JUSIFOR: Jurnal Sistem Informasi dan Informatika
Vol. 1, No. 1, Juni 2022, hal. 1 - 8
E-ISSN: 2830-2443 | P-ISSN: 2830-3393

Komparasi Metode Klasifikasi Algoritma C5.0 dan Naïve Bayes untuk Menentukan Jurusan Siswa

Muhammad Zainuri[1], Muhammad Hanif Fahmi[2], Raka Anugrah Hamdhana[3]

[1,2,3] Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Raden Rahmat, Indonesia

Info Artikel

Riwayat Artikel:
Diterima : **30-04-2022**
Direvisi : **02-05-2022**
Disetujui : **12-05-2022**

Kata Kunci:

Data Mining,
Algoritma C5.0,
Naïve Bayes,
K-fold Cross Validation,
Confusion Matrix,
Knowledge Discovery in Database

ABSTRAK

Penentuan jurusan siswa yang diimplementasikan di SMAN 1 Gondanglegi berdasarkan tes potensi akademik (TPA), tes *intelligence* (IQ) dan wawancara. Penjurusan tersebut dilakukan setelah siswa melakukan pendaftaran atau sebelum siswa diterima sebagai kelas X. Beberapa metode *data mining* dalam penentuan atau prediksi jurusan yang telah dilakukan oleh peneliti terdahulu diantaranya algoritma C4.5, C5,0 dan *naïve bayes* serta dalam perbandingan yaitu algoritma ID3 dan C5.0 kemudian perbandingan *naïve bayes* dan decision tree. Yang mana dalam perbandingan antara metode algoritma C5.0 dan *naïve bayes* belum dilakukan. Dari permasalahan tersebut maka peneliti bertujuan untuk melakukan analisa perbandingan *data mining* menggunakan klasifikasi *algoritma C5.0* dan *naïve bayes* dalam memprediksi jurusan siswa. Adapun metode *data mining* yang digunakan yaitu *knowledge discovery in database* (KDD). Berdasarkan hasil perbandingan pengujian yang telah dilakukan melalui berbagai skenario terhadap kedua metode tersebut, pengujian *10-fold cross validation* yang kemudian dicatat dalam *confusion matrix* menghasilkan nilai akurasi yaitu sebesar 60,87% untuk algoritma *C5.0* sedangkan untuk *naïve bayes* sebesar 56,52%. Dari hasil yang diperoleh algoritma *C5.0* merupakan metode paling baik dibanding *naïve bayes* yang dibuktikan dengan nilai tingkat akurasi yang didapatkan lebih tinggi.

***Keywords*:**

Data Mining,
C5.0 Algorithm,
Naïve Bayes,
K-fold Cross Validation,
Confusion Matrix,
Knowledge Discovery in Database

ABSTRACT

Determination of student majors implemented in SMAN 1 Gondanglegi based on academic potential tests (TPA), intelligence tests (IQ) and interviews. The assignment is done after the student registers or before the student is accepted as class X. Some data mining methods in determining or predicting majors that have been done by previous researchers include the algorithms C4.5, C5.0 and naïve bayes and in comparison, namely the ID3 and C5.0 algorithms then the comparison of naïve bayes and decision trees. Which in comparison between the C5.0 algorithm method and naïve bayes has not been done. From these problems, researchers aim to analyze comparative data mining using the classification of C5.0 algorithms and naïve bayes in predicting student majors. The data mining method used is knowledge discovery in database (KDD). Based on the results of comparisons of tests that have been conducted through various scenarios against the two methods, the 10-fold cross validation test which is then recorded in the confusion matrix produces an accuracy value of 60.87% for the C5.0 algorithm while for naïve bayes at 56.52%. From the results obtained by the C5.0 algorithm is the best method compared to naïve bayes as evidenced by the value of the accuracy level that is higher.

Penulis Korespondensi:
M. Hanif Fahmi,
Program Studi Sistem Informasi,
Universitas Islam Raden Rahmat Malang
Email: hanif@uniramalang.ac.id

1. PENDAHULUAN

SMAN 1 Gondanglegi adalah lembaga yang merupakan satu kesatuan dari Sekolah Menengah Atas (SMA) yang termasuk salah satu bentuk pendidikan formal. Sekolah menengah atas negeri (SMAN) 1 Gondanglegi berdiri sejak 22 Desember 1986 dengan terdapat tiga jurusan yaitu: IPA, IPS dan Bahasa. Kurikulum di SMAN I Gondanglegi yang digunakan adalah kurikulum 2013, di mana penjurusannya dimulai sejak kelas X (Sepuluh). Dalam mengelola sistem penjurusan ada beberapa pihak yang terlibat yaitu: waka kurikulum, tim guru yang bertugas serta guru bimbingan konseling. Sistem penjurusan yang diterapkan di SMAN 1 Gondanglegi adalah tes potensi akademik (TPA), tes *intelligence* (IQ) dan wawancara. Untuk penentuan jurusan pada SMA tersebut yang lebih dominan dilihat adalah dari nilai rata-rata tes potensi akademik siswa, yang mana pada saat wawancara siswa akan diberi rekomendasi oleh guru berdasarkan hasil dari nilai rata-rata tes potensi akademik. Dalam proses penjurusan yang telah dilakukan oleh SMAN 1 Gondanglegi terdapat kasus siswa yang meminta untuk pindah jurusan meskipun di awal penjurusan sudah diberitahu batas waktu tiga bulan tetapi ada juga siswa yang meminta pindah jurusan lebih dari batas waktu yang telah diberikan pihak sekolah. Hal tersebut bisa dibilang bahwa sistem penjurusan yang digunakan di SMA tersebut kurang efektif dan kurang efisien, yang mana dalam pemilihan jurusan terdapat masalah baik dari siswa maupun guru.

Untuk mengatasi permasalahan tersebut, dapat menggunakan *data mining* dengan model klasifikasi untuk menguji data siswa. *Data mining* adalah proses menganalisa data dari sudut pandang yang berbeda dan mengubahnya menjadi informasi-informasi penting yang dapat digunakan untuk meningkatkan keuntungan, mengurangi biaya atau dari keduanya. *Data mining* merupakan aktivitas yang melibatkan pengumpulan data dan penggunaan data historis untuk menemukan keteraturan, pola atau hubungan dalam data besar kemudian mengekstrak data tersebut menjadi informasi-informasi yang nantinya dapat digunakan[8].Dalam buku *data mining* teori dan aplikasi rapidminer menjelaskan bahwa *data mining* merupakan bagian proses untuk menggali nilai tambah berupa informasi yang belum diketahui secara manual dari basis data. Informasi yang didapat diperoleh dengan cara mengekstrak dan mengenali pola yang penting dari data yang ada pada basis data. Tujuan utama *data mining* adalah digunakan untuk mencari pengetahuan yang ada dalam basis data yang berukuran besar[9]. Klasifikasi merupakan salah satu model *data mining* yang sering digunakan untuk menyelasaikan masalah penentuan dalam mengambil keputusan. Dalam jurnalnya, Salean (2020) menyatakan klasifikasi juga disebut *supervised learning* yakni memasukkan klasifikasi nilai respon kategori untuk memisahkan data ke dalam kelas-kelas tertentu [7]. Klasifikasi adalah proses dua langkah. Langkah pertama menggunakan algoritma klasifikasi. Pada langkah kedua, model dilatih untuk mengukur kinerja dan akurasi, yang mana *input*-an akan diklasifikasikan dan menghasilkan *output* berupa label kelas. Tujuan utama dari klasifikasi adalah untuk mengklasifikasikan data ke dalam kelas yang berbeda sesuai dengan batasannya yang digunakan dalam memprediksi kelas sasaran dengan menganalisis data latih untuk menghasilkan nilai respon.

Algoritma C5.0 merupakan proses pembuatan pohon keputusan yang mirip dengan algoritma C4.5, dimana kemiripan tersebut adalah dalam perhitungan *entrophy* dan *gain*. Jika dalam perhiitungan algoritma *C4.5* berhenti sampai pada perhitungan *entrophy* dan *gain*, maka dalam algoritma *C5.0* perhitungan tersebut masih dilanjutkan perhitungan *gain ratio* dengan menggunakan hasil dari *gain* dan *entrophy* yang telah dihitung sebelumnya[6]. Berikut proses perhitungan algoritma C5.0 dengan menggunakan rumus persamaan (1).

$$entropy\ (s) = \ -\sum_{i=1}^{n} Pi * log_2\ p_i \qquad (1)$$

Dimana keterangan dari rumus tersebut adalah:
S = Himpunan Kasus
n = Jumlah banyaknya partisi S
pi = Porsi dari Si kepada S

Setelah tahap pertama selesai dan mendapatkan hasil entrophy dari setiap kelasnya, maka selanjutnya adalah mencari *information gain* dengan rumus persamaan (2).

$$information\ gain\ (S, A) = \ entrophy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * entropy(S_i) \qquad (2)$$

Dimana keterangan dari rumus tersebut adalah:
S = Himpunan Kasus
A = Atribut
n = Jumlah banyaknya atribut A
$|Si|$ = Jumlah kasus pada nilai atribut/kriteria
$|S|$ = Jumlah kasus atau total kasus

Kemudian setelah ke dua tahap tersebut selesai, maka dapat mencari nilai *GainRatio* menggunakan perhitungan dengan rumus persamaan (3) berikut ini:

$$GainRatio = \frac{information\ gain\ (S,A)}{\sum_{i=1}^{n} entrophy\ (Si)} \quad (3)$$

Dari hasil perhitungan *gain ratio, gain ratio* terbesar akan terpilih sebagai akar (*root*) dan yang rendah akan menjadi cabang. Kemudian proses *gain ratio* diulang sampai masing-masing cabang pada semua kelas memiliki kelasnya.

Naïve bayes dalam [3] adalah metode klasifikasi berdasarkan *Teorema Bayes*. Metode klasifikasi ini menggunakan metode probabilistik dan statistik yang dikenalkan pertama kali oleh seorang ilmuwan Inggris bernama Thomas Bayes, yakni metode untuk memprediksi peristiwa masa depan berdasarkan pengalaman masa lalu. Oleh karena itu, metode ini dikenal sebagai *Teorema Bayes*. *Naive bayes* juga sering memperoleh hasil yang jauh lebih baik dari yang diharapkan dalam situasi kehidupan nyata yang kompleks. Berikut poses penyelesaian perhitungan *naïve bayes* dapat menggunakan rumus persamaan (4) berikut:

$$P\ (H|X) = \frac{P\ (X|H)\ .\ P(H)}{P\ (X)} \quad (4)$$

Dimana keterangan dari rumus tersebut adalah:
X = Data dengan class yang belum diketahui
H = Hipotesis data merupakan suatu class spesifik
$P\ (H|X)$ = Probabilitas hipotesis H berdasar kondisi X (Posteriori probabilitas)
$P\ (H)$ = Probabilitas hipotesis H (Prior probabilitas)
$P\ (X|H)$ = Probabilitas X berdasarkan kondisi pada hipotesis H
$P\ (X)$ = Probabilitas X

Adapun langkah-langkah dalam perhitungan *naïve bayes* adalah sebagai berikut:
1. Tahap pertama yang dilakukan adalah menghitung jumlah kelas
2. Tahap kedua menghitung jumlah kasus yang sama dengan kelas yang sama
3. Tahap ketiga kalikan semua variable kelas yang akan dihitung
4. Tahap keempat membandingkan hasil kelas, yang mana nilai terbesar dari kelas tersebut merupakan hasil.

Hal ini selaras dengan penelitian terdahulu yang terkait antara lain: tentang Perbandingan algoritma *ID3* dan *C5.0* Dalam Identifikasi Penjurusan Siswa SMA[4]. Dalam penelitian ini data yang digunakan sebanyak 200 data siswa yang dibagi menjadi 2 yaitu 150 *data training* dan 50 sebagai *data testing*, yang mana hasil nilai akurasi pada algoritma *C5.0* sebesar 95% dan algoritma *ID3* sebesar 93%. Hal ini membuktikan bahwa kinerja algoritma *C5.0* lebih baik dibandingkan *algoritma ID3*. Selanjutnya dalam penelitian [10] dengan judul Analisis Komparasi *Algoritma* Klasifikasi *Data Mining* untuk Prediksi Penjurusan Siswa Menengah Atas (SMA) Pramita Karawaci Tangerang. Dalam penelitian ini data yang digunakan sebanyak 365 data siswa yang dibagi menjadi data training dan data testing dengan menguji data sebanyak 4 kali. Hasil nilai akurasi pada *naïve bayes* sangat baik yaitu sebesar 87,19% dibandingkan dengan *decision tree* sebesar 82,11%, sehingga *naïve bayes* merupakan metode yang cocok untuk diterapkan dalam menentukan jurusan siswa. Dari kedua penelitian yang telah dilakukan menunjukan bahwa algoritma C5.0 dan naïve bayes merupakan metode klasifikasi paling baik dengan hasil tingkat akurasi yang sangat tinggi yaitu sebesar 93% untuk algoritma C5.0 dan 87,19% untuk *naïve bayes*.

Berdasarkan uraian permasalahan-permasalahan dalam latar belakang di atas, peneliti tertarik untuk membandingkan metode *algoritma C5.0* dan *naïve bayes* pada SMAN 1 Gondanglegi. Peneliti akan mengkomparasi metode klasifikasi *data mining* algoritma *C5.0* dan *naïve bayes* dalam memprediksi jurusan siswa untuk mendapatkan hasil tingkat akurasi paling tinggi, yang mana dapat digunakan sebagai solusi dari permasalahan yang ada di SMAN 1 Gondanglegi. Diharapkan penelitian ini akan memberi kemudahan dan dapat menjadi salah satu sarana untuk rekomendasi penjurusan siswa yang efisien dan efektif.

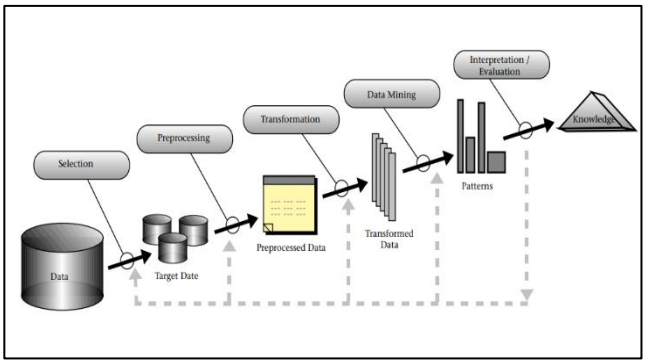
2. METODE PENELITIAN

A. Metode Pengumpulan Data

1. Studi literartur, metode ini dilakukan untuk mengumpulkan data yang terkait dengan topik penelitian dengan mencari sumber referensi seperti buku, jurnal dan skripsi. Referensi yang peneliti gunakan adalah terkait dengan data mining dan metode klasifikasi untuk penjurusan siswa.
2. Wawancara, metode ini dilakukan melalui tatap muka dan tanya jawab langsung antara pengumpul data terhadap narasumber. Adapun narasumber dari wawancara ini adalah operator pendataan dan guru bimbingan konseling di SMAN 1 Gondanglegi dengan tujuan untuk mendapatkan informasi terkait sekolah dan data siswa yang diperlukan oleh peneliti.
3. Observasi, metode ini dilakukan dengan cara melakukan kunjungan langsung pada SMAN 1 Gondanglegi untuk mendapatkan data siswa kelas X tahun ajaran 2021/2022.

B. Metode *knowledge discovery in database* (KDD) yang digunakan dalam penelitian ini:

Metode *knowledge discovery in database* (KDD) memiliki proses sebagai berikut [5]:



Gambar 1. Proses Tahapan Metode KDD

a. Data Selection

Tidak semua data yang ada dalam basis data dapat digunakan pada proses *data mining*. Tahap ini akan dilakukan seleksi data atau memilih atribut yang akan digunakan pada proses *data mining*. Struktur basis data yang tersedia antara lain: NISN, nama, JK, tanggal lahir, alamat, asal sekolah, domisili, nilai TPA, nilai tes IQ dan data-data pribadi siswa lainnya. Adapun data atau atribut yang dipilih adalah nilai tes IQ, nilai TPA, nama dan kelas. Atribut yang tidak dipilih merupakan atribut yang tidak perlu atau tidak penting dan juga bersifat rahasia.

b. Data Cleaning (Pre-processing)

Untuk mendapatkan data yang relevan harus melakukan pembersihan data. Pembersihan data yang dilakukan antara lain: memeriksa, memperbaiki dan menghapus data yang tidak relevan. Proses pada penelitian menggunakan metode *filter examples* yaitu menghapus data yang kosong (*missing value*). Setelah data yang diperlukan melalui tahap *data cleaning* maka data akan disimpan pada *dataset* baru dengan cara mengimpor data atau *write* excel sehingga akan tersimpan ke dalam file excel atau xlsx.

c. Data Transformation

Dalam tahap transformasi data yaitu dilakukan pada setiap atribut untuk merubah data yang sesuai dengan data yang di-*mining*. Pada tahap ini dilakukan pengkategorian terhadap data dengan mengubah data menjadi tipe *numeric* ke data dengan tipe nominal. Kemudian data akan diubah ke dalam format *csv* atau *xlsx* melalui Microsoft Excel.

d. Data Mining

Data mining merupakan kajian yang mencakup proses pengumpulan data, seleksi data, pembersihan data dan transformasi data, sehingga dengan aktivitas tersebut dapat memperolah pengertian yang mendalam akan data. Setelah melakukan proses pembersihan data dan transformasi data yang sesuai untuk penggunaan proses *data mining* selanjutnya akan dilakukan dengan metode klasifikasi yang mana data tersebut akan dibagi menjadi 2 yaitu data latih dan data uji. Data latih adalah data yang akan digunakan sebagai *data training* dalam menentukan pola data. Sedangkan data uji adalah data yang akan digunakan untuk menghasilkan performa dari model yang digunakan. Kemudian proses klasifikasi dilakukan menggunakan *algoritma C5.0* dan *naïve bayes*.

e. Evaluation

penelitian ini sebelum melakukan tahap evaluasi akan dilakukan pengujian dengan metode *k-fold cross validation* dengan k=10. Data akan dibagi menjadi 10 bagian secara *random* sehingga diperoleh akurasi yang optimal dari data tersebut. Selanjutnya tahap evaluasi yaitu hasil dari *cross validation* akan dicatat ke dalam *confusion matrix* yang akan menghasilkan nilai performa *accuracy, precision* dan *recall*. *Confusion matrix* yang digunakan yaitu matrix 3 *class*, maka untuk menghitung tingkat akurasi tersebut digunakan rumus persamaan berikut ini:

$$Accuracy = \frac{(TP)}{(TP+TN+FN+FP)} * 100 \quad (5)$$

Setelah proses evaluasi selesai akan dilakukan perbandingan performa antara algoritma C5.0 dan naïve bayes dengan melihat nilai *accuracy, precision* dan *recall* yang paling tinggi, sehingga didapat metode yang paling baik. Proses pengujian dan evaluasi dalam penelitian ini dilakukan menggunakan *software* Rapidminer.

3. HASIL DAN ANALISIS

3.1. Hasil Pembagian Data

Sebelum melakukan proses *data mining*, maka data dipecah menjadi 2 bagian:

1. Data Latih
2. Data Uji

Keduanya dibagi dengan proporsi 80% data latih dan 20% data uji. Adapun hasil dari pembagaian data dapat dilihat pada tabel berikut:

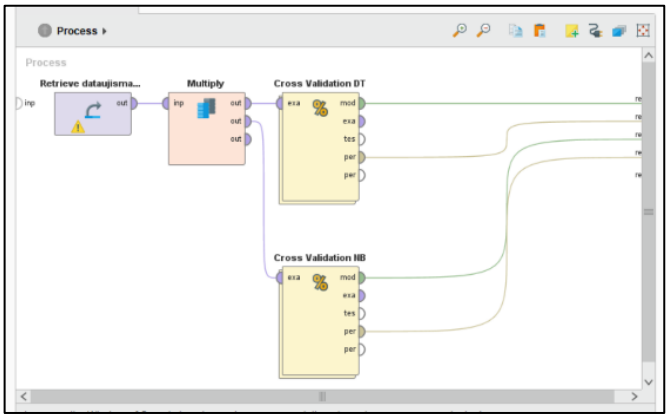
Tabel 1. Hasil Pembagian Data

Klasifikasi	Jumlah Data	Data Latih (80%)	Data Uji (20%)
MIPA	105	84	21
IPS	104	83	21
BHS	22	18	4
Total	231	185	46

Jumlah data latih yang diperoleh adalah 185 data, sedangkan data uji sebanyak 46 data. Data tersebut sudah diolah sehingga data siap untuk digunakan ke tahap selanjutnya.

3.2. Pengujian Cross Validation

Cross Validation adalah teknik validasi yang membagi data secara acak menjadi k dalam setiap bagian yang dilakukan menggunakan proses klasifikasi. Teknik ini dapat memperkirakan akurasi prediksi suatu model ketika diterapkan pada proses prediksi. Metode ini digunakan dalam proses pengujian untuk mengukur kinerja dari metode klasifikasi yang digunakan[1]. Dalam tahap ini data yang digunakan adalah data uji. Berikut merupakan proses *cross validation*:



Gambar 2. Proses Cross Validation C5.0 dan NB

Gambar di atas merupakan proses *cross validation* menggunakan metode *multiply* untuk membandingkan hasil akurasi antara *algoritma C5.0* dan *naïve bayes*. Jenis pengujian *cross validation* yang digunakan yakni *k-fold cross validation* dengan k-10 yang berarti data akan di uji sebanyak 10 kali.

3.3. Evaluasi Hasil Klasifikasi

Evaluasi hasil klasifikasi yang digunakan adalah pada data yang telah dilakukan pengujian menggunakan *10-fold cross validation***.** Metode evaluasi yang digunakan ialah *confussion matrix* yang menghasilkan nilai performa akurasi, *recall* dan *precision*. Dalam penelitian metode *confussion matrix* 3x3 dikarenakan hasil dari penelitian ada 3 kategori yaitu jurusan MIPA, IPS dan Bahasa. Adapun rumus dalam [2] *multiclass confusion matrix* 3x3 atau data yang terdiri dari tiga kelas dapat dilihat pada tabel berikut:

Tabel 2. Rumus Confusion Matrix 3 Class

		PREDIKSI		
		Class 1	Class 2	Class 3
AKTUAL	Class 1	TP	FN	FN
	Class 2	FP	TP	TN
	Class 3	FP	TN	TP

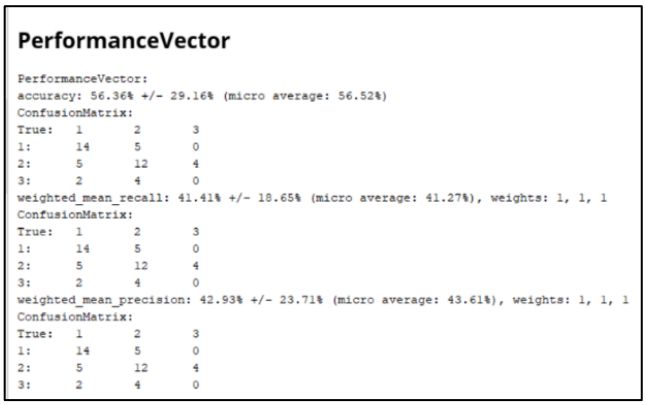
Dengan bantuan software rapidminer, maka diperoleh hasil confusion matix untuk metode algoritma C5.0 dapat dilihat pada gambar 3 dan *naïve bayes* dapat dilihat pada gambar 4.

PerformanceVector

```
PerformanceVector:
accuracy: 60.91% +/- 29.82% (micro average: 60.87%)
ConfusionMatrix:
True:   1       2       3
1:      16      7       2
2:      5       12      2
3:      0       2       0
weighted_mean_recall: 43.43% +/- 21.35% (micro average: 44.44%), weights: 1, 1, 1
ConfusionMatrix:
True:   1       2       3
1:      16      7       2
2:      5       12      2
3:      0       2       0
weighted_mean_precision: 44.44% +/- 22.08% (micro average: 42.39%), weights: 1, 1, 1
ConfusionMatrix:
True:   1       2       3
1:      16      7       2
2:      5       12      2
3:      0       2       0
```

Gambar 3. Confusion Matrix C5.0

Gambar di atas merupakan hasil evaluasi klasifikasi algoritma C5.0 menggunakan *confussion matrix*. Hasil akurasi yang diperoleh yakni sebesar 60,91%. Nilai *precision* 44,44% sedangkan untuk hasil nilai *recall* yaitu sebesar 43,43%.



PerformanceVector

```
PerformanceVector:
accuracy: 56.36% +/- 29.16% (micro average: 56.52%)
ConfusionMatrix:
True:   1       2       3
1:      14      5       0
2:      5       12      4
3:      2       4       0
weighted_mean_recall: 41.41% +/- 18.65% (micro average: 41.27%), weights: 1, 1, 1
ConfusionMatrix:
True:   1       2       3
1:      14      5       0
2:      5       12      4
3:      2       4       0
weighted_mean_precision: 42.93% +/- 23.71% (micro average: 43.61%), weights: 1, 1, 1
ConfusionMatrix:
True:   1       2       3
1:      14      5       0
2:      5       12      4
3:      2       4       0
```

Gambar 4. Confusion Matrix NB

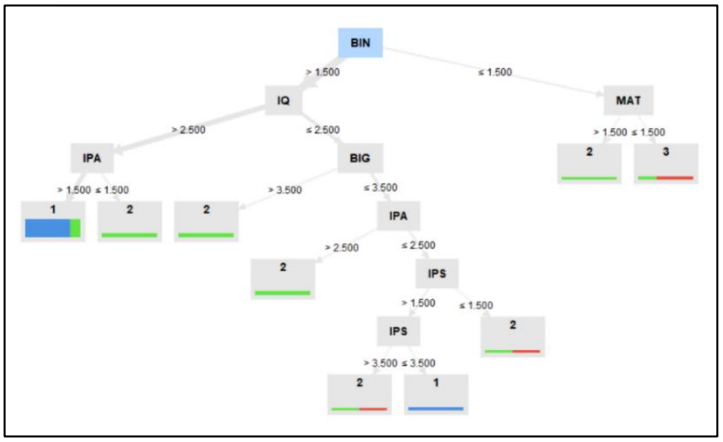
Gambar di atas merupakan hasil evaluasi klasifikasi dari naïve bayes menggunakan metode *confusion matrix*. Hasil akurasi yang didapat yakni sebesar 56,36%. Nilai *precision* 42,93% sedangkan untuk hasil nilai *recall* yaitu 41,41%. Dari hasil evaluasi klasifikasi, maka akan dilakukan perbandingan dengan nilai rata-rata akurasi, *precision* dan *recall* hasil dari performa algoritma *C5.0* dan *naïve bayes*. Adapun hasil dari masing-masing performa adalah sebagai berikut:

Tabel 3. Hasil Performa klasifikasi Algoritma C5.0 dan Naive Bayes

Klasifikasi	Akurasi (%)	***Precision (%)***	***Recall (%)***
Algoritma C5.0	60,87%	42,39%	44,44%
Naïve Bayes	56,52%	43,61%	41,27%

Hasil yang diperoleh dari performa tersebut untuk nilai akurasi algoritma *c5.0* sebesar 60,87%, *precision* 42,39% dan *recall* 44,44%. Sedangkan pada klasifikasi *naïve bayes* nilai akurasi sebesar 56,52% *precision* 43,61% dan *recall* 41,27%. Dari hasil tersebut dapat dilihat bahwa performa dari algoritma *c5.0* memiliki hasil yang lebih baik dibandingkan dengan metode *naïve bayes*. Sehingga metode klasifikasi paling baik adalah algoritma *C5.0*.

3.4. Implementasi Algoritma Terbaik



Gambar 5. Pohon Keputusan

Gambar di atas merupakan hasil implementasi algoritma *C5.0* dalam membentuk pohon keputusan. Dari hasil pohon keputusan terdapat 9 aturan (*rule*) dengan akar pertama atau *node* 1 yaitu BIN. Yang berarti nilai Bahasa Indonesia paling berpengaruh dalam klasifikasifikasi memprediksi jurusan siswa. Faktor-faktor yang signifikan mempengaruhi dalam memprediksi jurusan adalah nilai tes TPA, nilai tes IQ dan untuk faktor yang paling mempengaruhi adalah nilai Bahasa Indonesia.

4. KESIMPULAN

Berdasarkan hasil perbandingan metode algoritma *C5.0* dan *naïve bayes* menghasilkan nilai akurasi masing-masing sebesar 60,87% dan 56,52%. Dengan nilai rata-rata *precision* dan *recall* sebesar 42,39% dan 44,44% untuk *algoritma C5.0* serta 43,61% dan 41,27% untuk *naïve bayes*. Dari hasil yang didapat metode algoritma *C5.0* memiliki tingkat akurasi yang lebih tinggi dibanding metode *naïve bayes*. Sehingga metode algoritma *c5.0* merupakan metode paling baik dalam klasifikasi menentukan jurusan siswa di SMAN 1 Gondanglegi.

Saran untuk penelitian selanjutnya adalah dapat menambah dataset penelitian yang digunakan dari tahun ajaran 2019/2020 sampai tahun ajaran 2021/2022 atau tiga tahun ajaran agar hasil lebih akurat dan pengetahuan yang dihasilkan menjadi lebih baik. Menambahkan faktor-faktor lain sebagai atribut. Menggunakan *software* yang lainnya seperti *Weka, Orange,* atau dikembangkan dengan membuat program sendiri.

UCAPAN TERIMAKASIH

Dalam kesempatan ini penulis menyampaikan ucapan terimakasih kepada Universitas Islam Raden Rahmat Malang, Fakultas Sains dan Teknologi, Program Studi Sistem Informasi, serta kepada SMAN 1 Gondanglegi.

REFERENSI

[1] Anestiviya, V., & Pasaribu, A. F. O., "Analisis Pola Menggunakan Metode C4. 5 Untuk Peminatan Jurusan Siswa Berdasarkan Kurikulum (Studi Kasus: SMAN 1 Natar)", *Jurnal Teknologi dan Sistem Informasi (JTSI),* vol 2, no 1, hal. 80–85, Maret 2021.

[2] Fauziah Afreyna, D., Maududie, A., & Nuritha, I., "Klasifikasi Berita Menggunakan Algoritma K-nearst Neighbor", *Berkala Saintek,* vol 6 no 2, hal. 106-114, 2018.

[3] Hidayanti, I., Kurniawan, T. B., & Afriyudi, "Perbandingan Dan Analisis Metode Klasifikasi Untuk Menentukan Konsentrasi Jurusan", *Jurnal Ilmiah Global*, vol 11 no 1, hal. 16–21, Juli 2020.

[4] Munawaroh, H., Khusnul, B., & Kustiyahningsih, Y, "Perbandingan Algoritma ID3 dan C5.0 dalam Indentifikasi Penjurusan Siswa SMA", *Jurnal Sarjana Teknik Informatika*, vol 4, no 1, hal. 1–12, Juni 2013.

[5] Nofriansyah, D., Erwansyah, K., & Ramadhan, M., "Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL (Studi Kasus di CV. Sumber Utama Telekomunikasi)", *Jurnal Ilmiah Saintikom*, vol 15, no 2, hal. 81–92, Mei 2016.

[6] Putri, R. Y., Mukhlash, I., & Hidayat, N., "Prediksi Pola Kecelakaan Kerja Pada Perusahaan Non Ekstraktif Menggunakan Algoritma Decision Tree: C4.5 dan C5.0", *Jurnal Sains Dan Seni Pomits*, vol 2 no 1, hal. 1-6, 2013.

[7] Saelan, M. R. R., Sahputra, D. A., & Gata, W., "Komparasi Algoritma Klasifikasi untuk Prediksi Minat Sekolah Tinggi Pelajar pada Students Alcohol Consumption", *Jurnal Sains Dan Informatika*, vol 6, no 2, hal. 120–129, November 2020.

[8] Saleh, A., "Penerapan Data Mining Dalam Menentukan Jurusan Siswa", *Seminar Nasional Informatika*, 2015, hal. 351–355.

[9] Vulandari, R. T., Data Mining Teori dan Aplikasi Rapidminer, Surakarta: Gava Media, 2017.

[10] Yulianti, H, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penjurusan Siswa Sekolah Menengah Atas (SMA) Pramita Karawaci Tangerang", *LENSA*, vol *2*, no 48, hal. 1–6, September 2019.