pohon keputusan yang didapat sesuai gambar 2. didapatkan aturan atau rule, rule tersebut dimanfaatkan untuk mengambil keputusan pada data yang baru.

Gambar 3 merupakan pengujian model algoritma C4.5 menggunakan *software Rapid Minner*. *Read excel* yang pada gambar merupakan tools untuk mengambil data training yang akan dibuat model. Data kemudian dihubungkan dengan *Validation*. Di dalam proses *validation* kemudian ditambahkan tools untuk model menggunakan C4.5 dan *performance* untuk performansi dari klasifikasinya.

Berikut adalah gambar pengujian menggunakan metode *K-Fold Cross Validation*:





Gambar 3 Pengujian *K-Fold cross validation* algoritma C4.5

4.2. Evaluasi Cofussion Matrix

Hasil uji terbaik pada pengklasifikasian data warga penerima PKH menggunakan Algoritma C4.5 dapat di lihat pada tabel berikut :

Tabel 2 Konversi *confusion matrix* algoritma klasifikasi *C4.5*

accuracy: 91.25% +/- 2.38% (mikro: 91.25%)

	true LAYAK	true TIDAK	class precision
pred. LAYAK	593	91	86.70%
pred. TIDAK	6	419	98.59%
class recall	99.00%	82.16%	

Penjelasan dari Tabel 2. diketahui dari 1.109 data, 593 diklasifikasikan Layak sesuai dengan prediksi yang dilakukan dengan metode Algoritma C4.5, lalu 91 data diprediksi Layak tetapi ternyata hasilnya Tidak, 419 data *class* Tidak diprediksi sesuai, dan 6 data diprediksi Tidak ternyata Layak. Berdasarkan Tabel IV.5 tersebut menunjukan bahwa, tingkat akurasi dengan menggunakan algoritma C4.5 adalah sebesar 91,25%, dan dapat dihitung untuk mencari nilai *accuracy*, *sensitivity*, *specificity*, *ppv*, dan *npv* pada persamaan sebagai berikut:

$$acc = \frac{tp + tn}{tp + tn + fp + fn} \qquad acc = \frac{593 + 419}{593 + 419 + 6 + 91}$$

$$sensitivity = \frac{tp}{tp + fn} \qquad sensitivity = \frac{593}{593 + 91}$$

$$specitivity = \frac{tn}{tn + fp} \qquad specitivity = \frac{419}{419 + 6}$$

$$ppv = \frac{tp}{tp + fp} \qquad ppv = \frac{593}{593 + 6}$$

$$npv = \frac{tn}{tn + fn} \qquad npv = \frac{419}{419 + 91}$$

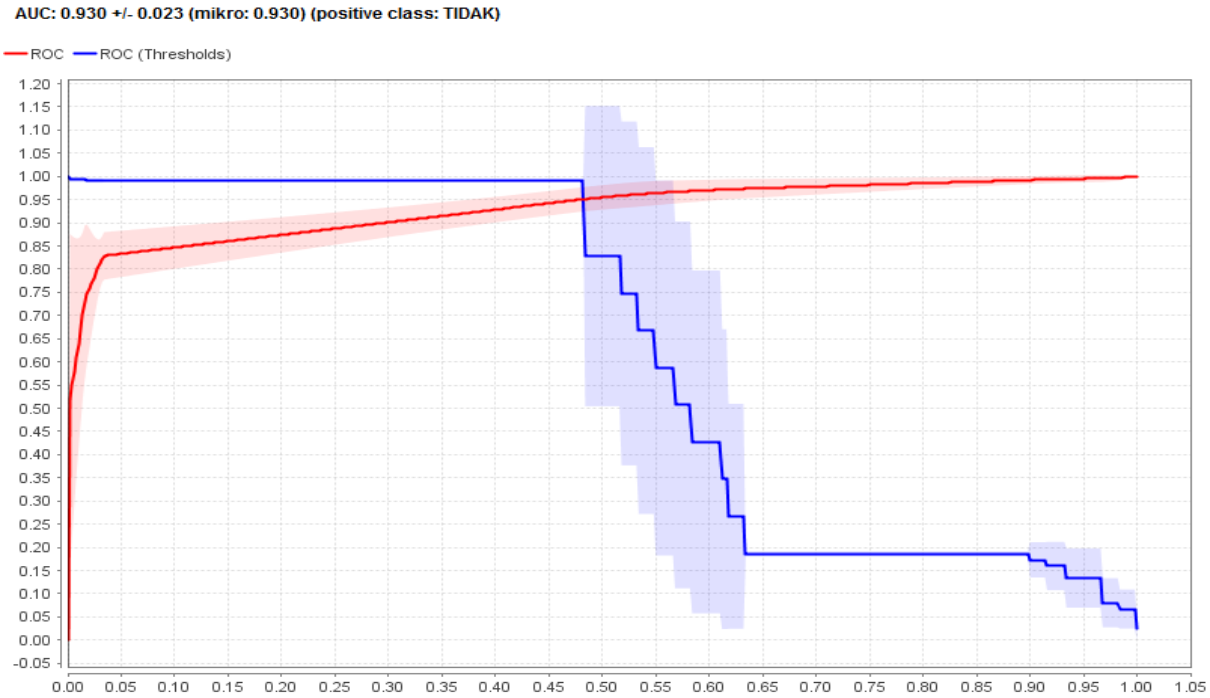
Tabel 3 Hasil Perhitungan Algoritma C4.5

	Nilai (%)
Accuracy	91,25
Sensitivity	86,69
Specitivity	98,58
PPV	98,99
NPV	82,15

Berdasarkan Tabel 3 menunjukan bahwa, tingkat akurasi menggunakan algoritma klasifikasi C4.5 adalah sebesar **91,25%.**

4.3. Evaluasi dengan ROC algoritma C4.5

Hasil perhitungan yang divisualisasikan dengan kurva ROC untuk algoritma C4.5 dapat di lihat pada Gambar 3. yang mengekspresikan *confusion matrix* dari Tabel 2. Garis horizontal adalah *false positive* dan garis vertikal *true positive*.



Gambar 4 Nilai AUC dalam grafik ROC algoritma C4.5

Pada Gambar 4 terdapat grafik ROC dengan nilai AUC (*Area Under Curve)* sebesar 0.930 dimana hasilnya dapat dinyatakan sebagai *Excellent Classification.* karena *Performance* keakurasian AUC 0.930 ada pada rentang 0.900 – 1.000 termasuk *performace Excellent Clasification.*

4.4. Hasil Eksperimen Algoritma *Naïve Bayes*

Dalam membuat model *Naive Bayes* terlebih dahulu kita mencari *probabilitas* hipotesis untuk masing-masing Kelas P(H). Hipotesis yang ada yaitu warga yang layak menerima bantuan PKH dan warga yang tidak layak menerima bantuan PKH. Data *tranning* yang digunakan sama seperti pengujian algortima C4.5, dengan total data yaitu 1.109 dengan 599 data warga yang layak PKH dan 510 data warga yang tidak layak PKH.

Perhitungan *probabilitas prior* dilakukan dalam bentuk persamaan dibawah ini:

P (Layak) =599:1109 = 0,54012624

P (Tidak Layak) =510:1109 = 0,45987376

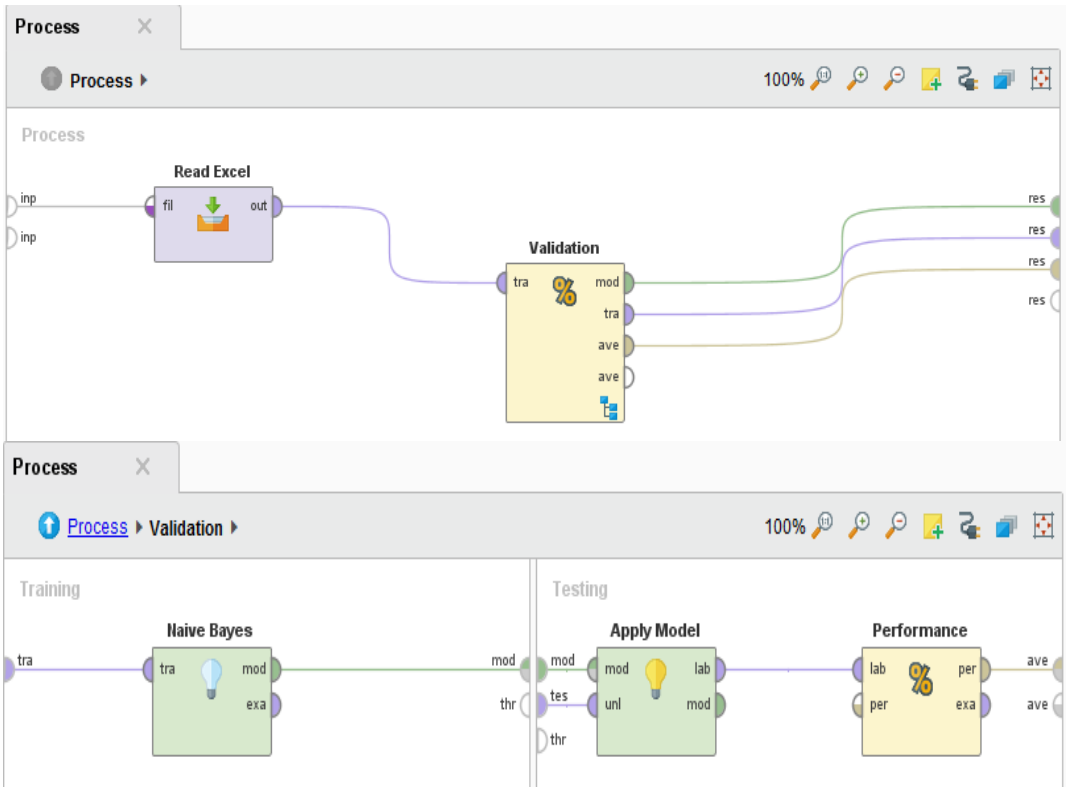
Setelah *probabilitas* untuk tiap hipotesis diketahui, langkah selanjutnya adalah menghitung *probabilitas* kondisi tertentu (*probabilitas* X) berdasarkan *probabilitas* tiap hipotesis (*probabilitas* H) atau dinamakan *probabilitas prior.* Selanjutnya untuk mengetahui hasil perhitungan dari *probabilitas prior*, maka dilakukan penghitungan dengan cara merinci jumlah kasus dari tiap-tiap atribut variabel data, adapun hasil perhitungan *probabilitas prior* dengan menggunakan algoritma *naïve bayes* dapat dilihat pada Tabel 4 berikut:

Tabel 4 Perhitungan *Probabilitas Prior*

Atribut		Jumlah Kasus (S)	LAYAK (S1)	TIDAK (S2)	P(X|Ci) LAYAK	P(X|Ci) TIDAK
TOTAL		1109	599	510	0.54013	0.45987
LANSIA						
	YA	419	344	75	0.5742905	0.1470588
	TIDAK	690	255	435	0.4257095	0.8529412
SUM		1109				
PENDIDIKAN TERAKHIR KK						
	TIDAK/BELUM SEKOLAH	5	4	1	0.0066778	0.0019608
	TIDAK TAMAT SD/MI	33	21	12	0.0350584	0.0235294
	TAMAT SD/MI	351	255	96	0.4257095	0.1882353
	TAMAT SLTP/MTSN	193	137	56	0.2287145	0.1098039
	TAMAT SLTA/MA	504	168	336	0.2804674	0.6588235
	TAMAT PT/AKADEMIK	16	12	4	0.0200334	0.0078431
	MASIH SLTP/MTSN	3	1	2	0.0016694	0.0039216
	MASIH SLTA/MA	4	1	3	0.0016694	0.0058824
SUM		1109				
ANAK SEKOLAH						
	YA	251	237	14	0.3956594	0.0274510
	TIDAK	858	362	496	0.6043406	0.9725490
SUM		1109				
PEKERJAAN						
	LAINNYA	27	19	8	0.0317195	0.0156863
	PEDAGANG	19	6	13	0.0100167	0.0254902
	PEGAWAI SWASTA	318	140	178	0.2337229	0.3490196
	PEKERJA LEPAS	101	63	38	0.1051753	0.0745098
	PENSIUNAN	79	45	34	0.0751252	0.0666667
	PETANI	33	22	11	0.0367279	0.0215686
	PNS/TNI/POLRI	3	1	2	0.0016694	0.0039216
	TIDAK/BELUM BEKERJA	202	185	17	0.3088481	0.0333333
	WIRASWASTA	327	118	209	0.1969950	0.4098039
SUM		1109				
IBU HAMIL						
	YA	58	56	2	0.0934891	0.0039216
	TIDAK	1051	543	508	0.9065109	0.9960784
SUM		1109				
MEMBELI PAKAIAN/PERTAHUN						
	YA	1105	597	508	0.9966611	0.9960784
	TIDAK	4	2	2	0.0033389	0.0039216
SUM		1109				
FREKUENSI MAKAN/HARI						
	YA	1103	595	508	0.9933222	0.9960784
	TIDAK	6	4	2	0.0066778	0.0039216
SUM		1109				
BEROBAT KE PUSKESMAS						
	YA	1103	595	508	0.9933222	0.9960784
	TIDAK	6	4	2	0.0066778	0.0039216
SUM		1109				
MEMBELI DAGING,AYAM,SUSU/MINGGU						
	YA	1104	596	508	0.9949917	0.9960784
	TIDAK	5	3	2	0.0050083	0.0039216
SUM		1109				
ASET						
	YA	1086	585	501	0.9766277	0.9823529
	TIDAK	23	14	9	0.0233723	0.0176471
SUM		1109				
JENIS DINDING						
	BAMBU	5	5	0	0.0083472	0
	TEMBOK	1103	594	509	0.9916528	0.9980392
	KAYU/SENG	1	0	1	0	0.0019608
SUM		1109				
JENIS LANTAI						
	LAINNYA	1	1	0	0.0016694	0
	SEMEN/PAPAN	5	4	1	0.0066778	0.0019608
	TANAH	1	1	0	0.0016694	0.0000000
	UBIN/KERAMIK/MARMER	1102	593	509	0.9899833	0.9980392
SUM		1109				
SUMBER PENERANGAN						
	LISTRIK	1108	598	510	0.9983306	1
	LAINNYA	1	1	0	0.0016694	0
SUM		1109				
SUMBER AIR MINUM						
	LAINNYA	1	1	0	0.0016694	0
	LEDENG/KEMASAN	1062	570	492	0.9515860	0.9647059
	SUMUR TERLINDUNG/POMPA	46	28	18	0.0467446	0.0352941
SUM		1109				
JENIS BAHAN BAKAR MEMASAK						
	LISTRIK/GAS	1103	595	508	0.9933222	0.9960784
	ARANG/KAYU	4	2	2	0.0033389	0.0039216
	MINYAK TANAH	1	1	0	0.0016694	0
	LAINNYA	1	1	0	0.0016694	0
SUM		1109				
FASILITAS BAB						
	JAMBAN SENDIRI	1103	595	508	0.9933222	0.9960784
	JAMBAN BERSAMA	4	2	2	0.0033389	0.0039216
	JAMBAN UMUM	1	1	0	0.0016694	0
	LAINNYA	1	1	0	0.0016694	0
SUM		1109				
LUAS LANTAI						
	YA	2	1	1	0.0016694	0.0019608
	TIDAK	1107	598	509	0.9983306	0.9980392
SUM		1109				

Perhitungan probabilitas prior dapat dibuatkan model *K-Fold Cross Validation* yaitu seperti terlihat pada Gambar 5.

E-ISSN : 2540-9719
ISSN : 2302-8149



Gambar 5 Pengujian *K-Fold Cross validation* algoritma *naïve bayes*

Gambar 5. merupakan pengujian model algoritma *naïve bayes* menggunakan *software rapidminner*. *Read excel* yang pada gambar merupakan *tools* untuk mengambil data training yang akan dibuat model. Data kemudian dihubungkan dengan *validation*. Di dalam proses *validation* kemudian ditambahkan *tools* untuk model menggunakan *naïve bayes* dan *performance* untuk performansi dari klasifikasinya.

4.5. Evaluasi Cofussion Matrix

Model *confusion matrix* yang kedua dengan menggunakan algoritma klasifikasi *naïve bayes*, kemudian masukan data *testing* yang sudah disiapkan kedalam *confusion matrix* sehingga didapatkan hasil pada Tabel 5 sebagai berikut:

Tabel 5 Konversi *Confusion matrix* Algoritma Klasifikasi *Naïve Bayes*

accuracy: 87.11% +/- 3.02% (mikro: 87.11%)

	true LAYAK	true TIDAK	class precision
pred. LAYAK	587	131	81.75%
pred. TIDAK	12	379	96.93%
class recall	98.00%	74.31%	

Penjelasan pada tabel tersebut, diketahui dari 1.109 data, 587 diklasifikasikan Layak sesuai dengan prediksi yang dilakukan dengan Algoritma *naïve bayes*, 131 data diprediksi Layak tetapi ternyata hasilnya Tidak, 379 data *class* Tidak diprediksi sesuai, dan 12 data diprediksi Tidak ternyata Layak. Berdasarkan Tabel 5 tersebut menunjukan bahwa, tingkat akurasi dengan menggunakan algoritma *naïve bayes* adalah sebesar 87,11%, dan dapat dihitung untuk mencari nilai *accuracy*, *sensitivity*, *specificity*, *ppv*, dan *npv* pada persamaan sebagai berikut:

$$\text{acc} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}} \qquad \text{acc} = \frac{587 + 379}{587 + 379 + 12 + 131}$$

$$\text{sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}} \qquad \text{sensitivity} = \frac{587}{587 + 131}$$

$$\text{specitivity} = \frac{\text{tn}}{\text{tn} + \text{fp}} \qquad \text{specitivity} = \frac{379}{379 + 12}$$

$$\text{ppv} = \frac{\text{tp}}{\text{tp} + \text{fp}} \qquad \text{ppv} = \frac{587}{587 + 12}$$

$$\text{npv} = \frac{\text{tn}}{\text{tn} + \text{fn}} \qquad \text{npv} = \frac{379}{379 + 131}$$

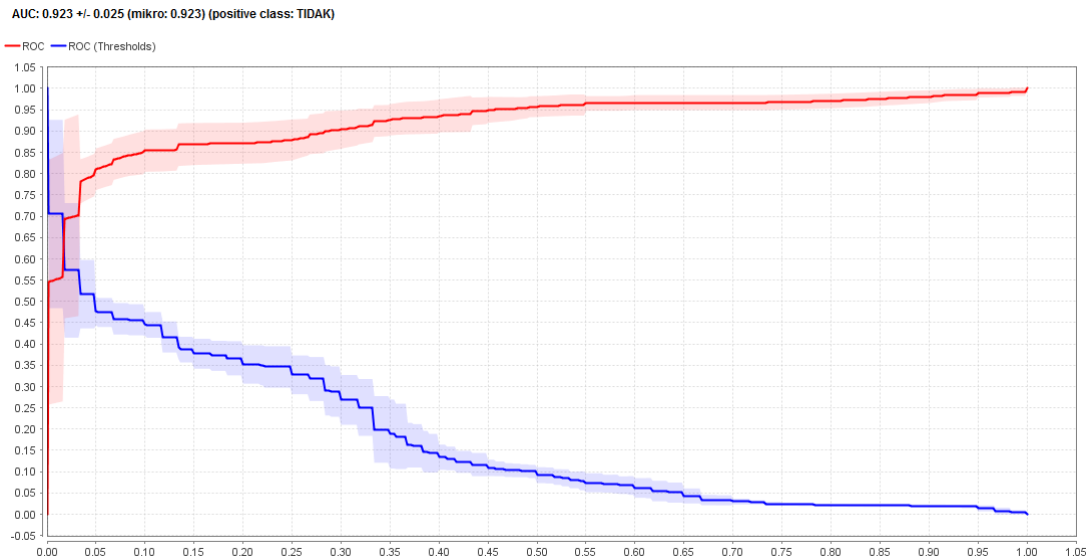
Tabel 6 Hasil Perhitungan Algoritma *Naive Bayes*

	Nilai (%)
Accuracy	87,11
Sensitivity	81,75
Specitivity	96,93
PPV	97,99
NPV	74,31

Berdasarkan Tabel 6 menunjukan bahwa, tingkat akurasi menggunakan algoritma klasifikasi *naïve bayes* adalah sebesar **87,11%.**

4.6. Evaluasi dengan *ROC Naïve Bayes*

Hasil perhitungan yang divisualisasikan dengan kurva ROC untuk algoritma *naïve bayes* dapat di lihat pada Gambar 5 yang mengekspresikan *confusion matrix* dari Tabel 5 Garis horizontal adalah *false positive* dan garis vertikal *true positive*.



Gambar 6 Nilai AUC dalam grafik ROC algoritma *Naïve Bayes*

Dari Gambar 6. terdapat grafik ROC dengan nilai AUC (*Area Under Curve)* sebesar 0.923 dimana hasilnya dapat dinyatakan sebagai *Excellent Classification* karena *Performance* keakurasian AUC 0.930 ada pada rentang 0.900 – 1.000 termasuk *performace Excellent Clasification.*

4.7. Analisis Evaluasi Hasil dan Validasi Model

Dari hasil pengujian diatas komparasi dari kedua algoritma klasifikasi *data mining* yaitu algoritma C4.5 dan *naïve bayes*, pengukuran akurasi menggunakan confusion matrix dan kurva ROC terbukti bahwa hasil pengujian algoritma C4.5 memiliki nilai akurasi yang lebih tinggi dibandingkan dengan algoritma Naïve Bayes. Algoritma C45 menghasilkan akurasi sebesar 91,25% sedangkan nilai akurasi untuk hasil pengujian model algoritma Naïve Bayes sebesar 87,11%.

Perbandingan hasil pengujian terbaik dapat di lihat pada Tabel 7:

Tabel 7 Komparasi Nilai Akurasi

	C4.5	NB
Accuracy	91,25%	87,11%
Precision	98,66%	97,11%
Recall	82,17%	74,33%

5. KESIMPULAN

Dalam penelitian ini dilakukan pengujian model dengan membandingkan dua metode data mining yaitu algoritma C4.5 dan *naïve bayes* dengan menggunakan data warga di salah satu kecamatan yang ada di Kota Karawang yang terdiri dari data warga yang layak menerima PKH dan warga yang tidak layak menerima PKH dengan total data sebanyak 1.109 data warga. Data diuji menggunakan *tools RapidMiner* kemudian model yang diuji akan menghasilkan nilai *accuracy*, *precision*, *recall* dan AUC dari setiap algoritma. Hasil evaluasi dan validasi diketahui bahwa algoritma C4.5 memiliki nilai *accuracy* 91,25% dan AUC 0,930 paling tinggi diantara metode yang lainnya, sedangkan untuk metode *naïve bayes* memiliki *accuracy* 87,11 dan AUC 0,923. Dari kedua algoritma *data mining* tersebut, tingkat AUC diagnosa *Excellent classification.* Dengan demikian algoritma C4.5 merupakan metode yang cukup baik dalam memprediksi kelayakan warga dalam menerima bantuan Program Keluarga Harapan (PKH).

REFERENSI

[1] I. Kurniawan And R. A. Saputra, "Penerapan Algoritma C5 . 0 Pada Sistem Pendukung Keputusan Kelayakan Penerimaan Beras Masyarakat Miskin," *J. Inform.*, Vol. 4, No. 2, Pp. 236–240, 2017.
[2] D. Iskandar And Y. K. Suprapto, "Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C 4.5 Dan Naïve Bayes," *Ilm. Nero*, Vol. 2, No. 1, Pp. 37–43, 2015.
[3] N. Iriadi And N. Nuraeni, "Kajian Penerapan Metode Klasifikasi Data Kelayakan Kredit Pada Bank," *J. Tek. Komput. Amik Bsi*, Vol. Ii, No. 1, Pp. 132–137, 2016.
[4] S. Anjarwati, "Penerima Bantuan Program Keluarga Harapan (Pkh) Dengan Metode Ahp Dan Promethee (Studi Kasus Pada Kelurahan Kudaile Slawi)," Vol. 4, No. 1, 2017.
[5] E. Rahmawati, "Vol. Xii No. 2, September 2015 Jurnal Techno Nusa Mandiri," *Techno Nusa Mandiri*, Vol. Xii, No. 2, Pp. 21–26, 2015.
[6] U. Pauziah, "Kajian Komparasi Algoritma C4 . 5 , Naïve Bayes Dan Neural Network Dalam Pemilihan Penerima Beasiswa (Studi Kasus Pada Sma Muhammadiyah 4 Jakarta)," *Vol . 1 No . 1*, Vol. 1, No. 1, Pp. 2527 – 9661, 2016.
[7] R. Thoplan, "International Journal Of Sciences : Random Forests For Poverty Classification," No. August, 2014.
[8] S. A. Purwanto, Sumartono, And M. Makmur, "Implementasi Kebijakan Program Keluarga Harapan (Pkh) Dalam Memutus Rantai Kemiskinan (Kajian Di Kecamatan Mojosari Kabupaten Mojokerto)," *Wacana*, Vol. 16, No. 2, Pp. 79–96, 2013.
[9] D. A. N. P. Kemiskinan, "Program Keluarga Harapan (Pkh): Antara Perlindungan Sosial," 2016.
[10] N. Nuraeni, "Penentuan Kelayakan Kredit Dengan Algoritma Naïve Bayes Classifier : Studi Kasus Bank Mayapada Mitra Usaha Cabang Pgc," Vol. Iii, No. 1, Pp. 9–15, 2017.
[11] A. Puspita And M. Wahyudi, "Algoritma C4.5 Berbasis Decision Tree Untuk Prediksi Kelahiran Bayi Prematur," Pp. 97–102, 2015.
[12] D. Iskandar And Y. K. Suprapto, "Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C4 . 5 Dan Naïve Bayes Clasifier," *Java J. Electr. Electron. Eng.*, Vol. 11, No. 1, Pp. 14–17, 2013.
[13] A. Mukminin And D. Riana, "Komparasi Algoritma C4 . 5 , Naïve Bayes Dan Neural Network Untuk Klasifikasi Tanah," Vol. 4, No. 1, Pp. 21–31, 2017.

[14] T. Arifin, "Metode Data Mining Untuk Klasifikasi Data Sel Nukleus Dan Sel Radang Berdasarkan Analisa Tekstur," *Informatika*, Vol. Ii, No. 2, Pp. 425–433, 2015.
[15] S. Hanggara, T. M. Akhriza, And M. Husni, "Aplikasi Web Untuk Analisis Sentimen Pada Opini Produk Dengan Metode Naive Bayes Classifier," *Semin. Nas. Inov. Dan Apl. Teknol. Di Ind. 2017*, Pp. 1–6, 2017.
[16] L. A. Utami, "Melalui Komparasi Algoritma Support Vector Machine Dan K-Nearest Neighbor Berbasis Particle Swarm Optimization," Vol. 13, No. 1, Pp. 103–112, 2017.
[17] M. D. Tree, R. Forest, R. D. L. P, C. Fatichah, D. Purwitasari, And A. Twitter, "Deteksi Gempa Berdasarkan Data Twitter," Vol. 6, No. 1, Pp. 159–162, 2017.
[18] A. Purwanto And E. A. Darmadi, "Perbandingan Minat Siswa Smu Pada Metode Klasifikasi Menggunakan 5 Algoritma," *J. Ikraith-Informatika*, Vol. 2, No. 1, Pp. 43–47, 2018.
[19] R. D. Probo, B. Irawan, R. Rumani, M. 3, P. S1, And S. Komputer, "Analisis Dan Implementasi Perbandingan Algoritma Knn (K- Nearest Neighbor) Dengan Svm (Support Vector Machine) Untuk Prediksi Penawaran Produk Comparative Analysis And Implementation Of Knn (K-Nearest Neighbor) With Svm (Support Vector Machine) Algorithm ," Vol. 3, No. 3, Pp. 4988–4995, 2016.