Jurnal Elektronik Ilmu Komputer Udayana
Volume 12, No 4. May 2024

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Perbandingan Kinerja Algoritma C4.5 dan Naive Bayes Untuk Klasifikasi Penerima Beasiswa

Lia Susanti[a1], Khoirunnisa[a2]

[a]Program Studi Teknik Informatika, Universitas Indraprasta PGRI
Kota Jakarta Timur, Daerah Khusus Ibukota Jakarta, Indonesia
[1]liasusanti.s4a.061@gmail.com
[2]khoirunnisa.1797@gmail.com

Abstract

One of the most important factors is education, but many students who have great abilities and potential cannot continue school because they are not financially capable but there are also many able-bodied students who get scholarships. In the educational environment, especially schools, there should be some rules or classification in determining students who receive scholarships Therefore, in this study, a comparison of the C4.5 and Naïve Bayes algorithms was applied to the data of students who received scholarships. This study aims to find a pattern that can determine the award of scholarships with predetermined criteria, in the selection of prospective scholarship recipients at SMKS 2 Adi Luhur using the Naive Bayes Algorithm and the C 4.5 Algorithm. The data will be tested using k-fold cross validation (k=10). From the comparison results, the results of the Naïve Bayes accuracy are higher than the C 4.5 Algorithm. The results obtained from the comparison of the two algorithms are the Naïve Bayes algorithm has an accuracy rate of 94.52% and the C4.5 algorithm has an accuracy rate of 92.52%.

Keywords: *Scholarship, C4.5 Algorithm, Naïve Bayes, K-Fold Cross Validation, Classification*

1. Pendahuluan

Salah satu masalah pendidikan yang dihadapi bangsa Indonesia saat ini adalah bagaimana meningkatkan mutu pendidikan di setiap jenjang, khususnya jenjang sekolah menengah atas agar mampu bersaing di era global. Ilmu pengetahuan seseorang dapat diperoleh melalui pendidikan di sekolah, karena dengan bersekolah kita akan mampu mewujudkan keberhasilan dan kesuksesan dalam kehidupan. Namun, kenyataan di lapangan menunjukkan bahwa ekonomi yang terbatas bagi sebagian orang tua menjadi faktor penghambat dalam mewujudkan kesuksesan anak, sehingga tidak semua anak usia wajib belajar dapat mengikuti pendidikan di sekolah. Untuk mengatasi permasalahan ini, sekolah mengadakan program beasiswa pendidikan. Yayasan Islam Adi Luhur 2 merupakan salah satu lembaga pendidikan yang memiliki SMA dan SMK, di mana sekolah ini mempunyai program Bantuan Dana Pendidikan bagi siswa yang dianggap kurang mampu secara status ekonomi. Namun, ada syarat dan ketentuan yang berlaku, dan semua kriteria pemilihan penentuan siswa yang memperoleh bantuan dana pendidikan itu dilakukan berdasarkan data siswa yang ada, lalu dianalisis secara manual, namun terkadang hasil yang diperoleh tidak sesuai. Seleksi penerima beasiswa merupakan salah satu kegiatan pengambilan keputusan yang cukup rumit karena ada beragam variabel yang digunakan sebagai bahan pertimbangan dalam proses seleksi. Variabel yang digunakan juga akan bergantung kepada jenis beasiswa yang diberikan. SMKS 2 Adi Luhur merupakan salah satu sekolah swasta di Jakarta yang berperan aktif dalam meningkatkan mutu pendidikan terhadap siswa. Saat ini, SMKS 2 Adi Luhur memiliki jumlah siswa yang cukup besar pada setiap tahun ajaran baru, kurang lebih mencapai 300-400 siswa. Dari jumlah siswa yang besar, maka dibutuhkan suatu metode yang tepat dalam menentukan beasiswa. Dengan demikian, pada penelitian ini penulis menggunakan dua Algoritma Klasifikasi Data Mining, yaitu Algoritma C4.5 dan Naive Bayes, untuk memprediksi siswa dalam memperoleh bantuan dana pendidikan agar dalam pemberian beasiswa sesuai sasaran. Sampel data diambil dari Sekolah Menengah Kejuruan (SMK) 2 Adi Luhur pada tahun 2018/2019 yang sudah di-cleaning sebanyak 148 data siswa.

Dari penelitian ini dapat dibandingkan dengan beberapa penelitian sebelumnya antara lain penelitian yang dilakukan tentang penelitian di Universitas Hamzanwadi menggunakan Naive Bayes Classifier dengan variabel seperti status DTKS, prestasi, pekerjaan orang tua, dan lainnya. Akurasi tertinggi

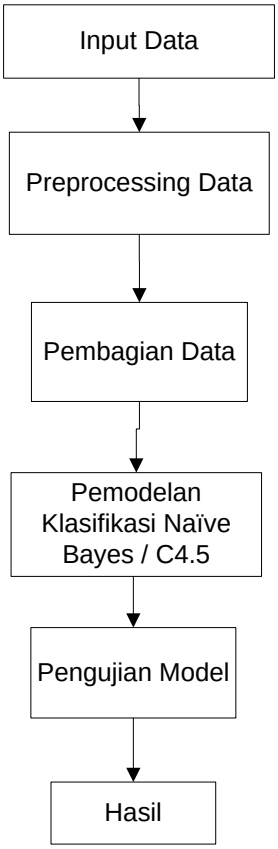
91,43% dengan AUC 0,996% yang menunjukkan algoritma ini sangat baik digunakan [1]. Kemudian Penelitian di Universitas Negeri Medan menggunakan Naive Bayes Classifier dengan variabel seperti pekerjaan orang tua, penghasilan, jumlah tanggungan, daya listrik, dan nilai UN. Akurasi tertinggi 79% dengan perbandingan data training dan testing 80:20 [2].

[3] Penelitian di SMK Swasta Anak Bangsa menggunakan C4.5 dengan variabel utama nilai siswa. Akurasi 92,7%, precision 92,05% untuk "Layak" dan 93,24% untuk "Tidak Layak".

Tujuan dari penelitian ini adalah untuk mencari pola yang dapat menentukan pemberian beasiswa dengan kriteria yang sudah ditetapkan, dengan membandingkan tingkat akurasi yang dihasilkan oleh teknik atau model data mining, yaitu Algoritma C4.5 dan Naive Bayes, di mana metode Naive Bayes digunakan untuk menentukan siswa yang berprestasi berdasarkan data identitas siswa dan data akademik siswa SMKS 2 Adi Luhur, sehingga dapat memberikan masukan kepada sekolah untuk mempermudah sistem dalam memberikan beasiswa.

2. Metode Penelitian

Penelitian ini menggunakan metodologi yang terdiri dari beberapa tahapan seperti, Pada gambar alur metode penelitian ini pada Gambar 1 berikut.



Gambar 1. Metode Penelitan
Sumber : Olah Pribadi

1. Input Data
   Dataset yang digunakan pada penelitian ini yaitu dataset siswa SMK , sebuah dataset dibagi menjadi dataset training dan testing Dataset ini diperoleh dari internal di SMKS 2 Adi Luhur tempat dilakukan penelitian.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

No	Kelas	Pekerjaan Orang Tua	Penghasilan orang Tua	Pengeluaran Orang tua	Jumlah Tanggungan orangtua	Rangking	Nilai Raport	Kerajinan	Kepribadian	Remark
1	XI	Wiraswasta	<1,000,000	<1,000,000	2-4 orang	tidak ada	>70	cukup	kurang	tidak dapat
2	XI	Tetap	>2,000,000	1500000 – 2000000	>4 orang	tidak ada	>70	cukup	cukup	tidak dapat
3	XI	Tetap	>2,000,000	1500000 – 2000000	>4 orang	2 s/d 4	>80	cukup	baik	tidak dapat
4	XI	Tidak Tetap	1500000 – 2000000	1500000 – 2000000	>4 orang	tidak ada	>70	cukup	cukup	tidak dapat
5	XI	Wiraswasta	<1,000,000	<1,000,000	2-4 orang	tidak ada	>60	cukup	kurang	dapat
6	XI	Wiraswasta	<1,000,000	<1,000,000	2-4 orang	5 s/d 10	>75	baik	baik	tidak dapat
7	XI	Wiraswasta	<1,000,000	<1,000,000	>4 orang	5 s/d 10	>75	cukup	cukup	tidak dapat
8	XI	Wiraswasta	<1,000,000	<1,000,000	2-4 orang	5 s/d 10	>75	baik	baik	tidak dapat
9	XI	Wiraswasta	<1,000,000	<1,000,000	2-4 orang	tidak ada	>70	cukup	kurang	dapat
10	XI	Tetap	>2,000,000	1500000 – 2000000	>4 orang	tidak ada	>70	cukup	cukup	dapat
11	X	Tetap	>2,000,000	1500000 – 2000000	2-4 orang	2 s/d 4	>80	baik	baik	tidak dapat
12	X	Tidak Tetap	1500000 – 2000000	1500000 – 2000000	2-4 orang	tidak ada	>70	cukup	baik	tidak dapat
13	X	Wiraswasta	>2,000,000	1500000 – 2000000	2-4 orang	2 s/d 4	>80	baik	baik	tidak dapat
14	X	Wiraswasta	>2,000,000	1500000 – 2000000	2-4 orang	tidak ada	>70	kurang	kurang	tidak dapat
15	X	Wiraswasta	>2,000,000	1500000 – 2000000	2-4 orang	tidak ada	>70	cukup	cukup	tidak dapat
16	X	Tidak Tetap	1500000 – 2000000	1500000 – 2000000	2-4 orang	5 s/d 10	>75	cukup	cukup	tidak dapat
17	X	Tidak Tetap	1500000 – 2000000	1500000 – 2000000	2-4 orang	tidak ada	>60	kurang	kurang	tidak dapat
18	X	Tidak Bekerja	<1,000,000	<1,000,000	2-4 orang	tidak ada	>70	cukup	kurang	tidak dapat
19	X	Tetap	>2,000,000	1500000 – 2000000	2-4 orang	1	>95	baik	baik	dapat
20	X	Tetap	>2,000,000	1500000 – 2000000	2-4 orang	2 s/d 4	>80	baik	baik	dapat

Gambar 2. Dataset Siswa SMK
Sumber : Olah Pribadi

Data tersebut memiliki 9 atribut. Berikut beberapa atribut-atribut yang ada pada dataset :

Tabel 1. Data Atribut Siswa

No	Atribut	Nilai Atribut
1	Kelas	X XI
2	Pekerjaan Orang Tua	Tidak Tetap Wiraswasta Tidak Bekerja Tetap
3	Penghasilan Orang Tua	>2.000.000 1.500.000-2.000.000 <1.000.000
4	Pengeluaran Orang Tua	1.500.000-2.000.000 <1.000.000
5	Jumlah Tanggungan Orang Tua	2s/d 4 orang
6	Rangking	1-10
7	Nilai Raport	60 s/d 95
8	Kerajinan	Baik Cukup Kurang
9	Kepribadian	Baik Cukup

		Kurang
	Remark/Hasil	Dapat/Tidak Dapat
10		

Pada saat penelitian ini dilakukan proses validasi untuk menemukan, dan mengkoversi data agar dapat digunakan dalam metode algoritma data mining dan memperoleh akurasi serta performasi yang baik. Dalam dataset yang akan digunakan, validasi data yang digunakan dengan hapus data yang tidak lengkap atau kosong yang tidak memiliki nilai (null). Setelah itu dilakukan seleksi atribut untuk memilih atribut mana saja yang dibutuhkan dari data set yang digunakan dalam proses menganalisis kelayakan pemberian beasiswa kepada calon siswa. Atribut yang diambil dari data pengajuan beasiswa.

2. Preprocessing Data
   Tahap Pada tahapain ini, dataset dilakukan serangkaian langkah untuk menyiapkan data yang diperlukan sebelum digunakan dalam pemodelan menggunakan algoritma C4.5 dan algoritma Naïve Bayes.
3. Pembagian Data
   Mula-mula data diberikan pelabelan secara manual dan setelah itu dilakukanlah pembagian data menjadi 2 yaitu data training dan data testing. Dimana terdapat 2 kelompok dalam penelitian ini yaitu kelompok data yang pertama adalah jumlah data testing lebih sedikit dari data training. Sedangkan kelompok data kedua adalah data lain lebih banyak dari data training.
4. Algoritma C4.5
   Merupakan salah satu metode klasifikasi yang digunakan. Melibatkan konstruksi pohon keputusan, koleksi node keputusan. Setiap cabang kemudian mengarah ke node lain baik keputusan atau ke node daun untuk mengakhiri [4]. C4.5 adalah algoritma yang mempunyai input berupa training samples berupa data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya dan samples yang merupakan field - field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data. Algoritma dasar dari C4.5 adalah sebagai berikut:
   a. Pohon yang dihasilkan berupa pohon terbalik,
   b. Pada tahap awal, semua contoh training adalah akar.
   c. Atribut adalah kategori.
   d. Contoh di partisi secara berulang berdasarkan atribut yang dipilih.
   e. Atribut tes dipilih dari data heuristic atau pengukuran statistik

Karena algoritma C4.5 digunakan untuk melakukan klasifikasi, jadi hasil dari pengolahan test dataset berupa pengelompokkan data ke dalam kelas-kelasnya Umumnya, langkah-langkah algoritma C4.5 yang digunakan untuk membentuk pohon keputusan adalah :

a. Pilih atribut sebagai root.
b. Buat cabang untuk setiap nilai
c. Bagi tiap cabang kedalam kelas.
d. Ulangi proses untuk setiap cabang sampai semua kasus pada tiap cabang memiliki kelas yang sama

5. Algoritma Naive Bayes
   Merupakan teknik probabilistik klasifikasi berdasarkan teorema Bayes dengan asumsi independensi diantara variabel prediktor. Secara sederhana, pengelompokan Naïve Bayes menganggap adanya suatu fitur tertentu dalam sebuah kelas tidak terkait dengan adanya fitur lainnya. Rumus Persamaan dari Teorema Bayes adalah [5]:

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

Keterangan:
X : Data dengan class yang belum diketahui
H : Hipotesis data X merupakan suatu class spesifik
P(H|X) : Probabilitas hipotesis *H* berdasar kondisi *X* (posteriori probability)
P(H) : Probabilitas hipotesis *H (*posteriori probability)
P(X|H) : Probabilitas *X* berdasarkan kondisi pada hipotesis *H*
P(X) : Probabilitas *X*

p-ISSN: 2301-5373
e-ISSN: 2654-5101

3. Hasil dan Pembahasan

3.1 Evaluasi dan Validasi Hasil

a. Hasil Pengujian Algoritma C4.5

Setelah model klasifikasi data mining prediksi penerima beasiswa terbentuk, untuk menilai keakuratan sebuah model akan dilakukan pengujian dengan teknik K-Fold Cross Validation dimana nilai k=10, yang berarti proses pengulangan sebanyak 10 kali pengulangan. Hasilnya model memiliki keakuratan sebesar 92.52%.

accuracy: 92.52% +/- 5.64% (mikro: 92.57%)

	true tidak dapat	true dapat	class precision
pred. tidak dapat	122	9	93.13%
pred. dapat	2	15	88.24%
class recall	98.39%	62.50%	

Gambar 1. Model Klasifikasi Algoritma C4.5 pada RapidMiner
Sumber : Olah Pribadi

Selain dilakukan pengujian model dengan teknik K-Fold Cross Validation, untuk mengevaluasi kinerja model klasifikasi yang dapat memprediksi benar atau tidak benarnya oleh model tersebut, maka digunakan teknik tabel Confusion Matrix. Berikut hasil tabel ukur Confusion Matrix terhadap algoritma C4.5 dengan menghasilkan akurasi sebesar 92.52%

PerformanceVector

```
PerformanceVector:
accuracy: 92.52% +/- 5.64% (mikro: 92.57%)
ConfusionMatrix:
True:   tidak dapat     dapat
tidak dapat:    122     9
dapat:  2       15
```

Gambar 4. Hasil Confusion Matrix Algoritma C4.5
Sumber : Olah Pribadi

a. Hasil Pengujian Algoritma Naive Bayes

Pada tahap ini peneliti menggunakan metode algoritma Naive Bayes untuk mengaplikasikan data yang telah mengalami proses preprocessing data atau pembersihan data pada aplikasi Rapidminer. Berdasarkan pengujian yang dilakukan menggunakan aplikasi Rapidminer didapatkan hasil yaitu: algoritma Naive Bayes mencapai akurasi 94.52%.

accuracy: 94.52% +/- 5.96% (mikro: 94.59%)

	true tidak dapat	true dapat	class precision
pred. tidak dapat	122	6	95.31%
pred. dapat	2	18	90.00%
class recall	98.39%	75.00%	

Gambar 5 Model Klasifikasi Naïve Bayes C4.5 pada RapidMiner
Sumber : Olah Pribadi

Selain dilakukan pengujian model dengan teknik *K-Fold Cross Validation*, untuk mengevaluasi kinerja model klasifikasi yang dapat memprediksi benar atau tidak benarnya oleh model

tersebut, maka digunakan teknik tabel *Confusion Matrix*. Berikut hasil tabel ukur *Confusion Matrix* terhadap Algoritma Naïve Bayes dengan menghasilkan nilai akurasi sebesar 94.52%

PerformanceVector

```
PerformanceVector:
accuracy: 94.52% +/- 5.96% (mikro: 94.59%)
ConfusionMatrix:
True:   tidak dapat     dapat
tidak dapat:    122     6
dapat:  2       18
```

Gambar 6 Confusion Matrix Naive Bayes
Sumber : Olah Pribadi

Hasil Perbandingan Model C4.5 dan Naïve Bayes. Dalam penelitian ini penulis membandingkan model klasifikasi dengan menggunakan Algoritma C4.5 dan Naïve Bayes meliputi akurasi model, tingkat error, precision, recall dan waktu komputasi. Hasil perbandingan ini diperoleh dengan bantuan *tools RapidMiner*. ini merupakan hasil perbandingan model klasfikasi antara Algoritma C.45 dan Algoritma Naïve Bayes yang ditunjukkan pada tabel 1 :

Tabel 2. Perbandingan Model Klasifikasi

Komponen	Algoritma C4.5	Naïve Bayes
Akurasi	92.52%	94.52%
Error	7.48%	5.48%
Precision	98.39%	98.39%
Recall	93.13%	95.31%
Waktu Komputasi	0.01 detik	0.01 detik

Berdasarkan tabel 4.6 dapat diperoleh hasil akurasi Algoritma C.45 lebih rendah dibandingkan Naïve Bayes yaitu 92.52% berbanding 94.52%. Tingkat error dari model klasifikasi Algoritma C.45 adalah lebih besar dibandingkan dengan Naïve Bayes yaitu sebesar 7.48% dan 5.48%. Begitupun halnya dari komponen precision dan recall Algoritma Naïve bayes lebih unggul dibandingkan Algoritma C4.5.

4. Kesimpulan

Penelitian ini mengembangkan metode untuk mendapatkan skema penerimaan beasiswa yang optimal dengan pemerataan tertinggi bagi penyelenggara sekolah. Metode tersebut dapat diterapkan karena memenuhi persyaratan pemerataan bahwa siswa yang berprestasi lebih baik harus menerima beasiswa yang sama atau lebih dari yang diterima oleh siswa yang kurang berprestasi; pemberian beasiswa meniadakan kebutuhan mahasiswa untuk mengajukan beasiswa tertentu secara manual, yang merupakan proses yang memakan waktu dan energi .Berdasarkan hasil penelitian dapat disimpulkan bahwa algoritma Naive Bayes memiliki performansi yang lebih baik. Dari hasil perbandingan tersebut didapat hasil akurasi Naïve Bayes lebih tinggi dibanding dengan Algoritma C4.5. Hasil yang didapat dari perbandingan kedua algoritma tersebut adalah Algoritma Naïve Bayes memiliki tingkat akurasi sebesar 94,52% dan algoritma C4.5 memiliki tingkat akurasi sebesar 92,52%.

Referensi

[1] N. I. Nurhidayati, Y. Yahya, F. Fathurrahman, L. . Samsu, and W. Amnia, “Implementasi Algoritma Naive Bayes Untuk Klasifikasi Penerima Beasiswa (Studi Kasus Universitas Hamzanwadi),” *Infotek J. Inform. dan Teknol.*, vol. 6, no. 1, pp. 177–188, 2023, doi: 10.29408/jit.v6i1.7529.

[2] F. Fatmawati, “Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes,” *None*, vol. 13, no. 1, pp. 50–59, 2016.

[3] M. Sari, A. Perdana Windarto, and H. Okprana, “Penerapan Data Mining Klasifikasi C4.5 Pada Penerima Beasiswa di SMK Swasta Anak Bangsa,” *BEES Bull. Electr. Electron. Eng.*, vol. 1, no. 3, pp. 115–121, 2021.

[4] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data*. 2014.

[5] Sunil Ray, “Penjelasan Pengklasifikasi Naive Bayes: Penerapan dan Masalah Praktik Pengklasifikasi Naive Bayes,” *Analytics Vidhya*, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/. [Accessed: 05-May-2024].

This page is intentionally left blank.